

# Small sample fault diagnosis method based on dual convolutional kernel feature fusion and channel attention weighted temporal convolutional network (DCK-CAM-TCN)

Shuangqiang Luo<sup>1</sup>, Xiaoyun Gong<sup>2</sup>, Wenliao Du<sup>3</sup>, Liangwen Wang<sup>4</sup>, Kunpeng Feng<sup>5</sup>,  
Yahong Qian<sup>6</sup>

<sup>1, 6</sup>Anyang Cigarette Factory, China Tobacco Industry Co., Ltd., Anyang, 455004, China

<sup>2, 3, 4, 5</sup>Henan Key Laboratory of Intelligent Manufacturing of Mechanical Equipment, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

<sup>2</sup>Corresponding author

**E-mail:** <sup>1</sup>595442140@qq.com, <sup>2</sup>2013011@zzuli.edu.cn, <sup>3</sup>dwenliao@zzuli.edu.cn, <sup>4</sup>w\_liangwen@sina.com, <sup>5</sup>jengkp12@163.com, <sup>6</sup>254829165@qq.com

Received 26 April 2025; accepted 29 August 2025; published online 29 September 2025

DOI <https://doi.org/10.21595/jve.2025.25034>



Copyright © 2025 Shuangqiang Luo, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract.** In actual industrial environments, equipment failures often occur sporadically during operation, resulting in insufficient labeled data for training. To address the issues of difficult feature extraction and poor generalization caused by insufficient data in small-sample fault diagnosis, a small sample fault diagnosis method based on dual convolutional kernel feature fusion and channel attention weighted temporal convolutional network (DCK-CAM-TCN) is proposed. Firstly, dual convolution kernels are employed to extract signal features, with the large kernel capturing low-frequency components and the small kernel extracting additional features to enhance the network's expressiveness. Secondly, the channel attention mechanism adaptively adjusts the feature responses of each channel, enabling the network to focus on the most informative and relevant features while suppressing unimportant ones. Finally, the Temporal Convolutional Network (TCN) is utilized to capture dependency features within long time series, further improving the model's ability to process sequential data. Experimental results demonstrate that the DCK-CAM-TCN model significantly outperforms traditional Convolutional Neural Networks (CNNs) and other comparison models in small-sample scenarios. The results indicate the significant advantages of the DCK-CAM-TCN model in small-sample fault diagnosis.

**Keywords:** small sample fault diagnosis, TCN, channel attention mechanisms, feature fusion.

## 1. Introduction

The reliable operation of industrial equipment is of great significance for production efficiency and safety [1], and as a key component in mechanical equipment, the fault diagnosis of bearings is particularly important [2]. However, in actual industrial production, it is difficult to obtain fault data for many devices, and data collection becomes more complex and costly when faults occur less frequently [3]. This makes it difficult for the fault diagnosis system to train effectively and make accurate judgments when faced with limited fault data [4]. Therefore, how to achieve efficient and accurate fault diagnosis with small sample has become one of the key challenges in current research [5].

Traditional bearing fault diagnosis methods are largely based on techniques such as time-frequency analysis [6] and empirical mode decomposition [7]. These methods rely on manually designed features and are difficult to adapt to complex operating conditions and diverse signal patterns. With the development of deep learning technology, convolutional neural networks (CNNs) have gradually become a research hotspot due to their powerful feature extraction capabilities. Common neural network models, such as auto-encoders [8], deep belief networks (DBNs) [9], recurrent neural networks (RNNs) [10], deep convolutional neural networks

(DCNNs) [11], and generative adversarial networks (GANs) [12], have made significant progress and have been applied in the field of fault diagnosis. The implementation of these methods typically requires designing novel and efficient network architectures or improving deep learning optimization algorithms. For example, Song et al. [13] proposed the Wide Convolutional Kernel Convolutional Neural Network (WKCNN) model to address the issue of low efficiency in traditional bearing fault diagnosis methods. Wang et al. [14] proposed a 1D-CNN based fault diagnosis method combining vibration and acoustic signals, achieving higher accuracy and robustness across various signal-to-noise ratios compared to single-modal approaches. Zhao et al. [15] put forward the CNN with mixed information model (MIXCN), which combines mixed information to enhance spatial feature extraction and reduce information loss in traditional convolution and residual connections, addressing the calculation efficiency limitations of existing complex convolutional neural networks.

In addition, small sample fault diagnosis has become a new research focus. For small sample fault diagnosis, researchers utilize feature extraction advantages of models or generate a large number of high-quality samples based on the distribution of real samples, or apply emerging machine learning techniques such as transfer learning. Lyu et al. [16] applied a novel data enhancement model gradient penalty separate classifier-Generative Adversarial Networks (GPSC-GAN), which addresses the challenge of generating high-quality multi-category fault samples by integrating a separate classifier and Wasserstein distance with gradient penalty. Li et al. [17] proposed a method based on two-dimensional vibration signal analysis and a Multi-Task Conditional Generative Adversarial Network (MTC-GAN), which effectively addresses the challenge of bearing fault diagnosis under small sample conditions through feature extraction and data augmentation, significantly improving diagnostic accuracy and robustness. Dong et al. [18] applied a fault diagnosis framework using dynamic models and transfer learning to address small sample problems, demonstrating improved fault identification through transferable features and reduced distribution discrepancies. Te et al. [19] developed a novel transfer learning framework for machinery fault diagnosis with sparse target data, utilizing paired source and target data to address distribution discrepancies and label mismatching, and demonstrate its superior performance over traditional methods through extensive experiments.

Although deep learning methods are widely used for efficient feature extraction in vibration signals, they lack the ability to dynamically adjust key feature channels. To address these issues, some researchers have introduced channel attention mechanisms into CNNs to enhance diagnostic performance through dynamic weighting. Huang et al. [20] proposed a CNN method with attention mechanisms that converts multivariate time series into images and incorporates prior knowledge, significantly improving fault diagnosis accuracy. Li et al. [21] proposed an attention mechanism-based improved CNN (AT-ICNN) to overcome the limitations of traditional diagnostic methods in extracting mechanical fault signals, enhancing fault feature extraction by combining CNN and hybrid attention mechanisms. Wang et al. [22] came up with a data-driven intelligent fault diagnosis method combining multi-head attention and CNN for automatic feature extraction and fault recognition of rolling bearings, achieving efficient bearing fault identification. Yang et al. [23] proposed a fault diagnosis method based on multi-layer Bidirectional Gated Recurrent Unit (BiGRU) and attention mechanism, combining CNN, GRU, and attention mechanism to improve the interpretability of neural networks in fault diagnosis.

Dual convolutional kernels extract multi-scale fault signatures by capturing features at different receptive fields, significantly improving the characterization and discernment of complex failure modes. In small sample fault diagnosis, large convolution kernels enhance robustness against noise-induced uncertainty [24], while deep small kernel stacks disentangle transferable fault representations from limited data. Li et al. [25] applied a dual-convolution kernel design, where parallel convolutional layers perform deep mining of fault features while the dynamic routing mechanism in capsule networks preserves spatial relationships among features, effectively addressing the challenges of small-sample dependency and poor generalization in rotating machinery fault diagnosis. Liu et al. [26] proposed a multi-scale kernel-based residual

Convolutional Neural Network (CNN) for motor fault diagnosis, aiming to address the complexity of vibration signals under non-stationary conditions, by incorporating a multi-scale kernel algorithm to capture vibration patterns of different fault types. Chang et al. [27] introduced a Concurrent Convolutional Neural Network (C-CNN) method based on dual convolutional kernels, which extracts features using convolutional kernels of different scales to effectively address noise interference in wind turbine bearing fault diagnosis. Liao [28] presents a fault diagnosis method (RACNN) for AHUs in HVAC systems, combining rule-based detection and multi-kernel 1D CNNs for feature selection, achieving 99.15 % accuracy in offline tests. However, these methods are mostly based on CNN and do not fully utilize the dynamic characteristics of the time series.

