

Article

An Integrated IEWT and CNN–Transformer Deep Architecture for Intelligent Fault Diagnosis of Bogie Axle-Box Bearings

Xiaoping Ding ^{1,2,3}, Zhongqi Li ^{1,2,*}, Minghui Tang ³, Xiaoxu Shen ³ and Liang Zhou ^{1,2}

¹ School of Electrical and Automation Engineering, East China Jiaotong University, Nanchang 330013, China; 2020029081100008@ecjtu.edu.cn (X.D.); zl971125@163.com (L.Z.)

² State Key Laboratory of Performance Monitoring and Protecting of Rail Transit Infrastructure, East China Jiaotong University, Nanchang 330013, China

³ Zhejiang Xingfu Rail Transit Operation Management Co., Ltd., Wenzhou 710064, China; tangminghui2000@163.com (M.T.); shenxiaoxubjtu@163.com (X.S.)

* Correspondence: lzq0828@163.com

Abstract

To address the strong nonstationarity and complex multi-source interference in vibration signals of bogie axle-box bearings, a fault diagnosis method combining Improved Empirical Wavelet Transform (IEWT) and a Convolutional Neural Network (CNN)–Transformer model is proposed. First, the vibration signals are decomposed using the IEWT method, where dynamic frequency-band division adaptively determines the decomposition bands. This yields multiple intrinsic mode functions, and key modes containing fault features are selected based on information entropy. Next, the selected key modes are fused and transformed into polar coordinate projection maps, further enhancing the distinctiveness of fault data features. Finally, CNN is employed to extract local features from the vibration signals, while the Transformer captures long-range dependencies through the self-attention mechanism, significantly improving feature modeling for complex signals. To validate the fault diagnosis performance of the IEWT and CNN–Transformer model, vibration signals from bogie axle-box bearings in urban railways are analyzed. Analysis of the experimental data suggests that the adaptive decomposition of bearing signals using IEWT effectively overcomes the fixed band boundary limitations of traditional EWT, enhancing the precision of signal feature extraction. The integration of polar coordinate projection maps more accurately illustrates frequency variations and amplitude differences in the signals, fully capturing their nonstationary characteristics. Among the five fault categories of bogie axle-box bearings, the proposed method achieves an accuracy of 99.46%, a recall rate of 99.52%, and an F1-score of 0.995, significantly outperforming five classic comparison methods. This demonstrates that the combined strengths of CNN and Transformer yield higher classification accuracy and better robustness in handling complex fault patterns, effectively solving the fault diagnosis challenges for bogie axle-box bearings.



Academic Editors: Yunfeng Yan,
Xian-Bo Wang and Stefano Scanzio

Received: 31 December 2025

Revised: 3 February 2026

Accepted: 5 February 2026

Published: 13 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: bogie axle-box bearings; fault diagnosis; improved empirical wavelet transform; convolutional neural network; transformer; polar coordinate projection

1. Introduction

Bogie axle-box bearings, as core components of the transmission system in rail transit vehicles, play a critical role in power transmission during train operation [1]. However, due to prolonged high-load, high-speed operation, and complex working conditions, these bearings are prone to wear, fatigue, and even failure. Such issues not only reduce

the service life of trains but also compromise operational safety and potentially endanger human lives. Therefore, conducting health monitoring and fault diagnosis research on bogie axle-box bearings to comprehensively understand their operational status and provide timely warnings of potential failures is a key measure to ensure the safe operation of rail vehicles and reduce the risk of accidents [2–4].

In recent years, with advancements in sensor measurement accuracy and computational power, equipment health monitoring has gradually evolved toward intelligent systems [5]. Previously, health monitoring research primarily relied on mathematical models and expert knowledge. For instance, reference [6] developed a state observer for wind turbines combined with adaptive threshold decision-making, and semi-physical simulation results demonstrated its effectiveness. However, that approach is highly dependent on precise mathematical models of the system. Discrepancies often exist between the model and the actual system's dynamics due to internal disturbances and external environmental influences. Similarly, reference [7] proposed a knowledge-based fault diagnosis method that used expert and operational experience to classify fault data. However, transforming expert knowledge into structured knowledge graphs is a complex and time-intensive process, necessitating a shift toward intelligent decision-making and precise management. Deep learning, with its advantages in automatic feature extraction and complex data handling, has become an important tool for equipment fault diagnosis [8]. Researchers worldwide have conducted extensive work in intelligent fault diagnosis for axle-box bearings, achieving valuable results. These efforts primarily leverage deep learning methods such as Convolutional Neural Networks (CNNs) [9], ResNet [10], Gated Recurrent Units [11], and Transformers [12] for feature extraction and efficient fault classification. For example, reference [13] introduced a fault diagnosis method combining Wavelet Transform (WT) and an optimized CNN, effectively improving the utilization of critical fault features. However, WT struggles with the decomposition and feature extraction of strongly nonlinear signals, and its accuracy depends on the selection of the wavelet basis function. Reference [14] proposed a method integrating Empirical Mode Decomposition (EMD) and Singular Value Decomposition, generating singular value feature matrices for fault classification using a CNN. While this method outperformed direct raw data input, EMD suffers from mode-mixing and under-enveloping issues during decomposition. Reference [15] combined a CNN with Long Short-Term Memory (LSTM) for axle-box fault diagnosis, considering various gears under different conditions. While effective in fault classification, LSTM's weak parallelization and dependency on sequential computations result in slower training and inference speeds. Finally, reference [16] presented a fault diagnosis method using an improved sparrow algorithm to optimize Support Vector Machines (SVMs). While this approach overcomes SVMs' tendency to fall into local optima, it relies on manual feature extraction, requiring significant expertise and complicating feature selection. The overall process remains cumbersome, limiting its efficiency in practical applications.

To overcome the limitations of the aforementioned methods while leveraging their strengths, this study integrates and optimizes signal processing, feature extraction, and model design to further enhance diagnostic performance under complex working conditions. Empirical Wavelet Transform (EWT) has emerged as a popular method in vibration signal analysis due to its adaptive filter bank construction, effectively addressing some inherent drawbacks of EMD. Reference [17] employed EWT to decompose partial discharge signals and calculated multiscale quantum entropy for each component, achieving more accurate spectral boundary determination and effectively avoiding mode-mixing issues. Reference [18] introduced a fault diagnosis method based on EWT, fuzzy entropy, and SVMs, where EWT decomposed vibration signals into multiple components, and the fuzzy entropy of AM-FM components served as input to the SVM model. The results

demonstrated that EWT outperformed EMD in signal decomposition. Compared to EMD, EWT provides finer decomposition of nonstationary signals and mitigates noise interference. However, traditional EWT often relies on fixed frequency-band boundaries, limiting its adaptability. Reference [19] proposed a fault diagnosis method combining improved EWT with fast spectral correlation, using maximum envelope processing for frequency-band division. Reference [20] addressed fan blade damage by preprocessing engine vibration signals from various flights, performing EWT decomposition, and extracting energy entropy and multiscale entropy with Hamming window functions. The experimental results indicated that the fitted data exhibited more significant entropy changes. Regarding feature extraction, reference [21] analyzed the development of a CNN in intelligent transportation. Due to its widespread application in image data, a CNN can automatically learn and extract fault features, significantly reducing the manual effort required for feature extraction. It also captures subtle fault patterns more effectively than other algorithms, making its extracted features more comprehensive and valuable. Transformers, known for their efficient fault detection and feature representation in images, excel at capturing global information. They address the gradient vanishing and weak parallelization issues of traditional LSTMs when handling long sequences. Reference [22] proposed a multivariable temporal labeling Transformer-based fault diagnosis method, incorporating multi-head attention mechanisms and residual connections to enhance training stability and feature extraction capabilities.

Inspired by the aforementioned studies and aiming to overcome the limitations of existing fault diagnosis methods, this paper proposes an efficient and intelligent diagnosis framework for bogie axle-box bearings by integrating Improved Empirical Wavelet Transform (IEWT) with a CNN–Transformer architecture. Existing methods usually suffer from three major issues: (i) fixed frequency-band division in traditional EWT lacks adaptability to nonstationary vibration signals; (ii) single-structure networks such as CNNs are limited in modeling long-range dependencies among multiscale fault features; and (iii) raw time-domain representations are insufficient to fully characterize complex spectral–temporal patterns under strong noise and interference.

To address these issues, the IEWT module is first introduced to adaptively decompose vibration signals by dynamically determining frequency-band boundaries based on energy distribution, thereby overcoming the rigidity of conventional EWT and enabling better representation of nonstationary characteristics. Information entropy is further employed to select informative modal components and suppress noise-dominated bands, ensuring that reconstructed signals preserve discriminative fault information. Subsequently, the reconstructed signals are transformed into polar coordinate projection maps, which encode frequency variation and amplitude distribution jointly in angular and radial dimensions, enhancing the visibility of subtle fault patterns in a two-dimensional form.

On this basis, a hybrid CNN–Transformer network is designed. The CNN focuses on extracting local spatial structures and texture-like fault signatures from the projection maps, while the Transformer employs a self-attention mechanism to capture global dependencies across different frequency bands and time segments. This complementary design enables the model to simultaneously learn fine-grained local features and long-range contextual relationships, which cannot be effectively achieved by CNN-based models alone.

The main contributions of this paper can be summarized as follows and the nomenclature is summarized in Table 1.

1. An adaptive IEWT-based decomposition strategy is proposed, which introduces energy-driven dynamic frequency partitioning to overcome the fixed-band limitation of traditional EWT for nonstationary bearing signals.