In addition, time step information cannot be ignored in vibration signals. To effectively capture temporal features and hidden information at different positions in time series, a common strategy is to adopt gated Recurrent Neural Network (RNN) structures, such as Long Short-Term Memory (LSTM), or alternative architectures like Temporal Convolutional Networks (TCN). Although LSTM excels in temporal modeling, its large number of parameters makes it prone to overfitting in small-sample scenarios. Additionally, assuming that signals propagate information in only one direction is not reasonable. Therefore, TCN, as an alternative with fewer parameters and the ability to capture both forward and backward information, becomes a more optimal choice. Guo et al. [29] developed a fault diagnosis method for IGBT open-circuit faults in Modular Multilevel Converter (MMC) systems, combining TCN with Adaptive Chirp Mode Decomposition (ACMD) and Silhouette Coefficient (SC) to effectively extract long-term sequence features and maintain robustness under noisy conditions. Ai et al. [30] proposed an automatic temporal convolution network method based on TCN and enhanced elite genetic algorithm (SEGA) optimization for sensor faults of hypersonic aircraft, and combined SPRT and wavelet packet transform (WPT) to improve the accuracy of fault diagnosis. However, TCN lacks the ability to extract multi-scale features and focus key features, limiting further improvements in its diagnostic performance. Previous methods have achieved relatively satisfactory results, deep learning models typically require a large number of samples to achieve ideal generalization performance. However, due to the relatively small amount of annotated data, models often cannot fully learn various effective features from limited samples and are prone to overfitting, which increases the difficulty of learning [31]. In addition, the comparative study of new activation functions in small sample fault diagnosis has not been thoroughly explored.

Therefore, in response to the above issues, this paper proposes a small-sample fault diagnosis method based on dual convolutional kernel feature fusion and channel attention weighting within a Temporal Convolutional Network. By combining dual convolutional kernel feature fusion and a channel attention mechanism, the model not only extracts multi-scale features comprehensively but also adaptively focuses on key channel features. Under small sample conditions, the proposed method still demonstrates good diagnostic performance.

The main contributions of the paper are as follows:

1) Multi-Scale Feature Extraction via Dual Convolutional Kernel Fusion and Channel Attention Mechanism. This innovation overcomes the limitations of single-kernel feature extraction and significantly strengthens feature representation, offering a robust solution for small-sample fault diagnosis.

2) Temporal Convolutional Network (TCN) with Dilated Convolutions for Long-Range Dependency Modeling. This approach maintains computational efficiency while mitigating gradient vanishing issues. This innovation provides a stable and effective framework for temporal feature extraction under small-sample constraints.

3) Adaptive Pooling and End-to-End Classification Framework Integration. By combining adaptive average pooling with a fully connected classifier, this framework eliminates the need for manual feature engineering and reduces reliance on fixed-dimensional inputs. This innovation streamlines the diagnostic workflow and significantly enhances generalization performance and classification accuracy under limited data conditions.

The structure of this paper is as follows:

Chapter 2 elaborates on the fundamental theoretical principles of techniques such as Temporal Convolutional Networks and Channel Attention Mechanisms. Chapter 3 describes the overall framework of the proposed method in this paper. Chapter 4 introduces the sources, settings, and environment of the experimental data, evaluates the model performance through comparative experiments, and validates the effectiveness of the relevant mechanisms. Chapter 5 primarily summarizes the research findings.

## 2. Theoretical background

### 2.1. Principle of feature fusion technology

Dual convolution kernels are used to extract signal features, where the large convolution kernel focuses on extracting low-frequency features from the signal, while the small convolution kernel is used to extract other features and deepen the expressive power of the neural network [3].

Path  $p1$  uses larger convolution kernels (Kernel size = 18) for convolution operations. The first layer of convolution, as shown in Eq. (1):

$$p1^{(1)} = \text{Conv1D}(x, W_1) + b_1, \quad (1)$$

where  $W_1$  is the convolution kernel weight,  $b_1$  is the bias term.

The second layer convolution, as shown in Eq. (2):

$$p1^{(2)} = \text{Conv1D}(p1^{(1)}, W_2) + b_2. \quad (2)$$

Afterwards, the maximum value is pooled in Eq. (3):

$$p1 = \text{MaxPool1D}(p1^{(2)}). \quad (3)$$

Path  $p2$  uses a smaller convolution kernel (Kernel size = 6) for convolution operation. The convolution result of channel  $p2$  is obtained through the following process:

$$\begin{aligned} p2^{(1)} &= \text{Conv1D}(x, W_3) + b_3, \\ p2^{(2)} &= \text{Conv1D}(p2^{(1)}, W_4) + b_4, \\ p2^{(3)} &= \text{MaxPool1D}(p2^{(2)}), \\ p2^{(4)} &= \text{Conv1D}(p2^{(3)}, W_5) + b_5, \\ p2^{(5)} &= \text{Conv1D}(p2^{(4)}, W_6) + b_6, \\ p2 &= \text{MaxPool1D}(p2^{(5)}), \end{aligned} \quad (4)$$

where  $W_i$  is the convolution kernel weight,  $b_i$  is the bias term.

Merge the outputs of  $p1$  and  $p2$  by element wise multiplication, as shown in Eq. (5):

$$X = p1 \odot p2, \quad (5)$$

where  $\odot$  represents element wise multiplication,  $X$  is the fused feature.

The features from the large and small convolution kernels are fused to form a combined feature rich in information. These fused features are input into the Channel Attention Mechanism-Temporal Convolutional Network (CAM-TCN) neural network model.

### 2.2. Channel attention mechanism (CAM)

The channel attention mechanism plays a pivotal role within convolutional neural networks, serving to adaptively recalibrate channel-wise feature responses [32]. A typical convolutional

network, through multiple layers of convolution, generates feature maps of dimensions  $H \times W \times C$  (where  $H$  denotes height,  $W$  denotes width, and  $C$  represents the number of channels), with each channel encapsulating distinct information.

When determining the importance of channels, a global average pooling operation is first applied to the feature map  $X \in \mathbb{R}^{H \times W \times C}$ , as shown in Eq. (6), resulting in a  $1 \times 1 \times C$  vector  $Z$ :

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_{i,j,c}, \quad (6)$$

where  $z_c$  represents the result of the global average pooling for the  $c$ -th channel,  $x_{i,j,c}$  denotes the element of the feature map at position  $(i, j)$  in the  $c$ -th channel. This operation aggregates the spatial information of each channel, thereby reducing the feature map to a  $1 \times 1 \times C$  vector  $Z \in \mathbb{R}^{1 \times 1 \times C}$ , which serves as a global descriptor for the information of each channel. Then, after passing through two fully connected layers, the first fully connected layer is shown in Eq. (7):

$$s = \sigma(W_7 s + b_7) \left( W_7 \in \mathbb{R}^{C \times \frac{C}{r}}, b_7 \in \mathbb{R}^{\frac{C}{r}}, s \in \mathbb{R}^{1 \times 1 \times C} \right), \quad (7)$$

where  $W_7 \in \mathbb{R}^{C \times \frac{C}{r}}$  is the weight matrix of the first fully connected layer,  $b_7 \in \mathbb{R}^{\frac{C}{r}}$  is the bias vector,  $\sigma$  is the activation function,  $s \in \mathbb{R}^{1 \times 1 \times C}$  represents the result after passing through the first fully connected layer. The second fully connected layer maps the  $C/r$  dimension back to the  $C$  dimension, as shown in Eq. (8):

$$e = \sigma(W_8 s + b_8) \left( W_8 \in \mathbb{R}^{C \times \frac{C}{r}}, b_8 \in \mathbb{R}^C, e \in \mathbb{R}^{1 \times 1 \times C} \right), \quad (8)$$

where  $W_8 \in \mathbb{R}^{C \times \frac{C}{r}}$  is the weight matrix of the second fully connected layer,  $b_8 \in \mathbb{R}^C$  is the bias vector,  $e \in \mathbb{R}^{1 \times 1 \times C}$  is the result after passing through the second fully connected layer, where it represents the channel attention weight.

Finally, the weight is multiplied by the original feature map by channel to obtain the recalibrated feature map  $X'$ , shown in Eq. (9):

$$x'_{i,j,c} = e_c \times x_{i,j,c}. \quad (9)$$

This adaptive channel recalibration enables the model to focus on the most relevant channel features in specific tasks, improving the network's discriminative ability and overall performance.

### 2.3. Temporal convolutional network (TCN)

The Temporal Convolutional Network (TCN) demonstrates outstanding performance in handling sequential data [33]. Its specific structure is shown in Fig. 1.

In its structure, causal convolutional layers strictly follow the principle of temporal causality. For the input sequence  $x = [x_1, x_2, \dots, x_T]$ , the output  $y_t$  of causal convolution is calculated according to Eq. (10):

$$y_t = \sum_{k=0}^{K-1} f(x_{t-k}), \quad (10)$$

where  $K$  is the size of the convolution kernel, and  $f$  is the function corresponding to the convolution kernel. This structure effectively avoids the interference of future information on the

current output, thereby accurately capturing the changing patterns in temporal data.

Dilated convolution expands the receptive field through the dilation factor  $d$ , for example, its output  $y_t = \sum_{k=0}^{K-1} w_k x_{t-kd}$  (where  $w$  is the dilated convolution kernel), successfully mining long-distance dependencies in the data without increasing computational complexity. The multi-layer stacking architecture of TCN enables each layer to gradually extract temporal features from local short-term to overall long-term.

The residual connection mechanism in the network (expressed as  $y = F(x) + x$ , where  $F(x)$  is the convolutional transformation function) effectively alleviates the gradient problem that is prone to occur as the network depth increases.

During the training phase, stochastic gradient descent (SGD) is used in TCN for training [34]. The parameter update formula can be expressed as Eq. (11):

$$\theta = \theta - \eta \frac{\partial L(\theta)}{\partial \theta}, \quad (11)$$

where  $\eta$  represents the learning rate. Usually, the training process is optimized by combining the learning rate decay strategy, momentum method (where the update formula for velocity variable  $v$  in momentum method is  $v = \beta v - \eta \frac{\partial L(\theta)}{\partial \theta}$ , and the parameter update formula is  $\theta = \theta + v$ , where  $\beta$  is the momentum coefficient), and adaptive learning rate algorithms such as Adagrad, Adadelata, Adam, etc. At the same time, the use of regularization techniques to prevent overfitting ensures the generalization ability and reliability of TCN in temporal data processing applications.

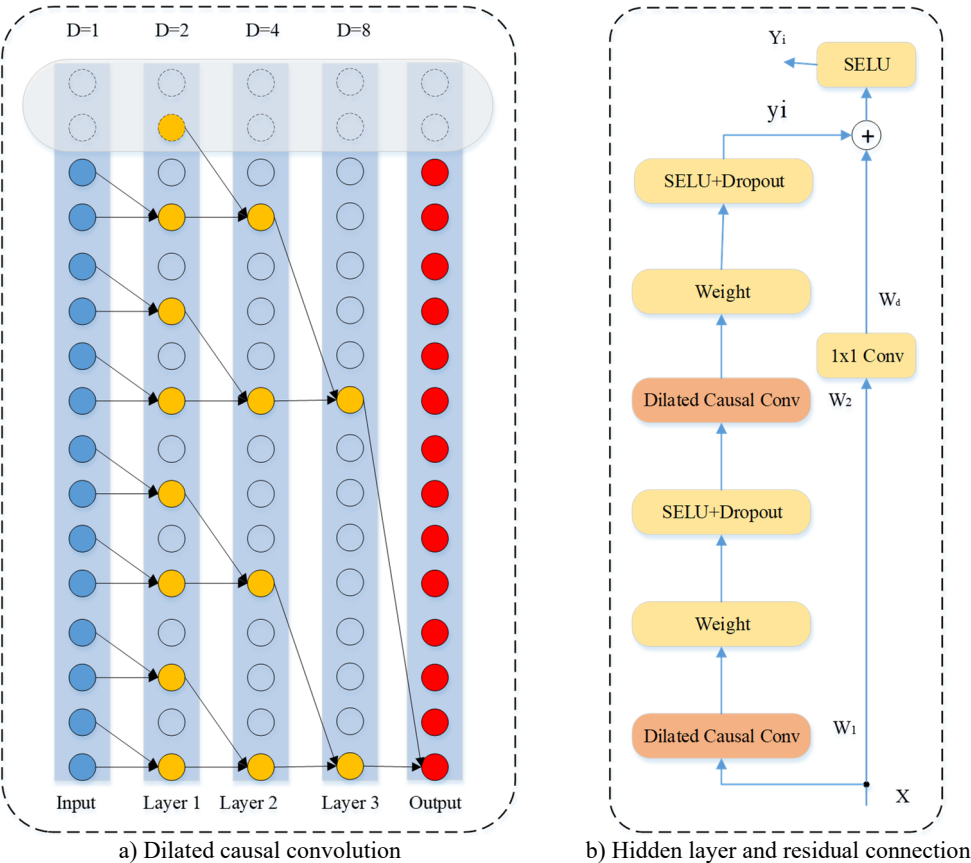


Fig. 1. Complete structure of TCN

3. The proposed fault diagnosis method

A small sample bearing fault diagnosis method based on dual convolutional kernel feature fusion and channel attention weighted temporal convolutional network is proposed, which combines feature fusion, attention mechanism, TCN and adaptive pooling. The specific steps of fault diagnosis based on DCK-CAM-TCN are as follows:

- 1) Data preprocessing: Convert the original one-dimensional signal into tensor format and normalize it to meet the input requirements of the model.
- 2) Multi scale convolution feature extraction: Extracting features from different frequency bands through two convolution branches, combined with pooling layers to reduce dimensionality.
- 3) Feature fusion and attention weighting: Multiply and fuse the element points at the corresponding positions extracted by two convolution kernels, and generate weighted features using attention mechanism.
- 4) Temporal Feature Modeling (TCN): Use Temporal Convolutional Network (TCN) to extract features with long-term dependencies and expand the receptive field through dilated convolution.
- 5) Reduce dimensions and classification decisions: After adaptive average pooling dimensionality reduction, it is mapped to the fault category space through a fully connected layer to complete classification.

The overall architecture of the DCK-CAM-TCN model is shown in Fig. 2.

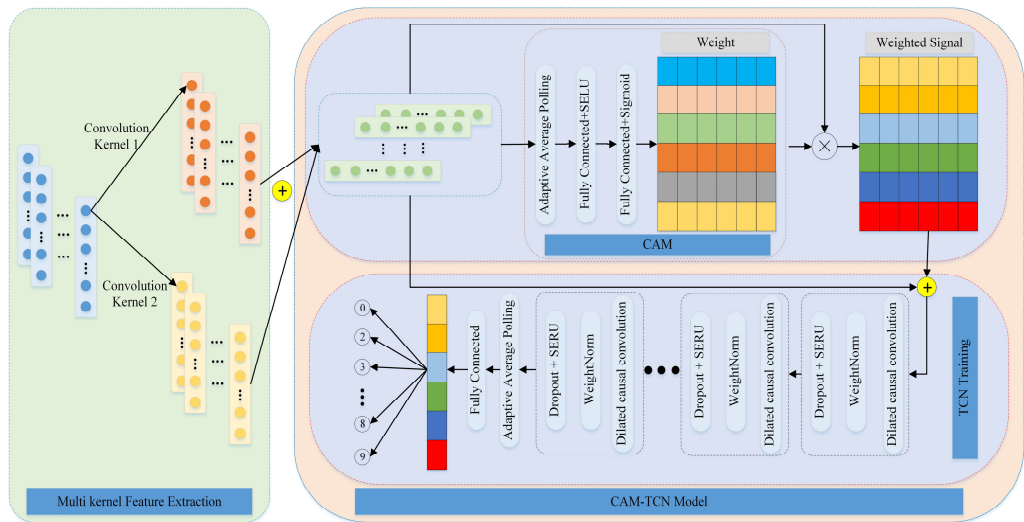


Fig. 2. Overall schema for the proposed network architecture of DCK-CAM-TCN

4. Experiment

4.1. Experimental setup

The experiments were implemented in PyTorch 1.12.0, and Python3.8, running on Intel (R) Core i7-10700K. Other parameters in the model are shown in Table 1.