2. A polar coordinate projection representation is constructed to transform one-dimensional vibration signals into informative two-dimensional maps, enhancing the expression of spectral–temporal fault characteristics.
3. A CNN–Transformer hybrid architecture is developed to jointly model local spatial patterns and global feature dependencies, improving discriminability under complex operating conditions.
4. Extensive experiments under noise and interference conditions verify that the proposed IEWT–CNN–Transformer framework achieves superior accuracy, stability, and robustness compared with several classical methods.

Table 1. Nomenclature.

Symbol	Description
CNNs	Convolutional Neural Networks
WT	Wavelet Transform
EMD	Empirical Mode Decomposition
LSTM	Long Short-Term Memory
SVMs	Support Vector Machines
EWT	Empirical Wavelet Transform
IEWT	Improved Empirical Wavelet Transform

2. Materials and Methods

2.1. Empirical Wavelet Transform

As an adaptive time–frequency analysis technique, EWT overcomes the limitations of traditional wavelet transforms that rely on predefined and fixed basis functions [23]. By constructing a data-driven filter bank according to the local variations in the signal spectrum, EWT decomposes the input signal into a series of sub-signals with distinct frequency-band characteristics. Each sub-signal reflects the local behavior of the vibration signal within a specific spectral range, which makes EWT particularly suitable for analyzing nonlinear and nonstationary signals.

In the context of bogie axle-box bearing fault diagnosis, vibration responses are typically nonstationary and affected by varying operating conditions and noise. Directly feeding such raw signals into deep models often leads to feature confusion and performance degradation. Therefore, EWT is introduced in this work as a front-end decomposition tool to separate mixed spectral components and highlight fault-related frequency structures before deep feature learning. However, conventional EWT still depends on fixed spectral segmentation strategies, which may fail to adapt to complex and time-varying signal distributions. This limitation motivates the development of the proposed Improved EWT (IEWT), which further enhances the adaptability of frequency-band partitioning and provides more discriminative modal components for subsequent CNN–Transformer-based fault identification.

Let the vibration signal to be decomposed be $s(t)$, with its corresponding Fourier transform expressed in Equation (1). By analyzing the amplitude spectrum of $\hat{s}(t)$, it is possible to identify significant changes in the spectral components, such as energy spikes or local extrema, denoted as $0 = \omega_0 < \omega_1 < \dots < \omega_N$. These boundary points are used to divide the entire spectrum into N sub-bands, where each sub-band corresponds to distinct faults or critical fault characteristics within the vibration signal.

$$\hat{s}(\omega) = \int_{-\infty}^{+\infty} s(t)e^{-j\omega t} dt \quad (1)$$

After splitting the spectrum, the next step is to automatically create a set of adaptive filters $H_k(\omega)$ based on the local features of the signal spectrum. This set consists of a low-pass filter and several band-pass filters. The design principle aims to smoothly and accurately extract signal components within different frequency bands while avoiding spectral leakage and signal distortion. The low-pass filter $H_0(\omega)$ targets the first band, extracting the lowest frequency components and passing them entirely within the cutoff frequency ω_1 . For subsequent bands $k = 1, 2, \dots, N - 1$, the band-pass filters $H_k(\omega)$ are designed based on two distinct frequency intervals. Within the interval (ω_{k-1}, ω_k) , a sine function is used to smoothly raise the filter response from 0 to 1, while within the interval (ω_k, ω_{k+1}) , a cosine function is applied to smoothly lower the response from 1 back to 0. The detailed derivation of the formula is as follows:

$$H_0(\omega) = \begin{cases} 1, & |\omega| \leq \omega_1 \\ \cos\left(\frac{\pi}{2} \frac{|\omega| - \omega_1}{\omega_2 - \omega_1}\right), & \omega_1 < |\omega| \leq \omega_2 \\ 0, & |\omega| > \omega_2 \end{cases} \quad (2)$$

$$H_k(\omega) = \begin{cases} \sin\left(\frac{\pi}{2} \frac{|\omega| - \omega_{k-1}}{\omega_k - \omega_{k-1}}\right), & \omega_{k-1} < |\omega| \leq \omega_k \\ \cos\left(\frac{\pi}{2} \frac{|\omega| - \omega_k}{\omega_{k+1} - \omega_k}\right), & \omega_k < |\omega| \leq \omega_{k+1} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Using the constructed filter set, the vibration signal is adaptively decomposed. For the k th frequency band, $H_k(\omega)$ is multiplied with the signal's spectrum, and an inverse Fourier transform is performed to obtain the corresponding sub-signal $q_k(t)$. The detailed derivation of the formula is as follows:

$$\begin{aligned} q_k(t) &= \mathcal{F}^{-1}\{\hat{q}(\omega)H_k(\omega)\} \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{q}(\omega)H_k(\omega)e^{j\omega t}d\omega \end{aligned} \quad (4)$$

where $\hat{q}(\omega)$ denotes the Fourier transform of the signal, while $\mathcal{F}^{-1}$ indicates the operation of applying the inverse Fourier transform.

2.2. CNN

CNNs have shown outstanding performance in image analysis and pattern recognition due to their ability to exploit local receptive fields, weight sharing, and hierarchical feature learning. These mechanisms enable CNNs to automatically capture multi-level local structures from data, effectively alleviating the reliance on handcrafted feature extraction [24,25].

In bogie axle-box bearing diagnosis, vibration signals are highly nonlinear, nonstationary, and often contaminated by noise, making direct learning from raw time-series challenging. To bridge this gap, the reconstructed signals are transformed into polar coordinate projection maps, which convert temporal and spectral variations into spatial patterns. This representation allows CNN to fully exploit their strength in visual feature perception, enabling the extraction of localized time–frequency textures and discriminative fault-related structures. In the proposed framework, CNN serves as the front-end feature extractor, focusing on capturing fine-grained local patterns, which provides informative and compact representations for subsequent global modeling by the Transformer module.

For the input polar coordinate projection map $I \in R^{H \times W}$, the convolution layer typically extracts local features from the image using multiple 2D convolution kernels.

Assuming a specific convolution kernel is $K \in R^{k_h \times k_w}$, performing a 2D convolution between the kernel and the input image yields:

$$D(m, n) = (W * K)(m, n) = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} I(m+q, n+e) \cdot K(q, e) + b \quad (5)$$

where $D(m, n)$ refers to the result of applying the convolution operation at position (m, n) , and b denotes the corresponding bias term.

Once the convolution operation is completed, a nonlinear activation function is typically employed to enhance the model's ability to handle nonlinear relationships. In this paper, the activation function chosen is the Rectified Linear Unit, which can be mathematically expressed as:

$$f(z) = \max(0, z) \quad (6)$$

where $f(z)$ indicates the result obtained after applying the activation function.

The pooling layer serves to decrease the dimensionality of the feature maps generated by the convolutional layer. This helps to lower the data dimensions and reduce computation. In this paper, the pooling layer uses max pooling S . Let the pooling window size be specified as $p \times p$; then, the pooling operation can be expressed as:

$$L(b, s) = \max_{(b,s) \in \mathcal{W}_{bs}} S(b, s) \quad (7)$$

where $\mathcal{W}_{bs}$ represents the pooling window region corresponding to the output position (b, s) .

Once the convolution and pooling processes are completed, the feature map is usually converted into a single-dimensional vector and forwarded to the fully connected layer for additional feature processing. In the output layer, the softmax function is used to output the probability for each class, which can be expressed as follows:

$$\hat{y}_c = \frac{\exp(z_c)}{\sum_{c'=1}^C \exp(z_{c'})} \quad (8)$$

where z_c represents the c th value from the fully connected layer's output, C is the total number of categories, and $\hat{y}_c$ indicates the predicted likelihood for each category.

2.3. Transformer

The Transformer model, relying on the self-attention mechanism, can effectively address challenges such as information loss, gradient vanishing, and low computational efficiency often encountered by traditional recurrent networks when processing long sequences. Through multi-head attention, the Transformer dynamically captures relationships between different parts of the input data, enabling precise modeling of global features. In this study, because the vibration signals of bogie axle-box bearings exhibit multiscale, nonstationary, and noise-interfered characteristics, CNN alone—focusing on local feature extraction—cannot fully capture the global dependencies. By incorporating the Transformer into the CNN framework, the model can extract local time–frequency features while simultaneously attending to global correlations across different frequency bands and time segments, thereby significantly enhancing the representation of fault features and improving classification performance.

The Transformer model successfully identifies the long-range relationships within the input data using the self-attention process and multi-head attention structure. Combined

with mechanisms such as positional encoding, feedforward networks, and residual connections, it constructs an efficient and flexible feature representation structure. The components are described in the following.

The self-attention mechanism dynamically calculates the correlations between each position based on the input sequence features. Let the input matrix be L . Through the query vector G , key vector H and value vector J are obtained, which can be represented as follows:

$$G = LW^Q, \quad H = LW^K, \quad J = LW^V \quad (9)$$

$$\text{Attention}(G, H, J) = \text{softmax}\left(\frac{GH^\top}{\sqrt{d_k}}\right)J \quad (10)$$

where W is weight; Equation (10) uses the scaled dot product to stabilize the gradient, and the softmax function generates the weight distribution. d_k represents the feature dimension.

Transformer uses a multi-head attention mechanism to capture information from different subspaces. Multi-head attention maps the input to different subspaces, calculates the attention for each, and then concatenates them. This can be expressed as:

$$H_i = \text{Attention}(GW_i^G, HW_i^H, JW_i^J) \quad (11)$$

$$\text{MultiHead}(G, H, J) = \text{Concat}(H_1, \dots, H_h)W^O \quad (12)$$

where W_i^G, W_i^H, W_i^J represents the weight matrix associated with each attention head, and W^O is the matrix used for mapping the output, Concat is the concatenation function, and MultiHead is the multi-head attention output.

3. Experimental Results and Analysis

3.1. Method for Bogie Axle-Box Bearings Based on IEWT and CNN-Transformer

Figure 1 shows the proposed fault diagnosis method for the bogie axle-box bearings, which combines IEWT and CNN-Transformer. Initially, the vibration signals of the bogie axle-box bearings were collected and adaptively decomposed using IEWT to obtain several modal components. Subsequently, the reconstructed signals were transformed into polar coordinate projection maps, where both angles and radii simultaneously expressed multidimensional fault information.

Finally, all polar coordinate projection maps were divided into training, validation, and test sets in a ratio of 6:2:2. The IEWT and CNN-Transformer fault diagnosis method was used to train on the polar coordinate projection map dataset. During training, accuracy and loss curves were generated across iterations. The optimal model parameters were determined and saved using the validation set. The test set was then employed to evaluate the performance of the optimal model, generating a confusion matrix and t-SNE clustering visualization results.

In actual train operations, the scarcity of real fault samples due to safety constraints and maintenance cycles poses challenges to the generalization capability of deep learning models. To address this, the proposed method enhanced model adaptability and robustness from three aspects: feature, structure, and training. First, the improved IEWT was introduced for adaptive signal decomposition, enabling the extraction of key features under different operating conditions within a unified frequency-band structure, thereby improving feature stability and transferability at the source. Second, the model integrated the local feature extraction capability of a CNN with the global attention mechanism of a Transformer to achieve efficient multiscale feature fusion and effectively mitigate overfit-

ting. Finally, dropout and L2 regularization were applied during training to enhance the model's robustness against small-sample limitations and noise disturbances.

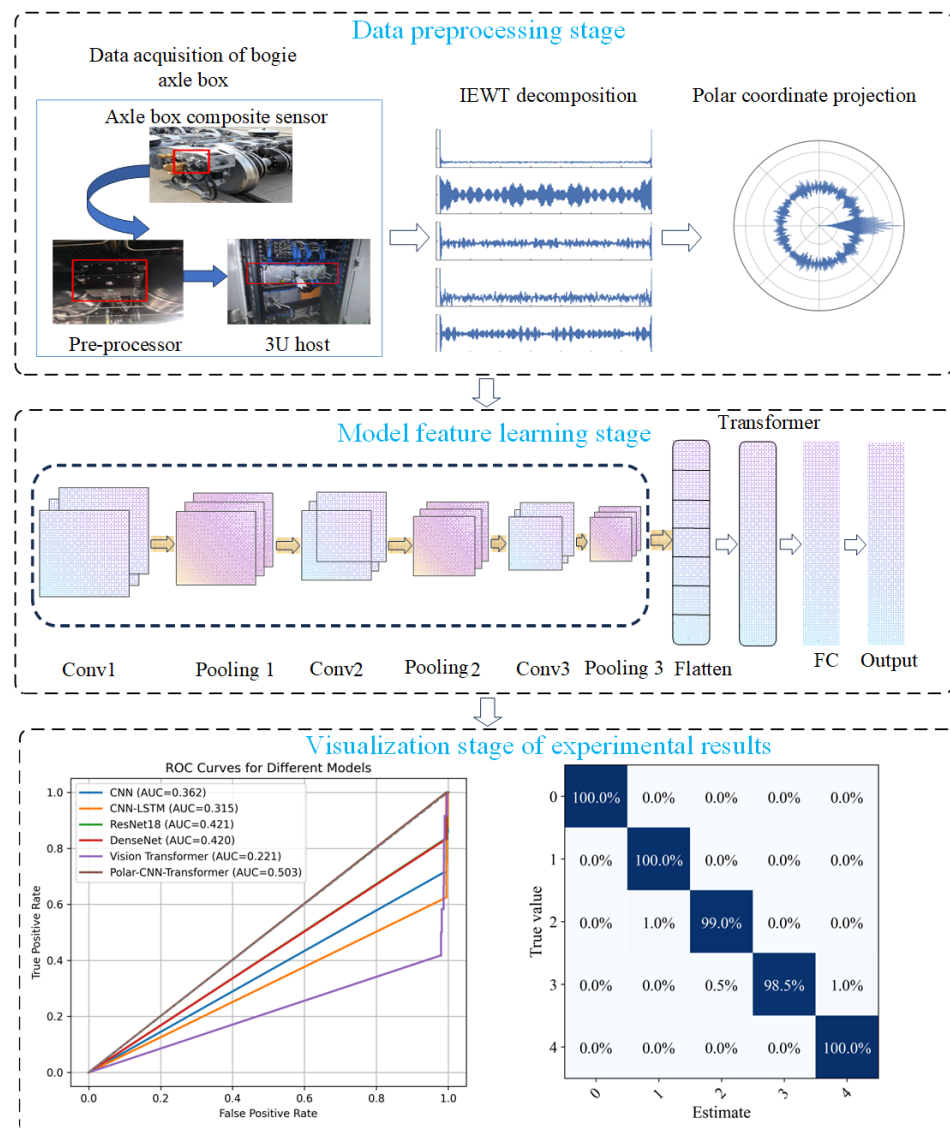


Figure 1. Fault diagnosis method based on IEWT and CNN-transformer.

3.1.1. The Main Steps of IEWT

To overcome the issues of preset frequency-band boundaries and noise sensitivity in traditional EWT decomposition, this paper introduces the IEWT method with a dynamic frequency-band division strategy based on energy distribution. The basic process includes the following main steps:

Step 1: Perform a Fast Fourier Transform on the vibration signal of the bogie axle-box bearing to compute the corresponding amplitude spectrum and frequency information.

Step 2: Process the amplitude spectrum using min-max normalization, allowing different amplitude spectra to be compared with each other.

Step 3: The processed amplitude spectrum is clustered using the K-means algorithm with the number of clusters set to $K = 5$, dividing the spectral energy into five adaptive frequency bands. As shown in Figure 2, t-SNE visualization indicates that the five clusters are well separated in the two-dimensional space, suggesting clear distinctions between frequency bands. This choice is further supported by the clustering evaluation metrics listed in Table 2: when $K = 5$, the silhouette coefficient reaches its maximum value of 0.8, indicating

the tightest intra-cluster cohesion, while the within-cluster sum of squares (WCSS) attains its minimum of 120.5, demonstrating the most concentrated spectral energy distribution within each cluster. Together, these results confirm that $K = 5$ provides an optimal balance between cluster compactness and separation for adaptive frequency-band division.

Table 2. Clustering evaluation metrics for different K values.

K Value	Silhouette Coefficient	WCSS
3	0.65	150.2
4	0.72	135.4
5	0.80	120.5
6	0.75	125.3
7	0.69	130.1

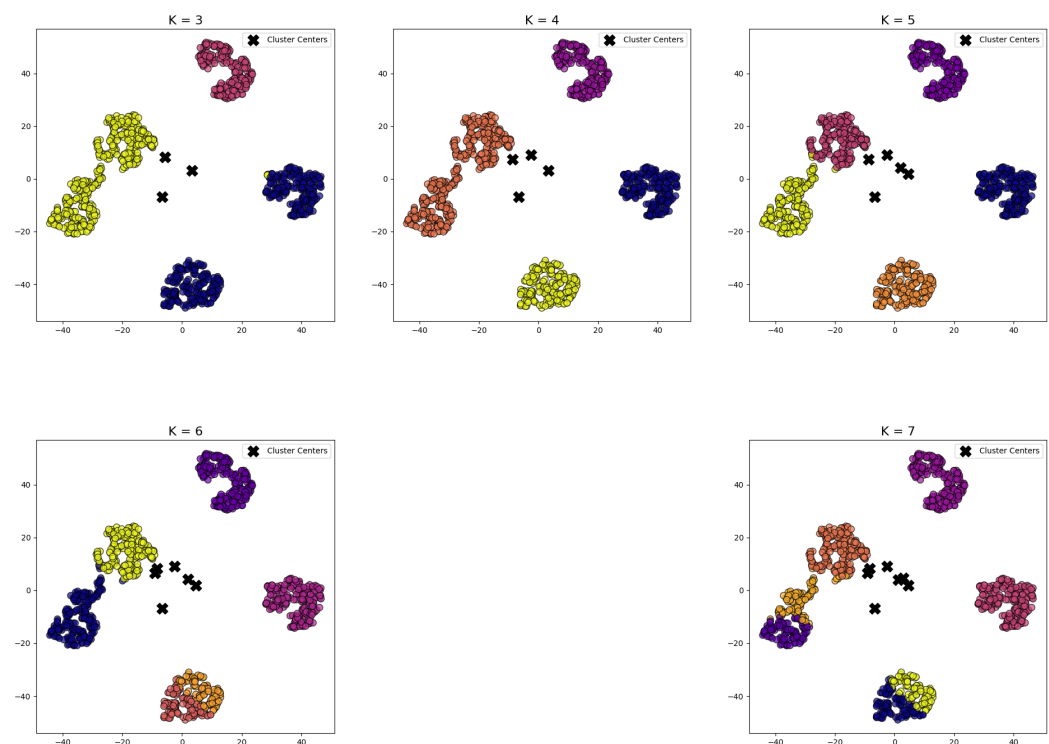


Figure 2. Clustering results under different K values.