Table 1. Critical parameters of the model

Parameter	Description	Parameter	Description	Parameter	Description
Learning rate	0.001	Input channels	30	Activation function	SELU
Optimizer	Adam	Output channels	[64, 64]	Dropout rate	0.2
Epochs	150	Kernel size	3	Network depth	2
Batch size	64	Dilation rates	[1, 2]	Normalization	BatchNorm1d

## 4.2. Case 1: bearing fault data of West Reserve University

### 4.2.1. Dataset description

The driving end rolling bearing data provided by Case Western Reserve University is collected by the equipment shown in Fig. 3, wherein the faults (Inner ring fault (IR), Outer ring fault (OR) and Rolling element fault (RE)) are generated by electric discharge machining (EDM), the sampling frequency is 12 kHz, the load is 1HP, and the three damage degrees (the fault diameter is 0.118/0.356/0.533 mm respectively). An acceleration sensor located at the drive end of the motor housing collects acceleration data. Each fault type includes 102400 data points, of which 1024 data points are taken as samples, and each fault type has 100 samples. The specific fault types and labels are shown in Table 2.

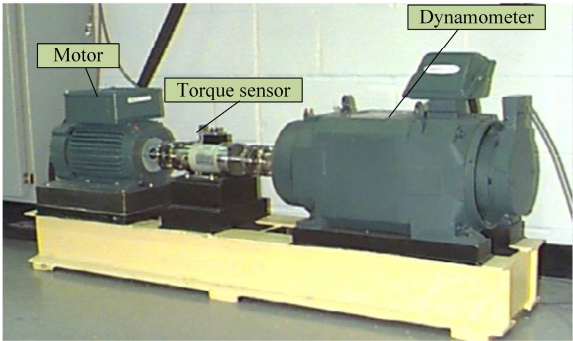


Fig. 3. CWRU test bench

### 4.2.2. Ablation comparative experiment

Vibration signals exhibit a nonlinear distribution, whereas neural networks inherently perform linear computations. To mitigate the issue of vanishing gradients, nonlinear non-saturating activation functions are commonly employed. To comprehensively evaluate the impact of activation functions on model performance, the nonlinear characteristics of these activation functions are first visualized in Fig. 4. Subsequently, the loss and accuracy curves under various activation functions, including ReLU, LeakyReLU, ELU, Softplus, and SELU, are systematically compared to assess their effectiveness in the proposed model.

Table 2. Fault dataset description of CWRU

Fault type	Fault description	Label	Number of samples
Normal	Normal bearing	0	100
OR/RE/IR	0.1778 mm	1/2/3	100
OR/RE/IR	0.3556 mm	4/5/6	100
OR/RE/IR	0.5334 mm	7/8/9	100

All results are carried out under the CWRU dataset with a training-set ratio of 0.4, and the training loss and transfer accuracy are obtained, as shown in Fig. 5 and 6. It can be seen that all models can converge, with LeakyReLU having the maximum loss of 0.61. In terms of convergence stability, except for LeakyReLU, the other five activation functions are relatively stable, where the differences in loss are small in the later epochs. The difference in final losses among ReLU, Softplus, and SELU is about 0.005, while SELU requires fewer epochs and achieves the fastest convergence. Therefore, in the subsequent experiments, SELU is used as the activation function.

This section aims to evaluate the performance differences among Adam, Adagrad, and Adadelta optimizers in deep learning tasks, with a focus on convergence speed, model accuracy,



and training stability. The rationale for selecting Adam as the primary optimizer is validated through systematic comparisons. The study employs CWRU datasets and measures key metrics including training loss, test accuracy, and gradient noise sensitivity. The results are summarized in Table 3, highlighting the performance of each optimizer under identical hyperparameter settings.

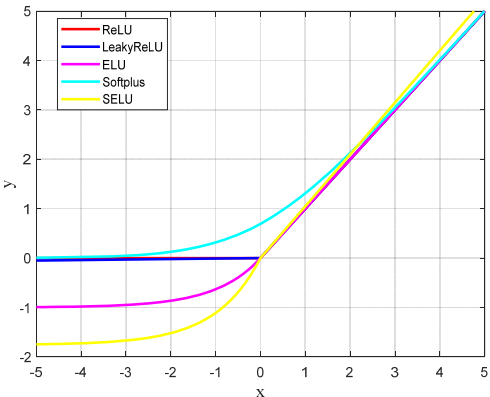


Fig. 4. Different activation functions

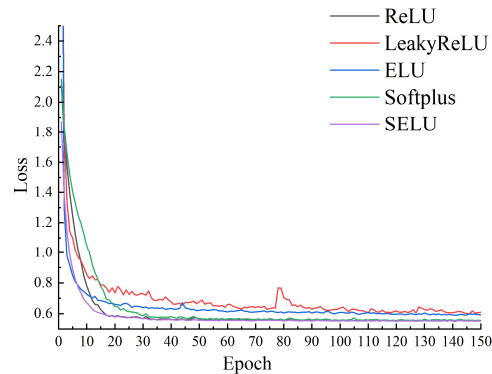


Fig. 5. Loss under different activation functions

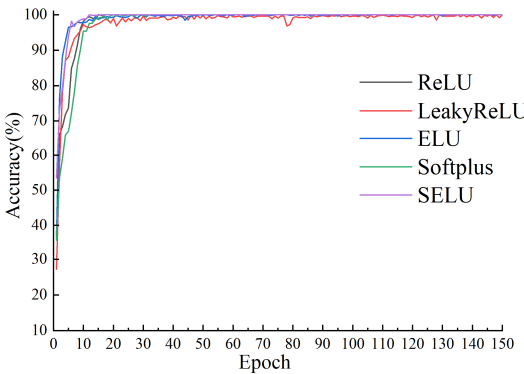


Fig. 6. Accuracy under different activation functions

Table 3. Test results of different optimizers

Optimizer	Test accuracy (%)	Convergence epochs	Final training loss
Adam	100	25	0.52
Adagrad	94.86	33	0.75
Adadelta	92.54	30	1.10

The ablation study confirms that Adam outperforms Adagrad and Adadelta in most critical metrics. Its adaptive learning rate mechanism, combined with momentum-based updates, provides superior convergence speed and stability, making it suitable for complex, high-dimensional tasks. While Adagrad remains effective for sparse data, its learning rate decay limits long-term training. Therefore, Adam is selected as the default optimizer in this study due to its robustness and efficiency.

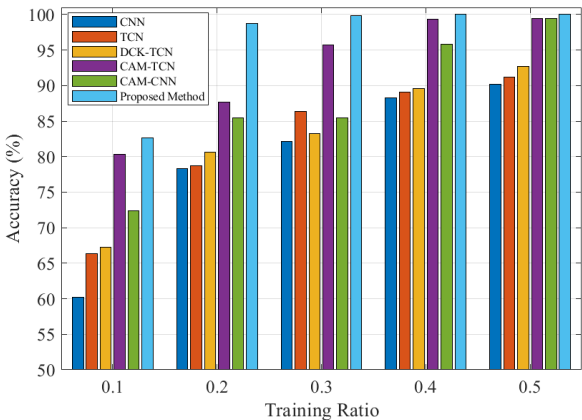
This section introduces a systematic evaluation of the learning rate (LR) impact on model performance. The goal is identify the optimal LR range for the model and highlight the trade-offs between training efficiency and final performance. The results are summarized in Table 4, with all other hyperparameters fixed.