Step 4: Construct corresponding band-pass filters for each frequency band based on the dynamically determined frequency-band boundaries from steps 1–3.

Step 5: Use the constructed filters to adaptively decompose the bogie axle-box bearing vibration data, obtaining modal components for each frequency band.

Step 6: Select the primary fault modal components using information entropy [26], then reconstruct the signal using these primary modes. The information entropy calculation is as follows:

$$H = - \sum_{i=1}^n p_i \log p_i \quad (13)$$

where H refers to the result of the information entropy, and p_i is the probability of the i th event.

3.1.2. The Main Steps of Polar Coordinate Projection

Polar coordinate projection maps the one-dimensional vibration signal data $s(t)$ of the bogie axle-box bearing to a two-dimensional polar coordinate space, which can intuitively reveal the periodicity, patterns, and abnormal characteristics of the signal [27]. The basic process includes the following main steps:

Step 1: Angle Mapping: Map the vibration signal sample points to the angle $\theta = 2\pi t/T$, where T represents the sample length (1024).

Step 2: Radius Mapping: Map the vibration signal to the radius $r = |s(t)|$.

Step 3: Plot Polar Coordinate Graph: Plot (r, θ) in the polar coordinate system, obtaining the polar coordinate projection map for each sample data.

In the parameter setting, signal resolution, amplitude scale, and sample comparability were comprehensively considered: the sample length was set to 1024 to balance resolution and computational efficiency; the radius was normalized to eliminate amplitude differences; and the polar origin and radial scale were kept consistent to ensure feature comparability.

3.1.3. CNN–Transformer

The input data for the CNN–Transformer diagnostic model are the polar coordinate projection map, a grayscale image (single-channel image) of size 32×32 . The CNN is used to automatically learn the local features in the polar coordinate projection map, while the Transformer captures the global dependencies. The combination of both enhances the fault diagnosis model's accuracy and robustness in multidimensional feature extraction and fault category classification tasks. The specific fault diagnosis structure is shown in Figure 3, and the network structure parameters are listed in Table 3.

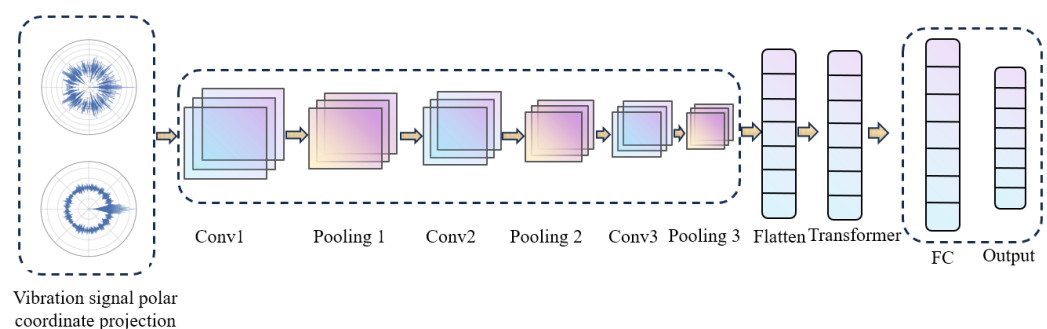


Figure 3. Fault diagnosis architecture of CNN–Transformer.

The size of the vibration signal's polar coordinate projection map was selected as 32×32 . This size can retain key information while allowing the model to train quickly. The CNN network consisted of three layers of convolution and pooling operations. The number of convolution kernels in each layer was 16, 32, and 64, respectively. The increasing number of kernels layer by layer allowed the model to extract and learn deeper, more abstract fault features. After the convolution and pooling operations, the data's spatial size was reduced to 4×4 .

In the Transformer network, the embedding dimension was set to 128, and four multi-head attention mechanisms were used, allowing the fault diagnosis model to perform well in terms of both expressive power and computational efficiency. To avoid introducing too many parameters, the encoder was set to two layers, with the feedforward network dimension and dropout set to 256 and 0.1, respectively, based on empirical values. This helped enhance the model's nonlinearity and prevents overfitting.

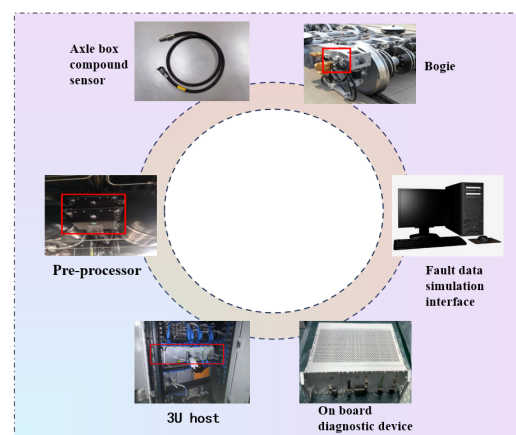
Table 3. Network structure parameters of CNN–Transformer.

Network Module	Main Model Parameters	Output Size
Input Data	-	(batch, 1, 32, 32)
Convolution Module 1	Conv2d(1, 16, 3×3 , stride = 1, padding = 1) β MaxPool2d(kernel _{size} = 2)	(batch, 16, 16, 16)
Convolution Module 2	Conv2d(16, 32, 3×3 , stride = 1, padding = 1) β MaxPool2d(kernel _{size} = 2)	(batch, 32, 8, 8)
Convolution Module 3	Conv2d(32, 64, 3×3 , stride = 1, padding = 1) β MaxPool2d(kernel _{size} = 2)	(batch, 64, 4, 4)
Flatten Layer	Linear(64, 128)	(16, batch, 128)
Transformer Module	Encoder layers: 2, multi-head: 4 (dimension per head: 32), feedforward network dimension: 256, dropout: 0.1	(16, batch, 128)
Global Average Pooling	-	(batch, 128)
FC	10	(batch, 10)
Output	5	(batch, 5)

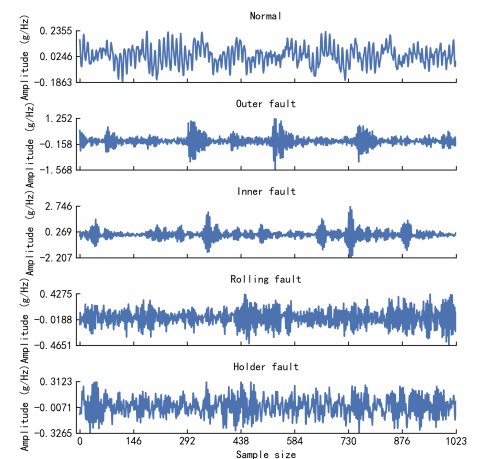
3.2. Case Analysis: Fault Diagnosis of Bogie Axle-Box Bearings

3.2.1. Description

The platform primarily consisted of the running components, bogie, axle-box composite sensors (e.g., SKF T-300 vibration sensor), front processor, 3U host (e.g., ComExpress 3U), on-board diagnostic device (e.g., CDS-3000 diagnostic system by Sifang Electric), and a fault data simulation interface, as illustrated in Figure 4a. The dataset comprised five fault categories: normal, outer-race fault, inner-race fault, rolling-element fault, and holder fault. Data were collected with a sampling frequency of 12 kHz, under AW0 load conditions, and with a motor speed of 3500 RPM. The details of the dataset are summarized in Table 4. The one-dimensional time-series fault data for the bogie axle-box bearing is shown in Figure 4b, with each fault signal consisting of 1024 sampled points, without overlap between samples. In total, 5000 samples were collected, divided into 3000 for training, 1000 for validation, and 1000 for testing.



(a) Data acquisition platform for axle-box bearings



(b) Vibration signals of axle-box bearings

Figure 4. Experimental setup and vibration characteristics of bogie axle-box bearings.

Table 4. Experimental bearing data.

Fault Types	Motor Speed/RPM	Load	Length	Identify Tags
Normal	3500	AW0	1024	0
Outer Fault	3500	AW0	1024	1
Inner Fault	3500	AW0	1024	2
Rolling Fault	3500	AW0	1024	3
Holder Fault	3500	AW0	1024	4

3.2.2. IEWT Adaptive Decomposition

Based on the decomposition steps of the IEWT in Section 3.1.1, the vibration signals of the bogie axle-box bearing were adaptively decomposed. Figure 5 shows the IEWT decomposition results for outer-race fault and cage fault cases. As seen in Figure 5, the modes after IEWT decomposition exhibited good frequency-band separation. Through adaptive frequency-band division, the key fault signals were more prominently expressed in specific modal components, while noise and other irrelevant components were separated.

Each mode component was evaluated using information entropy to select the main modes containing the primary fault information. The information entropy results for the modes after IEWT decomposition for the outer race fault and cage fault are shown in Table 5. From Table 5, it can be seen that IMF1, IMF3 and IMF4 for the outer race fault, and IMF1 and IMF4 for the cage fault, were the main fault modal components.

According to the polar coordinate projection procedure described in Section 3.1.2, the reconstructed signals of the outer-race and holder faults were transformed into two-dimensional polar projection maps, as shown in Figure 6. Both fault types exhibited a regular annular distribution in the polar domain. To quantitatively verify the capability of this transformation in highlighting nonstationary features, the time–frequency entropy was calculated according to Equation (13). The entropy values of the outer-race and cage faults decreased by 26.4% and 22.7%, respectively, compared with those in the Cartesian domain, indicating that the energy becomes more concentrated in the angular direction and the nonstationary characteristics are effectively enhanced, as shown in Figure 7. This symmetrical and energy-concentrated pattern facilitates rapid localization of key fault features and clearer visualization of subtle fault modes.

Table 5. Information entropy of each mode.

Outer-Fault Modal Component	Information Entropy/Bit	Holder-Fault Modal Component	Information Entropy/Bit
IMF1	0.4475	IMF1	0.4094
IMF2	2.1283	IMF2	1.7663
IMF3	1.3723	IMF3	2.2059
IMF4	1.4884	IMF4	1.4736
IMF5	2.0693	IMF5	2.5798

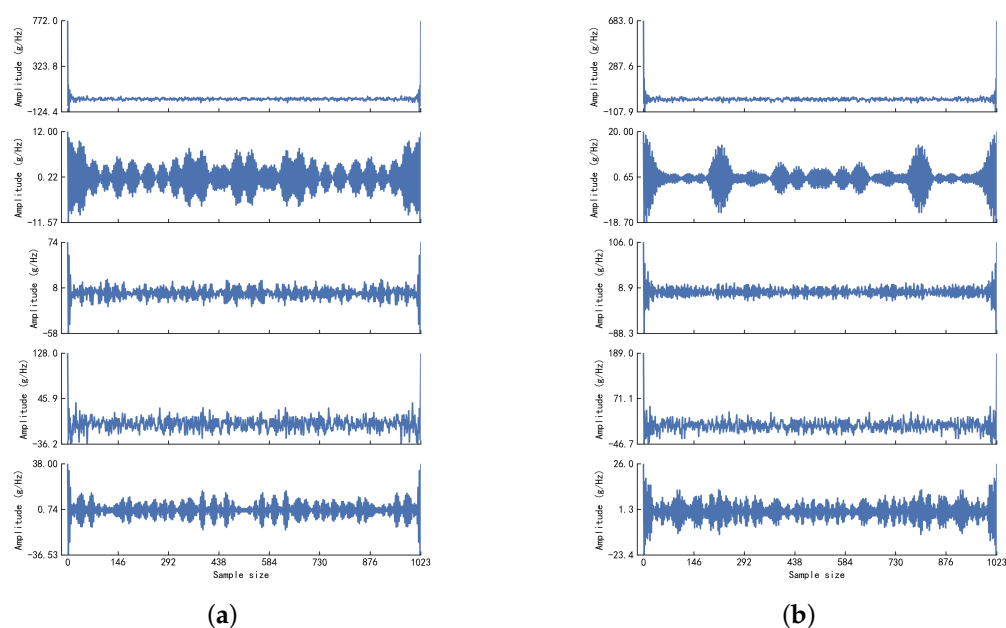


Figure 5. IEWT decomposition results of outer-race fault (a) and cage fault (b).

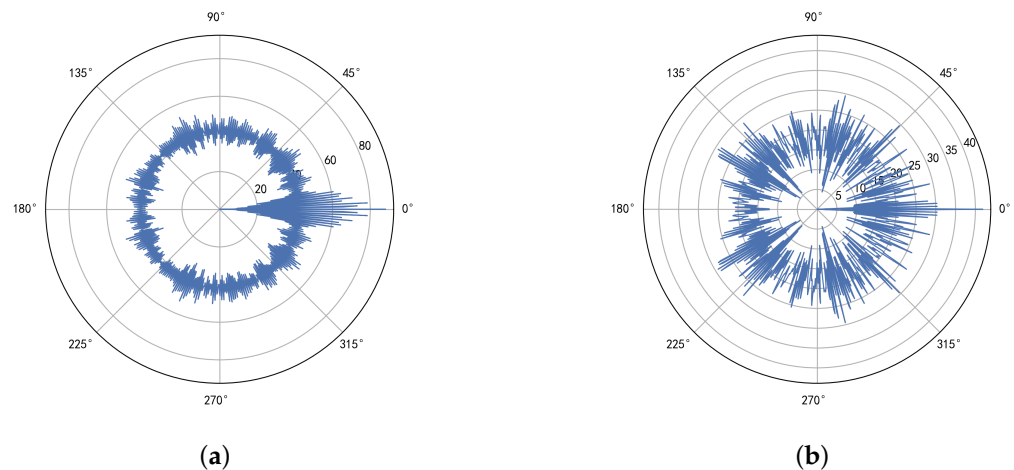


Figure 6. Polar coordinate projection maps of outer-race fault (a) and cage fault (b).

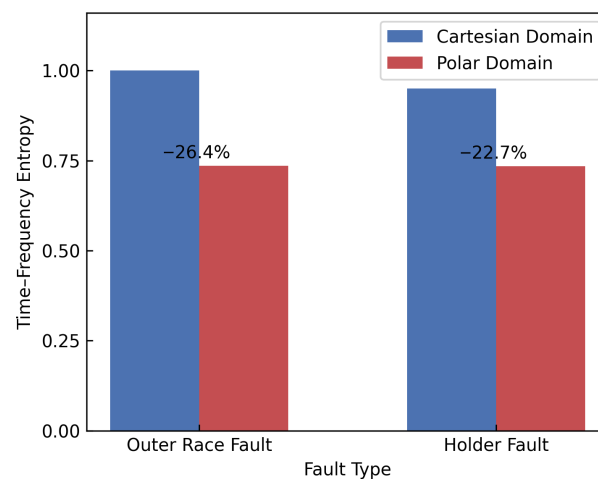


Figure 7. Comparison of time–frequency entropy in Cartesian and polar domains.

3.2.3. Results and Discussion

To validate the performance and advantages of the CNN–Transformer model in fault diagnosis, it was compared with five typical deep learning methods: CNN [24], ResNet18 [10], CNN-LSTM [15], DenseNet [28], and Vision Transformer [29]. These methods were chosen based on their ability to address different aspects of the vibration signal analysis task for steering gear axle-box bearings. CNN and ResNet18 are highly effective for image-based data processing, as their convolutional layers excel at extracting local spatial features. However, they are limited by their inability to capture long-range dependencies in the data, which can hinder their performance in fault diagnosis tasks that require global context understanding. CNN-LSTM, which combines CNN's local feature extraction with LSTM's ability to model temporal dependencies, is well suited for sequential data. Nevertheless, LSTM suffers from gradient vanishing and weak parallelization issues, which can limit its scalability and efficiency. DenseNet, with its dense layer connections, captures rich, detailed features effectively, making it suitable for complex signal data. However, its higher memory usage and computational complexity pose challenges when dealing with large datasets or constrained computational resources. Vision Transformer, utilizing a self-attention mechanism, excels at capturing global information and long-range dependencies, which is advantageous for tasks requiring global context. However, Vision Transformer typically requires large quantities of training data and has weaker local feature extraction capabilities compared to CNN-based models. These methods were selected to highlight

different strengths in processing vibration signals, and the comparison demonstrated how the CNN–Transformer model, by combining both local feature extraction and global dependency modeling, offered a more efficient and scalable solution for fault diagnosis. All experiments were conducted in Python 3.6, using PyTorch 1.10.2, on a system equipped with an NVIDIA GeForce RTX 3050 GPU.

Firstly, the accuracy and loss value variation curves over 100 iterations of training for the proposed method and the five comparison methods are shown in Figure 8. From Figure 8a, it can be seen that Vision Transformer performed weaker than the other comparison methods, indicating that its ability to extract local features and learn fault modes may be insufficient. ResNet18, DenseNet, CNN–LSTM, and CNN–Transformer all performed well in the early stages of training. However, ResNet18 experienced significant fluctuations in accuracy and loss values around the 40th and 80th iterations, which may be related to its deeper network structure and gradient issues during training. CNN–LSTM performed slightly better than the proposed CNN–Transformer in the early stages, but as training progressed, CNN–LSTM became less stable, especially in Figure 8b, where the loss value fluctuation curve of CNN–LSTM exhibited large oscillations in later stages, indicating possible instability during the optimization process. In contrast, CNN–Transformer showed a stable loss value in the later stages, demonstrating stronger stability and suggesting that the Transformer was more suitable for long-term training, maintaining model stability and reducing overfitting and oscillations. This ultimately enhanced the model’s performance in complex tasks.

In contrast, the CNN–Transformer method proposed in this paper combines the advantages of a CNN in local feature extraction with the powerful capability of a Transformer in modeling global dependencies, ensuring the model’s stability and efficiency throughout the training process, and ultimately exhibiting superiority in fault diagnosis tasks.

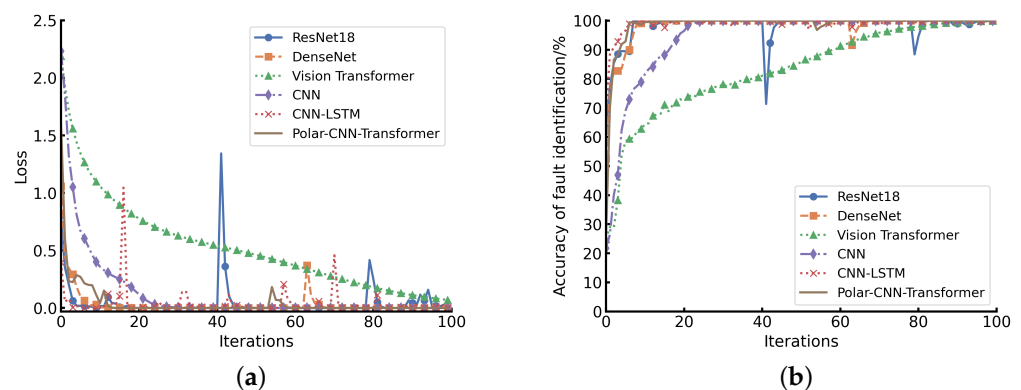


Figure 8. Accuracy (a) and loss (b) variation during the training process of different comparison methods.