The learning rate sensitivity analysis confirms that 0.001 is the optimal LR for this model, achieving the highest test accuracy and lowest training loss while maintaining rapid convergence

and stability.

**Table 4.** Experimental results of different learning rates

Learning rate	Test accuracy (%)	Final training loss	Convergence speed (epochs)
0.0001	97.50	1.05	54
0.0005	98.83	0.61	53
0.001	100	0.52	25
0.005	97.83	0.62	48
0.01	98.50	0.63	53



**Fig. 7.** Accuracy values of test-set under different model and training-set ratio

To verify the diagnostic performance of the model in small samples, compares the diagnostic accuracy of the test set under different training set ratios in this paper. Fig. 7 illustrates the highest accuracy achieved upon model convergence. The experimental results demonstrate that the traditional CNN and TCN performs the worst under small sample conditions. As the proportion of the training set increases, the accuracy of CNN and TCN gradually improves, reaching 90.17 % and 91.23 % respectively, when 50 % of the training samples are used. However, they still lag behind the other models presented in this study. Comparative analysis between TCN and DCK-TCN demonstrates that the DCK module significantly enhances the model's classification accuracy. The CAM-TCN and CAM-CNN models significantly enhance feature extraction by incorporating a Channelized Attention Mechanism (CAM), with a particularly notable improvement under small sample conditions. CAM-TCN achieves accuracies of 80.36 % and 87.67 % at 10 % and 20 % of the training set, respectively, outperforming CAM-CNN, which reaches 72.34 % and 85.46 %. This suggests that the time series modeling capability of CAM-TCN better facilitates fault identification in small sample scenarios. The proposed method consistently outperforms the other models, achieving an accuracy of 82.67 % across all training set proportions, and reaching 100 % test accuracy at 40 % of the training set. These results demonstrate better diagnostic performance in small sample tests.

**4.2.3. Visual analysis of results**

Optimal performance is achieved with 100 % accuracy when utilizing the full 40 % training dataset, establishing this as a benchmark for comparative analysis. Therefore, the dataset was divided into a training set and test set in a 4:6 ratio, with the experimental results presented below. Fig. 8 and 9 illustrate the accuracy curve and the loss curve, respectively. The accuracy curve basically converges after 25 epochs. To better evaluate the contribution of each module in fault diagnosis, this study applies visual dimension reduction on the raw data, feature fusion data, CAM data, and data processed by the TCN sequentially through t-SNE analysis, as shown in Fig. 10.

The results reveal that the data distribution becomes more concentrated after each processing layer, with the distinction between categories significantly enhanced. Notably, the diagnostic performance between fault categories is greatly improved, particularly in the data weighted by channel attention. The TCN module further optimizes feature representation, leading to a substantial improvement in fault classification performance. The confusion matrix results for the test set (Fig. 11) demonstrate that the collaboration of the modules effectively enhances the model's ability to identify small sample fault diagnosis tasks, confirming the effectiveness of the proposed method in fault diagnosis.

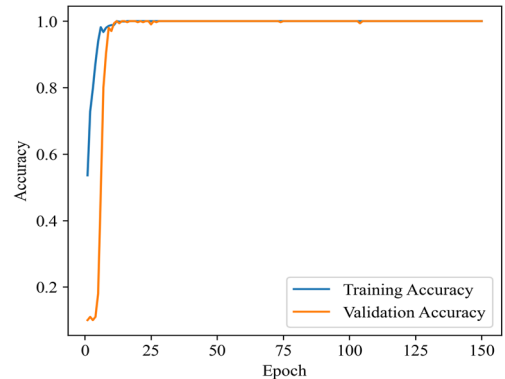


Fig. 8. Accuracy curve

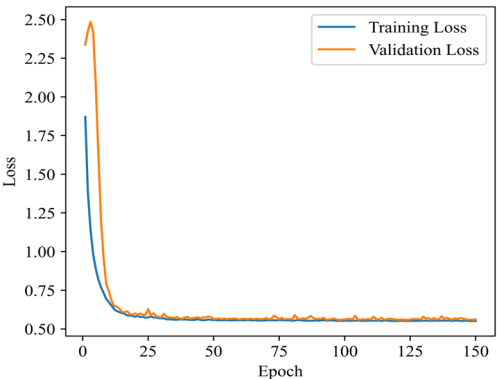
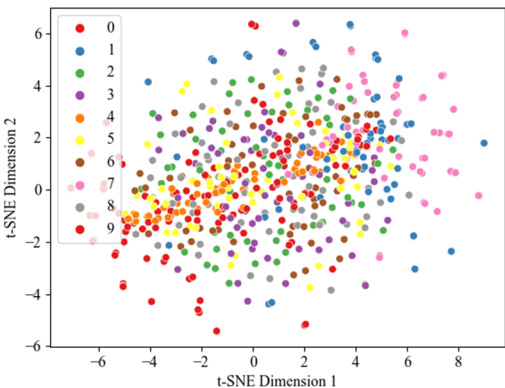
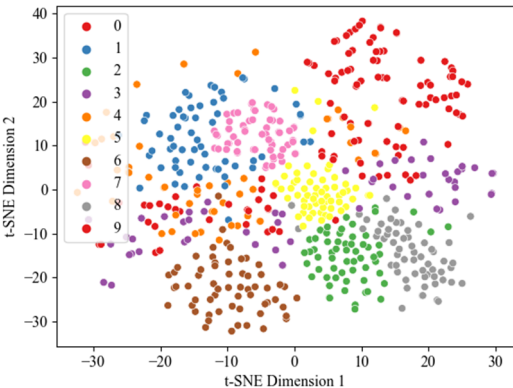


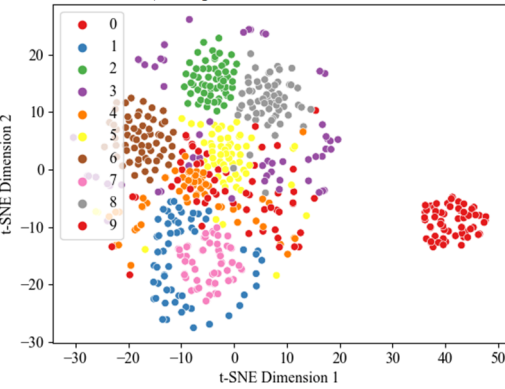
Fig. 9. Loss curve



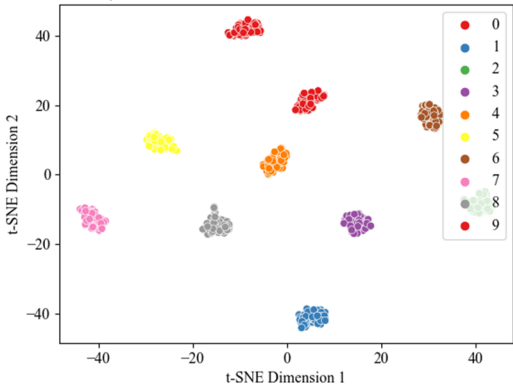
a) Original data T-SNE



b) t-SNE after convolution fusion



c) t-SNE after attention mechanism



d) t-SNE after TCN

Fig. 10. t-SNE visualization

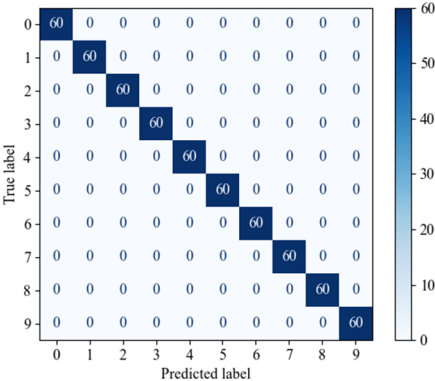


Fig. 11. Confusion matrix diagram

4.3. Case 2: Fault data of mechanical fault comprehensive test bench

To validate the generalization ability of the proposed method, this study further conducted experimental verification using the fault data from the mechanical fault comprehensive test bench.

4.3.1. Dataset description

The mechanical failure comprehensive test stand from SQ Company (USA) is employed to simulate bearing failure data at the load end. The Mechanical Fault Comprehensive Test Bench is illustrated in Fig. 12. The bearings at the load end is ER-12K. A 0.5 mm deep groove is machined on the outer and inner rings, as well as the rolling elements, to simulate failure of the bearing outer, inner rings and rolling elements fault, respectively. The motor operates at speeds of 1200 rpm, 1800 rpm, and 2400 rpm, with a sampling frequency of 12,800 Hz. The vibration signals utilized in this study are acquired by an acceleration sensor located at the load end. Each fault type consists of 102,400 data points, from which 1,024 data points are selected as samples, with each fault type containing 100 samples (training set: test set = 4:6). The corresponding labels for the various data types are presented in Table 5.

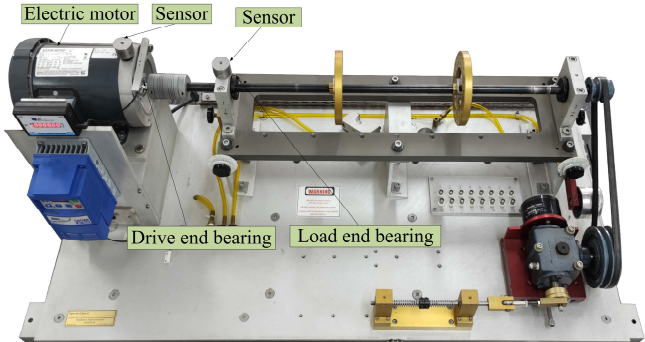


Fig. 12. Mechanical fault comprehensive test bench

Table 5. Dataset description of comprehensive test bench

Fault type	Rotational speed	Label	Number of samples
Normal	1200 r/min	0	100
RE /IR/OR		1/2/3	100
RE /IR/OR	1800 r/min	4/5/6	100
RE /IR/OR	2400 r/min	7/8/9	100

4.3.2. Result and analysis

This experiment compares the performance of four models in classifying a dataset with a training-to-test set ratio of 4:6, with the results presented in Fig. 13. Fig. 13 illustrates the highest accuracy achieved upon model convergence. The traditional CNN achieved an accuracy of 88.14 %, demonstrating its ability to handle basic tasks, though it does not fully capture the deep correlations within the feature data. The DCK module improves the model's classification accuracy, as shown by comparing TCN and DCK-TCN. In contrast, the model combining CAM-TCN significantly improved performance, attaining an accuracy of 96.78 %. This indicates that the attention mechanism enhances the model's ability to process temporal data. Further optimization of the CNN model by integrating the attention mechanism (CAM-CNN) resulted in an accuracy of 98.83 %, highlighting its superior capability in feature extraction and the distribution of importance weights. Finally, the DCK-CAM-TCN model achieved perfect accuracy of 99.98 %, demonstrating that this combination effectively captures data features and addresses the classification challenges in the experiment.

These results suggest that the introduction of model complexity and specific mechanisms, such as CAM and DCK, significantly enhances classification performance. The loss curve shown in Fig. 14 indicates that the loss steadily decreases and stabilizes as training progresses, signifying the successful convergence of the model. Additionally, the accuracy curve in Fig. 15 illustrates the continuous improvement in both training and validation accuracy, ultimately reaching 100 % accuracy.