To comprehensively assess the performance of different algorithms, Table 6 presents a comparison based on accuracy, recall, and F1-score. From Table 6, it can be observed that ResNet18 and DenseNet performed similarly, with DenseNet slightly outperforming ResNet18, indicating that its dense connections enhanced feature representation. While Vision Transformer had a lower accuracy, it excels at capturing long-term dependencies in sequential data, which is beneficial for complex time-series problems. The CNN demonstrated good performance in identifying individual faults, and combining CNN with LSTM further improved the model’s performance, showing that LSTM compensated for the CNN’s limitations in sequential dependency modeling. Notably, the proposed method, which combines a CNN and a Transformer, achieved the highest accuracy of 99.46%, recall of 99.52%, and an F1-score of 0.995, demonstrating significant advantages in handling complex time-series data and long-term dependency feature

extraction. With NVIDIA GeForce RTX 3050 GPU acceleration, all methods benefited from reduced training and inference times, enabling real-time performance. Specifically, ResNet18 showed fast training times of 200–300 s and inference of 5–10 ms, suitable for simpler fault detection. DenseNet achieved training times of 300–500 s and inference of 10–20 ms. Vision Transformer had longer training times of 400–600 s and inference of 20–50 ms but excelled at capturing intricate sequential patterns. The CNN was lightweight, with training times of 150–250 s and inference of 5–15 ms, ideal for individual fault detection. CNN-LSTM improved sequential modeling with training times of 300–500 s and inference of 15–30 ms. Finally, the proposed CNN–Transformer method, with its superior accuracy and long-term dependency handling, had training times of 500–700 s and inference times of 20–40 ms, making it particularly suitable for real-time fault diagnosis in industrial applications. Overall, the combination of CNN and Transformer, enhanced by GPU acceleration, not only achieved high performance but also provided fast, real-time inference, making it highly effective for the fault diagnosis of the bogie axle-box bearing.

Table 6. Fault classification evaluation metrics of different methods.

Method	Accuracy/%	Recall/%	F1-Score	Training Time (s)	Inference Time (ms)	Parameters (M)
ResNet18	99.01	99.06	0.990	200–300	5–10	11.7
DenseNet	99.28	99.32	0.993	300–500	10–20	28
Vision Transformer	96.96	97.04	0.970	400–600	20–50	50–150
CNN	98.56	98.63	0.986	150–250	5–15	5–15
CNN-LSTM	99.21	99.24	0.992	300–500	15–30	10–25
Polar–CNN–Transformer	99.46	99.52	0.995	500–700	20–40	20–60

The ROC curves of all six models for the bogie axle-box bearing fault diagnosis task are shown in Figure 9. As observed, the CNN–Transformer model achieves the highest true positive rates across almost all false positive rates, resulting in the largest AUC value among the comparison methods. This indicates its superior ability to distinguish between fault categories under the one-vs-rest evaluation. Traditional convolutional models like ResNet18, DenseNet, and CNN exhibit relatively high discrimination capability due to their effective local feature extraction, but their ROC curves lie slightly below that of CNN–Transformer, suggesting less robustness in distinguishing subtle fault differences. The CNN-LSTM model improves over standard CNN in capturing temporal dependencies, reflected in moderately higher ROC curves. In contrast, the Vision Transformer shows lower TPR at moderate FPR levels, likely due to its weaker local feature extraction on small-scale polar projection inputs and the limited dataset size.

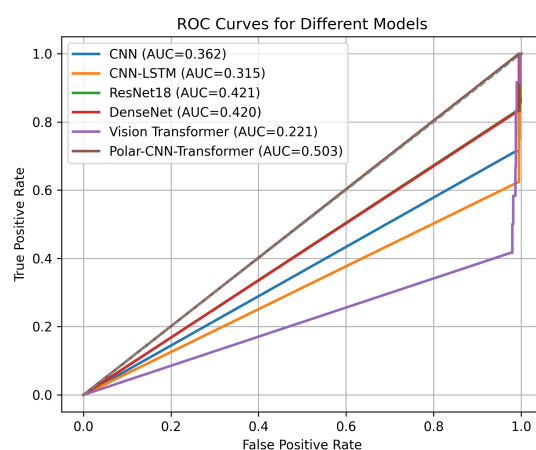


Figure 9. ROC curves for different models.

Next, Figure 10 presents the confusion matrix results for the six comparison methods, providing a more intuitive view of their fault classification capabilities on the bogie axle-box bearing. The diagonal elements of the matrix represent the overall accuracy for each class, while the columns and rows correspond to the true label values and predicted results, respectively. By analyzing the correspondence between the predicted results and true labels, we can clearly identify which categories are more prone to misclassification. Specifically, inner-race faults (label 2) and rolling-element faults (label 3) are more likely to be misclassified, indicating that these fault features have lower distinguishability across the different methods. As shown in Figure 10f, the diagnostic results from the proposed method outperform the other comparison methods, demonstrating its superiority in the bogie axle-box bearing fault diagnosis task.

Finally, to further explore the performance differences of various algorithms in the bogie axle-box bearing fault classification task, this study conducted a t-SNE clustering visualization analysis of the fault classification results for each algorithm, as shown in Figure 11. Figure 11 visually demonstrates the distribution of different categories, allowing for a clear observation of how well the different methods distinguish between fault categories. It is evident that the clustering effect of Vision Transformer in Figure 11c is the worst among the six comparison methods, with considerable overlap between many categories. In contrast, the CNN–Transformer method proposed in this study shows tighter clustering between categories, indicating its strong classification capability and good generalization ability. The clustering results of other algorithms also exhibit some overlap between a few categories. In conclusion, the CNN–Transformer method proposed in this study is more suitable for fault diagnosis and classification of the bogie axle-box bearing compared to the other five compared methods.

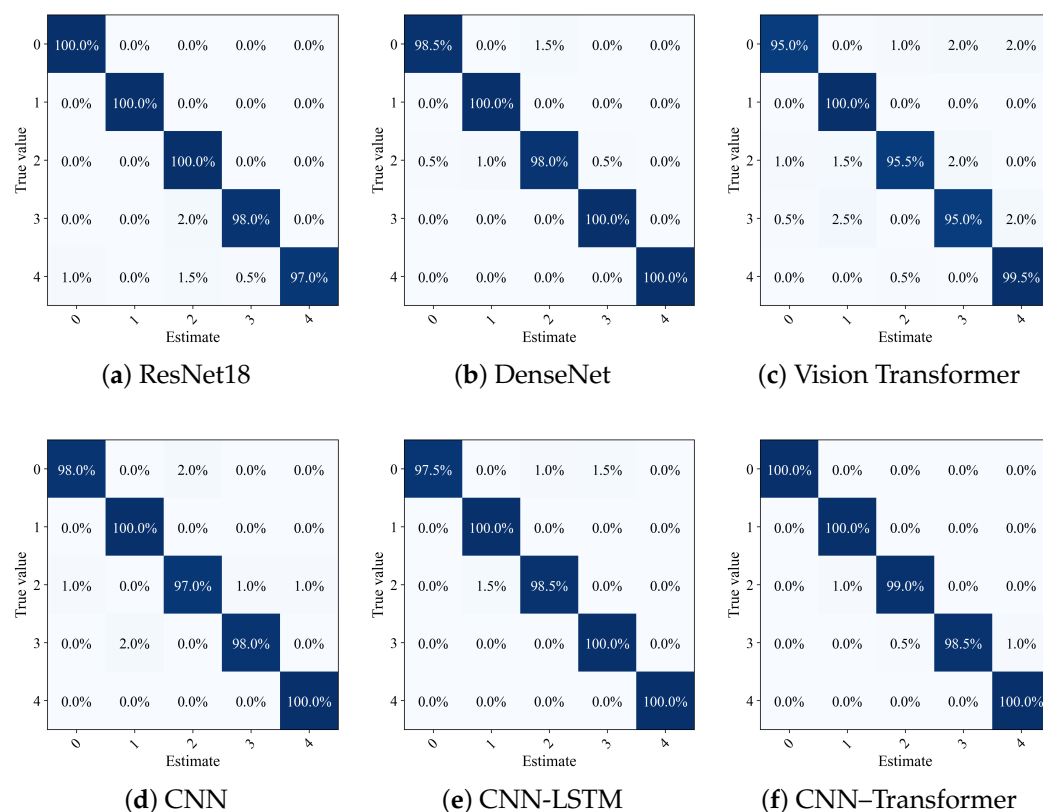


Figure 10. Confusion matrices of fault classification for different comparison methods.