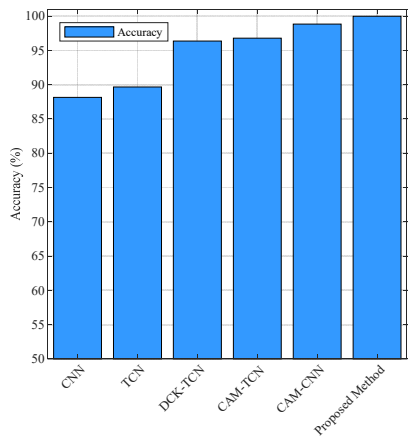


Fig. 13. Accuracy values of test-set under different mode

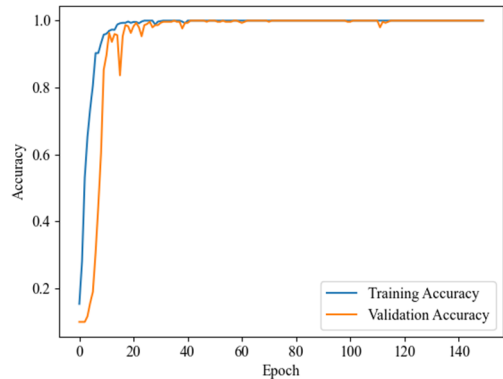


Fig. 14. Accuracy curve

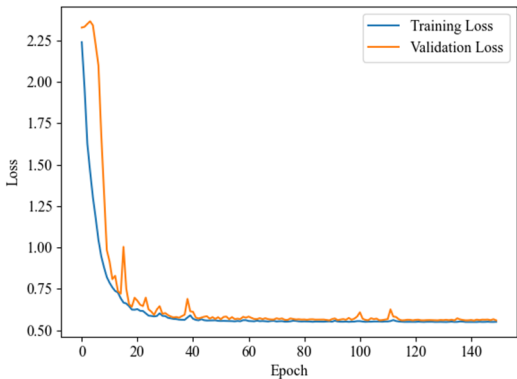


Fig. 15. Loss curve

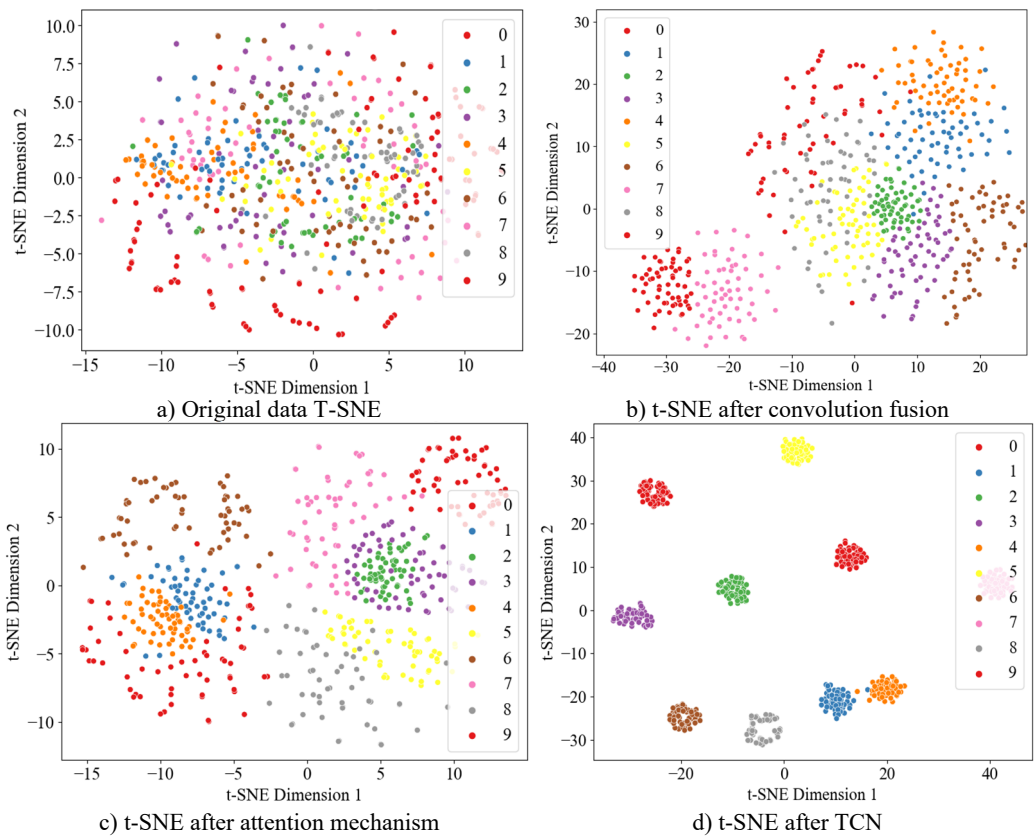


Fig. 16. t-SNE visualization

4.3.3. Visual analysis of results

The dataset also was divided into a training set and test set in a 6:4 ratio, with the experimental results presented in the Fig. below. The t-SNE dimensionality reduction visualization results are shown in Fig. 16, where the learning effects of each layer's features can be intuitively observed. In the DCK-CAM-TCN model, the t-SNE plots for each module clearly demonstrate the separation between categories, proving that the introduction of the CAM and DCK mechanisms effectively enhances the model's ability to process data features. Particularly in the deeper feature mappings of the model, the data distribution becomes more concentrated, and the differences between categories become more pronounced, further validating the effectiveness of the model. Overall, with the increase in model complexity and the introduction of these mechanisms, the DCK-CAM-TCN demonstrates exceptional performance and superior generalization ability in fault diagnosis.

5. Conclusions

This paper presents a novel small-sample fault diagnosis method based on dual convolutional kernel feature fusion and channel attention weighted temporal convolutional network (DCK-CAM-TCN) to address the challenges of insufficient labeled data and low accuracy in fault diagnosis of industrial equipment. Dual convolutional kernel feature fusion enables the extraction of both low-frequency and high-frequency features, providing comprehensive multi-scale information. The channel attention mechanism adaptively weights the channels to focus on crucial features, enhancing the model's discriminative ability. Temporal Convolutional Network (TCN)

is employed to handle long-sequence data effectively, capturing long-term dependencies and expanding the receptive field.

Experimental results on two datasets demonstrate the superiority of the proposed method. It shows that the data distribution becomes more concentrated and the category distinctions are enhanced after each processing step, indicating that the model can extract more discriminative features. In comparison with traditional Convolutional Neural Networks (CNNs) and other models, DCK-CAM-TCN achieves significantly higher accuracies, especially in small-sample scenarios. For instance, with only 20 % of the training data, it reaches 97 % convergence accuracy on the test set in certain cases, highlighting its excellent diagnostic capabilities. Performance peaks at 100 % accuracy when 40 % of the training data is used. Visual analysis through t-SNE further validates the effectiveness of each module in the model.

In summary, the DCK-CAM-TCN model proposed in this paper provides an effective solution for small-sample fault diagnosis. However, the model's generalization capability in extreme small-sample regimes below 10 % training data requires further exploration. Future research could focus on further optimizing the model, exploring its application in more complex industrial scenarios, and potentially combining it with other emerging techniques to achieve even better performance in the field of fault diagnosis.

## Acknowledgements

This research was funded in part by the National Nature Science Foundation of China under Grant (52275138), in part by Postgraduate Education Reform and Quality Improvement Project of Henan Province(YJS2025AL32), and in part by Science and Technology Innovation Project of China Tobacco Industry Co., Ltd. (AW2023024; AYBW202404).

## Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Author contributions

Shuangqiang Luo: supervision, conceptualization, methodology. Xiaoyun Gong: methodology, investigation, validation, writing-original draft preparation, writing-review and editing. Wenliao Du: project administration, writing-review and editing. Liangwen Wang: data curation, software. Kunpeng Feng: investigation, validation, translating, editing. Yahong Qian: supervision, resources.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

- [1] Q. Li, "Fractional quaternion central difference Kalman filter with random walk motion for deterioration prognosis of rolling bearing with multiple jump outliers," *Mechanical Systems and Signal Processing*, Vol. 234, p. 112844, Jul. 2025, <https://doi.org/10.1016/j.ymssp.2025.112844>
- [2] I. Misbah, C. K. M. Lee, and K. L. Keung, "Fault diagnosis in rotating machines based on transfer learning: Literature review," *Knowledge-Based Systems*, Vol. 283, p. 111158, Jan. 2024, <https://doi.org/10.1016/j.knosys.2023.111158>
- [3] X. Zhang, C. He, Y. Lu, B. Chen, L. Zhu, and L. Zhang, "Fault diagnosis for small samples based on attention mechanism," *Measurement*, Vol. 187, p. 110242, Jan. 2022, <https://doi.org/10.1016/j.measurement.2021.110242>

- [4] H. Su, L. Xiang, A. Hu, Y. Xu, and X. Yang, "A novel method based on meta-learning for bearing fault diagnosis with small sample learning under different working conditions," *Mechanical Systems and Signal Processing*, Vol. 169, p. 108765, Apr. 2022, <https://doi.org/10.1016/j.ymssp.2021.108765>
- [5] T. Zhang et al., "Intelligent fault diagnosis of machines with small and imbalanced data: A state-of-the-art review and possible extensions," *ISA Transactions*, Vol. 119, pp. 152–171, Jan. 2022, <https://doi.org/10.1016/j.isatra.2021.02.042>
- [6] Z. Feng, M. Liang, and F. Chu, "Recent advances in time-frequency analysis methods for machinery fault diagnosis: A review with application examples," *Mechanical Systems and Signal Processing*, Vol. 38, No. 1, pp. 165–205, Jul. 2013, <https://doi.org/10.1016/j.ymssp.2013.01.017>
- [7] J. Chen, D. Zhou, C. Lyu, and C. Lu, "An integrated method based on CEEMD-SampEn and the correlation analysis algorithm for the fault diagnosis of a gearbox under different working conditions," *Mechanical Systems and Signal Processing*, Vol. 113, pp. 102–111, Dec. 2018, <https://doi.org/10.1016/j.ymssp.2017.08.010>
- [8] L. Chen, Y. Ma, H. Hu, and U. S. Khan, "An effective fault diagnosis approach for bearing using stacked de-noising auto-encoder with structure adaptive adjustment," *Measurement*, Vol. 214, p. 112774, Jun. 2023, <https://doi.org/10.1016/j.measurement.2023.112774>
- [9] Z. Zhu et al., "A review of the application of deep learning in intelligent fault diagnosis of rotating machinery," *Measurement*, Vol. 206, p. 112346, Jan. 2023, <https://doi.org/10.1016/j.measurement.2022.112346>
- [10] T.-T. Vo, M.-K. Liu, and M.-Q. Tran, "Harnessing attention mechanisms in a comprehensive deep learning approach for induction motor fault diagnosis using raw electrical signals," *Engineering Applications of Artificial Intelligence*, Vol. 129, p. 107643, Mar. 2024, <https://doi.org/10.1016/j.engappai.2023.107643>
- [11] S. Manikandan and K. Duraivelu, "Vibration-based fault diagnosis of broken impeller and mechanical seal failure in industrial mono-block centrifugal pumps using deep convolutional neural network," *Journal of Vibration Engineering and Technologies*, Vol. 11, No. 1, pp. 141–152, May 2022, <https://doi.org/10.1007/s42417-022-00566-0>
- [12] Z. Yin, F. Zhang, C. Yin, G. Xu, and S. Liu, "A bearing fault diagnosis method for sample imbalance," *Engineering Applications of Artificial Intelligence*, Vol. 157, p. 111171, Oct. 2025, <https://doi.org/10.1016/j.engappai.2025.111171>
- [13] X. Song, Y. Cong, Y. Song, Y. Chen, and P. Liang, "A bearing fault diagnosis model based on CNN with wide convolution kernels," *Journal of Ambient Intelligence and Humanized Computing*, Vol. 13, No. 8, pp. 4041–4056, Apr. 2021, <https://doi.org/10.1007/s12652-021-03177-x>
- [14] X. Wang, D. Mao, and X. Li, "Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network," *Measurement*, Vol. 173, p. 108518, Mar. 2021, <https://doi.org/10.1016/j.measurement.2020.108518>
- [15] Z. Zhao and Y. Jiao, "A fault diagnosis method for rotating machinery based on CNN with mixed information," *IEEE Transactions on Industrial Informatics*, Vol. 19, No. 8, pp. 9091–9101, Aug. 2023, <https://doi.org/10.1109/tii.2022.3224979>
- [16] P. Lyu, Y. Cheng, M. Zhang, W. Yu, L. Xia, and C. Liu, "GPSC-GAN: a data enhanced model for intelligent fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, Vol. 73, pp. 1–16, Jan. 2024, <https://doi.org/10.1109/tim.2024.3457925>
- [17] J. Li, Y. Wei, and X. Gu, "MTC-GAN bearing fault diagnosis for small samples and variable operating conditions," *Applied Sciences*, Vol. 14, No. 19, p. 8791, Sep. 2024, <https://doi.org/10.3390/app14198791>
- [18] D. Xiao, Y. Huang, C. Qin, Z. Liu, Y. Li, and C. Liu, "Transfer learning with convolutional neural networks for small sample size problem in machinery fault diagnosis," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, Vol. 233, No. 14, pp. 5131–5143, Mar. 2019, <https://doi.org/10.1177/0954406219840381>
- [19] T. Han, C. Liu, R. Wu, and D. Jiang, "Deep transfer learning with limited data for machinery fault diagnosis," *Applied Soft Computing*, Vol. 103, p. 107150, May 2021, <https://doi.org/10.1016/j.asoc.2021.107150>
- [20] Y. Huang, J. Zhang, R. Liu, and S. Zhao, "Improving accuracy and interpretability of CNN-based fault diagnosis through an attention mechanism," *Processes*, Vol. 11, No. 11, p. 3233, Nov. 2023, <https://doi.org/10.3390/pr11113233>



- [21] X. Li, S. Xiao, F. Zhang, J. Huang, Z. Xie, and X. Kong, "A fault diagnosis method with AT-ICNN based on a hybrid attention mechanism and improved convolutional layers," *Applied Acoustics*, Vol. 225, p. 110191, Nov. 2024, <https://doi.org/10.1016/j.apacoust.2024.110191>
- [22] H. Wang, J. Xu, R. Yan, C. Sun, and X. Chen, "Intelligent bearing fault diagnosis using multi-head attention-based CNN," *Procedia Manufacturing*, Vol. 49, pp. 112–118, Jan. 2020, <https://doi.org/10.1016/j.promfg.2020.07.005>
- [23] Z.-B. Yang, J.-P. Zhang, Z.-B. Zhao, Z. Zhai, and X.-F. Chen, "Interpreting network knowledge with attention mechanism for bearing fault diagnosis," *Applied Soft Computing*, Vol. 97, p. 106829, Dec. 2020, <https://doi.org/10.1016/j.asoc.2020.106829>
- [24] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, Vol. 17, No. 2, p. 425, Feb. 2017, <https://doi.org/10.3390/s17020425>
- [25] D. Li et al., "Fault diagnosis of rotating machinery based on dual convolutional-capsule network (DC-CN)," *Measurement*, Vol. 187, p. 110258, Jan. 2022, <https://doi.org/10.1016/j.measurement.2021.110258>
- [26] R. Liu, F. Wang, B. Yang, and S. J. Qin, "Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions," *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 6, pp. 3797–3806, Jun. 2020, <https://doi.org/10.1109/tii.2019.2941868>
- [27] Y. Chang, J. Chen, C. Qu, and T. Pan, "Intelligent fault diagnosis of wind turbines via a deep learning network using parallel convolution layers with multi-scale kernels," *Renewable Energy*, Vol. 153, pp. 205–213, Jun. 2020, <https://doi.org/10.1016/j.renene.2020.02.004>
- [28] H. Liao, W. Cai, F. Cheng, S. Dubey, and P. B. Rajesh, "An online data-driven fault diagnosis method for air handling units by rule and convolutional neural networks," *Sensors*, Vol. 21, No. 13, p. 4358, Jun. 2021, <https://doi.org/10.3390/s21134358>
- [29] Q. Guo, X. Zhang, J. Li, and G. Li, "Fault diagnosis of modular multilevel converter based on adaptive chirp mode decomposition and temporal convolutional network," *Engineering Applications of Artificial Intelligence*, Vol. 107, p. 104544, Jan. 2022, <https://doi.org/10.1016/j.engappai.2021.104544>
- [30] S. Ai, J. Song, and G. Cai, "A real-time fault diagnosis method for hypersonic air vehicle with sensor fault based on the auto temporal convolutional network," *Aerospace Science and Technology*, Vol. 119, p. 107220, Dec. 2021, <https://doi.org/10.1016/j.ast.2021.107220>
- [31] J. Feng et al., "A self-improving fault diagnosis method for intershaft bearings with missing training samples," *Mechanical Systems and Signal Processing*, Vol. 225, p. 112260, Feb. 2025, <https://doi.org/10.1016/j.ymssp.2024.112260>
- [32] Y.-J. Huang, A.-H. Liao, D.-Y. Hu, W. Shi, and S.-B. Zheng, "Multi-scale convolutional network with channel attention mechanism for rolling bearing fault diagnosis," *Measurement*, Vol. 203, p. 111935, Nov. 2022, <https://doi.org/10.1016/j.measurement.2022.111935>
- [33] G. Qu, M. Song, G. Xin, Z. Shang, and L. Sun, "Time-convolutional network with joint time-frequency domain loss based on arithmetic optimization algorithm for dynamic response reconstruction," *Engineering Structures*, Vol. 321, p. 119001, Dec. 2024, <https://doi.org/10.1016/j.engstruct.2024.119001>
- [34] Y. Tian, Y. Zhang, and H. Zhang, "Recent advances in stochastic gradient descent in deep learning," *Mathematics*, Vol. 11, No. 3, p. 682, Jan. 2023, <https://doi.org/10.3390/math11030682>



**Shuangqiang Luo** earned his master's degree from Jilin University in 2016. He is currently affiliated with Anyang Cigarette Factory, China Tobacco Henan Industrial Co., Ltd. His research focuses on tobacco primary processing technology control, equipment automation systems, and machine condition monitoring, with an emphasis on optimizing production efficiency and quality assurance in cigarette manufacturing processes.



**Xiaoyun Gong** received the Ph.D. degree from Zhengzhou University, Zhengzhou, China, in 2013. She is currently a professor in mechanical engineering at the Zhengzhou University of Light Industry, Zhengzhou. Her current research interests include vibration signal processing and rotating machinery fault diagnosis.



**Wenliao Du** received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2013. He is currently a Professor with the Zhengzhou University of Light Industry, Zhengzhou, China. His research interests include signal processing and rotating machinery fault diagnosis.



**Liangwen Wang** is a Second-Level Professor and Doctoral Supervisor at Zhengzhou University of Light Industry. His research focuses on intelligent robotics, mechanical equipment design, and mechanical systems analysis. He has led 7 National Natural Science Foundation projects and 10+ provincial initiatives, publishing over 200 papers. With 80 invention patents and 12 provincial/ministerial science awards, he was honored as National Outstanding Teacher and Henan Province's Outstanding Craftsman Talent.



**Kunpeng Feng** received the master's degree from Zhengzhou University of Light Industry in 2024. His current research interests include vibration signal processing and rotating machinery fault diagnosis. In recent years, he has published more than 4 academic papers at home and abroad, including 2 papers included in SCI.



**Yahong Qian** received his master's degree in mechanical engineering from Jilin University in Changchun, Jilin Province, China, in 2016. Now he works at Anyang Cigarette Factory of Henan Tobacco Industry Co., Ltd. His current research interests include tobacco machinery, automatic control, and fault diagnosis.