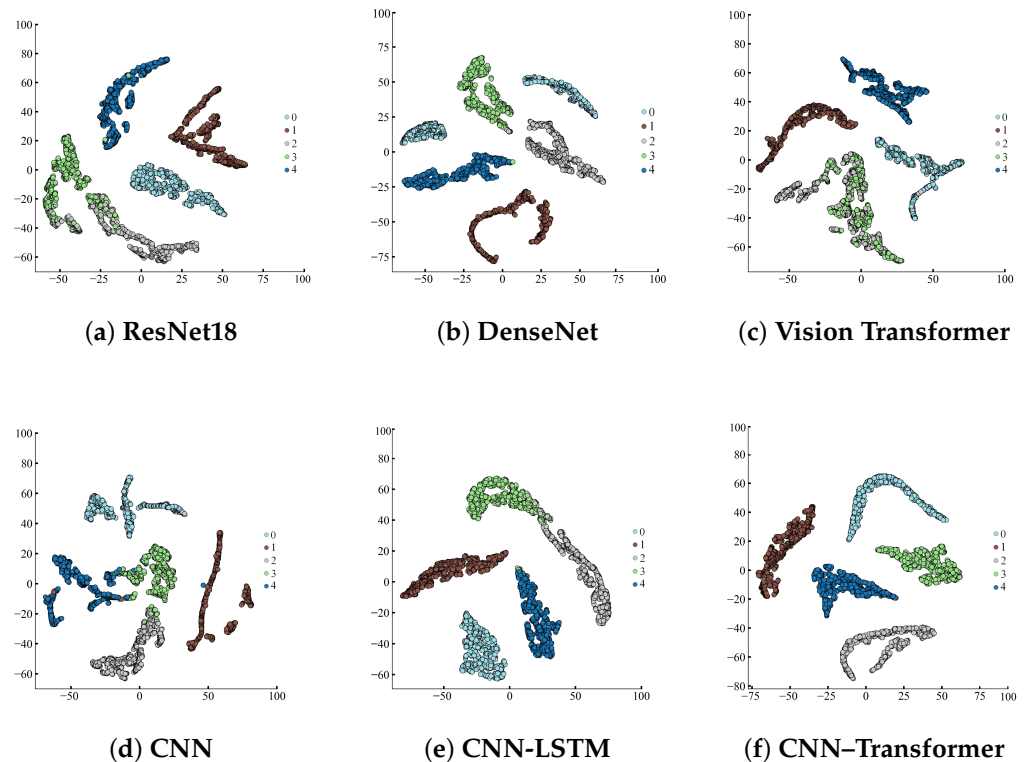


Figure 11. t-SNE plots of fault classification results for different comparison methods.

3.2.4. Analysis of the Necessity of Polar Coordinate Charts

To evaluate the necessity of polar coordinate projection maps in fault diagnosis, the original time-domain signals were fed into the CNN–Transformer model and compared with the polar coordinate input. All models shared the same architecture and hyperparameters, and training was accelerated using an NVIDIA GeForce RTX 3050 GPU. As shown in Figure 12 and Table 7, the time-domain CNN–Transformer achieved 99.38% accuracy, 99.42% recall, and an F1-score of 0.994, with a training time of 400–500 s, inference time of 15–30 ms, and 15–50M parameters. It converged rapidly in the early iterations but exhibited minor loss fluctuations around the 60th iteration, indicating some training instability.

In contrast, the polar coordinate CNN–Transformer achieved slightly higher performance—99.46% accuracy, 99.52% recall, and 0.995 F1-score—with a training time of 500–700 s, inference time of 20–40 ms, and 20–60 M parameters. Although convergence was slower, the training process was smoother, and the loss decreased steadily with minimal fluctuation, demonstrating stronger stability.

Time-domain signals capture the instantaneous temporal features of vibration data, enabling fast early-stage convergence. However, their training is more sensitive to noise and local minima, leading to occasional instability. Polar coordinate projection maps, by transforming signals into structured 2D representations, emphasize nonstationary features and facilitate more robust feature extraction. Consequently, while time-domain CNN–Transformer is advantageous for scenarios prioritizing training speed, polar coordinate CNN–Transformer provides superior stability and enhanced diagnostic capability.

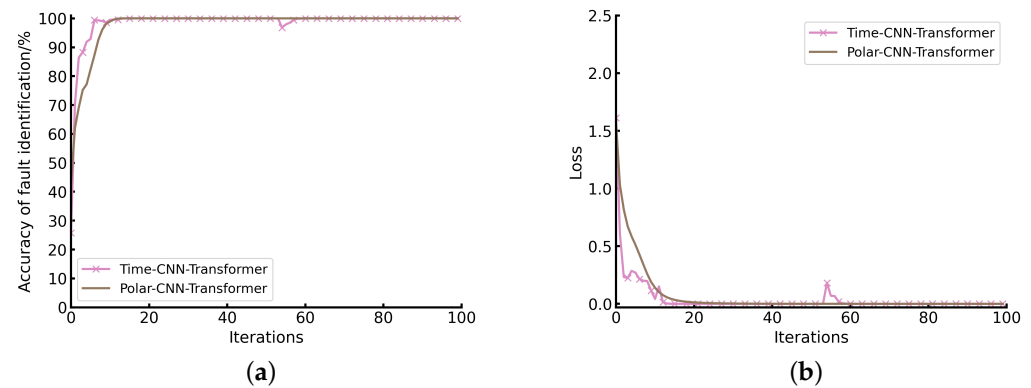


Figure 12. Comparison chart of accuracy and loss values for polar coordinates and time-domain signals. (a) Accuracy rate, (b) Loss value.

Table 7. Comparison of time-domain and polar coordinate CNN-Transformer models.

Method	Accuracy/%	Recall/%	F1-Score	Training Time (s)	Inference Time (ms)	Parameters (M)
Time-CNN-Transformer	99.38	99.42	0.994	400–500	15–30	15–50
Polar-CNN-Transformer	99.46	99.52	0.995	500–700	20–40	20–60

3.2.5. Discussion on Model Robustness Testing

In practical engineering applications, vibration signals are often affected by various types of noise, which may compromise the reliability of fault diagnosis. To evaluate the noise robustness of the proposed model, Gaussian noise with signal-to-noise ratios (SNRs) of 20 dB, 15 dB, and 10 dB was added to the dataset while keeping all trained model parameters unchanged. These noise levels correspond to mild and moderate noise conditions commonly encountered in actual measurement scenarios. The objective of this experiment was to assess the model's classification stability under noise interference and thereby demonstrate its generalization ability and practical applicability.

As shown in Table 8, the classification accuracy of the proposed model decreased only slightly after the addition of Gaussian noise, 99.02% at SNR = 20 dB and 98.90% at SNR = 15 dB, indicating only a minor drop compared with the noise-free condition. In contrast, the CNN-Transformer baseline suffered a more pronounced decline—its accuracy dropped to 98.26% at SNR = 15 dB and further decreased to 97.12% at SNR = 10 dB.

However, in real railway operations, vibration signals are also frequently affected by more complex non-Gaussian noise sources, such as impulsive disturbances caused by wheel-rail impacts or sensor malfunctions that introduce salt-and-pepper noise. Such non-Gaussian noise typically exhibits sparsity, heavy-tailed characteristics, and unpredictable distributions, making it more challenging for fault diagnosis models than Gaussian noise. To further evaluate the model's robustness under these more difficult noise conditions, this study additionally considers salt-and-pepper noise—one of the most common non-Gaussian impulsive noise types in industrial environments. Unlike Gaussian noise, salt-and-pepper noise consists of discrete random impulses, and its intensity is measured by “noise density” rather than SNR. Therefore, noise densities of 1% and 3% were introduced into the dataset for evaluation.

The results showed that the proposed model achieved classification accuracies of 98.56% and 97.84% under 1% and 3% noise densities, respectively, only a slight decrease compared with the noise-free case. In contrast, the CNN-Transformer achieved 98.15% under 1% noise density and dropped significantly to 96.43% under 3% noise density.

Although the CNN–Transformer demonstrated a certain degree of robustness in noisy environments, its performance remained consistently inferior to that of the IEWT–CNN–Transformer. This performance gap confirms that the integration of IEWT substantially enhances the model’s ability to adapt to noise. Overall, the results demonstrate that the proposed model is not only highly robust against continuous noise (e.g., Gaussian noise) but also effectively resists impulsive non-Gaussian disturbances. This further verifies the effectiveness of the IEWT module in signal mode decomposition and denoising, as well as the robustness of the CNN–Transformer framework in extracting critical fault-related features.

Table 8. Classification accuracy under different noise levels.

Noise Type	Noise Intensity Index	Accuracy/%	Model Type
None	–	99.46	IEWT–CNN–Transformer
Gaussian Noise	SNR = 20 dB	99.02	IEWT–CNN–Transformer
Gaussian Noise	SNR = 15 dB	98.90	IEWT–CNN–Transformer
Gaussian Noise	SNR = 15 dB	98.26	CNN–Transformer
Gaussian Noise	SNR = 10 dB	97.65	IEWT–CNN–Transformer
Gaussian Noise	SNR = 10 dB	97.12	CNN–Transformer
Salt and pepper noise	Density = 1%	98.56	IEWT–CNN–Transformer
Salt and pepper noise	Density = 1%	98.15	CNN–Transformer
Salt and pepper noise	Density = 3%	97.84	IEWT–CNN–Transformer
Salt and pepper noise	Density = 3%	96.43	CNN–Transformer

3.2.6. Ablation Experiment

To evaluate the specific contributions of each module in the proposed method to fault diagnosis performance, ablation experiments were conducted. By removing the IEWT module or the Transformer attention module, or by using only CNN for feature extraction, the impact of each component on overall model performance could be quantified. As shown in Table 9, the contributions of different modules to fault diagnosis were evident. Using only the CNN model yielded an accuracy of 98.56% on the test set; adding the Transformer attention module increased the accuracy to 99.11%, demonstrating that the attention mechanism enhanced the model’s ability to extract key fault features. Combining IEWT with CNN achieved 98.96% accuracy, indicating that the IEWT module played a crucial role in denoising the signal and extracting key modal components. The full IEWT–CNN–Transformer model achieved the highest accuracy of 99.46%, confirming the synergistic effect of the IEWT and the Transformer attention module, and enabling superior fault diagnosis performance and stability under complex vibration signals.

Table 9. Ablation experiment results.

Model Configuration	Accuracy (%)
IEWT–CNN–Transformer	99.46
IEWT–CNN	98.96
CNN	98.56
CNN–Transformer	99.11

4. Conclusions

- (1) Considering the strong nonstationarity and complex multi-source interference of bogie axle-box vibration signals, this paper proposed a fault diagnosis method based on an improved EWT and CNN–Transformer combination. To overcome the limitations of fixed frequency-band boundaries in traditional EWT methods, this study introduced a dynamic frequency-band division strategy based on energy distribution, which

can adaptively adjust the frequency-band boundaries to better align with the actual changes in the signal characteristics. First, the IEWT was applied to decompose the bogie axle-box vibration signal to extract fault-related feature signals. Then, information entropy was used to filter the decomposed modal components, and key modal signals were reconstructed to extract important feature information.

- (2) The reconstructed signal was converted into a polar coordinate projection map, displaying the signal's frequency variation and amplitude differences through both angle and radius, highlighting its nonstationary characteristics. To extract local spatial features, a CNN was used for feature learning, while Transformer's self-attention mechanism captured the global dependencies within the signal. The combination of CNN and Transformer leveraged the advantages of image feature extraction and global modeling. The diagnostic performance of various comparison methods was analyzed using confusion matrices and t-SNE clustering algorithms. Experimental results showed that the proposed method significantly outperformed the five comparison methods in terms of accuracy for bogie axle-box bearing fault diagnosis, effectively improving fault recognition performance.
- (3) Based on the comparison between signal lengths of 1024 and 2048, both settings preserved the main fault features, while 2048-length samples provided finer polar coordinate representations. Therefore, future work will prioritize longer signal segments to enhance feature expression and diagnostic stability. Meanwhile, efforts will focus on improving computational efficiency, reducing hardware dependence, and expanding fault categories and operating conditions to strengthen the generalization and practical applicability of the proposed method.

Author Contributions: X.D. and L.Z. provided ideas and wrote the first draft; Z.L. supervised the entire project process and gave methodological guidance; Z.L. and X.D. revised the first draft; X.D., Z.L., M.T. and X.S. wrote the experimental code; X.D., Z.L., M.T. and X.S. integrated the experimental data. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 52472342.

Data Availability Statement: The data used in this article can be downloaded from the following webpage: <https://engineering.case.edu/bearingdatacenter/download-data-file>, accessed on 4 February 2026.

Conflicts of Interest: Authors Xiaoping Ding, Minghui Tang and Xiaoxu Shen were employed by the company Zhejiang Xingfu Rail Transit Operation Management Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Jin, Z.; He, D.; Wei, Z. Intelligent fault diagnosis of train axle box bearing based on parameter optimization VMD and improved DBN. *Eng. Appl. Artif. Intell.* **2022**, *110*, 104713. [\[CrossRef\]](#)
- Guo, L.; Yu, Y.; Chen, Z.; Liu, Y.; Gao, H. Study on dynamic characteristics of urban rail transit vehicle considering faulty axle-box bearings under variable speeds. *Mech. Syst. Signal Process.* **2024**, *206*, 110849. [\[CrossRef\]](#)
- Wang, Z.; Cheng, Y.; Allen, P.; Yin, Z.; Zou, D.; Zhang, W. Analysis of Vibration and Temperature on the Axle Box Bearing of a High-Speed Train. *Veh. Syst. Dyn.* **2020**, *58*, 1605–1628. [\[CrossRef\]](#)
- Zheng, Z.; Song, D.; Zhang, W.; Jia, C. A Fault Diagnosis Method for Bogie Axle Box Bearing Based on Sound-Vibration Multiple Signal Fusion. *Appl. Acoust.* **2025**, *228*, 110336. [\[CrossRef\]](#)
- Zhong, L.; Xue, L.; Deng, X.; Liu, S.; Zhu, Z. High power-efficiency readout circuit for MEMS capacitive accelerometer. *IEEE Trans. Circuits Syst. II Express Briefs* **2023**, *71*, 76–80. [\[CrossRef\]](#)
- Wei, X.; Verhaegen, M. Sensor and actuator fault diagnosis for wind turbine systems by using robust observer and filter. *Wind. Energy* **2011**, *14*, 491–516. [\[CrossRef\]](#)

7. Nan, C.; Khan, F.; Iqbal, M.T. Real-time Fault Diagnosis Using Knowledge-Based Expert System. *Process Saf. Environ. Prot.* **2008**, *86*, 55–71. [\[CrossRef\]](#)
8. Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G. Automatically designing CNN architectures using the genetic algorithm for image classification. *IEEE Trans. Cybern.* **2020**, *50*, 3840–3854. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Zhao, J.; Shen, J.; Liu, L. Bus Passenger Flow Classification Prediction Driven by CNN-GRU Model and Multi-Source Data. *J. Traffic Transp. Eng.* **2021**, *21*, 265–273.
10. Du, X.; Li, K.; Lv, Y.; Qiu, S. Motor imaging EEG signal recognition of ResNet18 network based on deformable convolution. *Electronics* **2022**, *11*, 3674. [\[CrossRef\]](#)
11. Abumohsen, M.; Owda, A.Y.; Owda, M. Electrical load forecasting using LSTM, GRU, and RNN algorithms. *Energies* **2023**, *16*, 2283. [\[CrossRef\]](#)
12. Bengtsson, C. Status and trends in transformer monitoring. *IEEE Trans. Power Deliv.* **2002**, *11*, 1379–1384. [\[CrossRef\]](#)
13. Lin, Y.; Chen, K.; Zhang, X.; Tan, B.; Lu, Q. Forecasting crude oil futures prices using BiLSTM-Attention-CNN model with Wavelet transform. *Appl. Soft Comput.* **2022**, *130*, 109723. [\[CrossRef\]](#)
14. Gai, J.; Hu, Y.; Shen, J. A Bearing Performance Degradation Modeling Method Based on EMD-SVD and Fuzzy Neural Network. *Shock Vib.* **2019**, *2019*, 5738465. [\[CrossRef\]](#)
15. Mohamad, T.H.; Abbasi, A.; Kim, E.; Nataraj, C. Application of Deep CNN-LSTM Network to Gear Fault Diagnostics. In Proceedings of the 2021 IEEE International Conference on Prognostics and Health Management (ICPHM), Detroit, MI, USA, 7–9 June 2021; IEEE: New York, NY, USA, 2021; pp. 1–6.
16. Li, J.; Wu, Y. Improved sparrow search algorithm with the extreme learning machine and its application for prediction. *Neural Process. Lett.* **2022**, *54*, 4189–4209. [\[CrossRef\]](#)
17. Shang, H.; Li, Y.; Lin, W. Partial discharge feature extraction of a transformer based on EWT-MQE. *Power Syst. Prot. Control* **2022**, *50*, 161–171. (In Chinese)
18. Deng, W.; Zhang, S.; Zhao, H.; Yang, X. A Novel Fault Diagnosis Method Based on Integrating Empirical Wavelet Transform and Fuzzy Entropy For Motor Bearing. *IEEE Access* **2018**, *6*, 35042–35056. [\[CrossRef\]](#)
19. Li, Z.; Zhou, W.; Chang, Y.; Liu, W.; Zhu, J. A noise reduction method based on EWT-ICA for spectrum induced polarisation data. *Explor. Geophys.* **2021**, *52*, 170–178. [\[CrossRef\]](#)
20. Xu, J.; Zhao, S.; Ma, C.; Ba X. Damage Monitoring of Engine Fan Blades Based on EWT-entropy Method. *J. Aerosp. Power* **2023**, *38*, 23–31. (In Chinese)
21. Lu, W.; Li, J.; Wang, J.; Qin, L. A CNN-BiLSTM-AM method for stock price prediction. *Neural Comput. Appl.* **2021**, *33*, 4741–4753. [\[CrossRef\]](#)
22. Tipirneni, S.; Reddy, C.K. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans. Knowl. Discov. Data* **2022**, *16*, 105. [\[CrossRef\]](#)
23. Cao, H.; Fan, F.; Zhou, K.; He, Z. Wheel-Bearing Fault Diagnosis of Trains Using Empirical Wavelet Transform. *Measurement* **2016**, *82*, 439–449. [\[CrossRef\]](#)
24. Gu, J.; Peng, Y.; Lu, H.; Chang, X.; Chen, G. A Novel Fault Diagnosis Method of Rotating Machinery via VMD, CWT and Improved CNN. *Measurement* **2022**, *200*, 111635. [\[CrossRef\]](#)
25. Zhang, Y.; Tang, Y.; Liu, Y.; Liang, Z. Fault Diagnosis of Transformer Using Artificial Intelligence: A Review. *Front. Energy Res.* **2022**, *10*, 1006474. [\[CrossRef\]](#)
26. He, J.; Qu, L.; Wang, P.; Li, Z. An oscillatory particle swarm optimization feature selection algorithm for hybrid data based on mutual information entropy. *Appl. Soft Comput.* **2024**, *152*, 111261. [\[CrossRef\]](#)
27. Rong, J.; Wen, L.; Xiao, J. Efficiency improvement of the polar coordinate transformation for evaluating BEM singular integrals on curved elements. *Eng. Anal. Bound. Elem.* **2014**, *38*, 83–93. [\[CrossRef\]](#)
28. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [\[CrossRef\]](#)
29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.