frontiers | Frontiers in Mechanical Engineering

TYPE Original Research
PUBLISHED 18 December 2025
DOI 10.3389/fmech.2025.1708745

OPEN ACCESS

EDITED BY
Nima Rezazadeh,
Università della Campania Luigi Vanvitelli, Italy

REVIEWED BY
Yanfeng Peng,
Hunan University of Science and Engineering, China
Shila Fallahy,
Polytechnic of Milan, Italy
Alireza Keyhani Asl,
Birmingham City University, United Kingdom

*CORRESPONDENCE
Jin Chen,
13519012451@163.com

RECEIVED 19 September 2025
REVISED 21 November 2025
ACCEPTED 28 November 2025
PUBLISHED 18 December 2025

CITATION
Chen J (2025) Fault diagnosis of large-scale electric pumping and irrigation electromechanical equipment by integrating CWT and swin transformer.
Front. Mech. Eng. 11:1708745.
doi: 10.3389/fmech.2025.1708745

Fault diagnosis of large-scale electric pumping and irrigation electromechanical equipment by integrating CWT and swin transformer

Jin Chen*

Gansu Jingtaichuan Electric Pumping Irrigation Water Resources Utilization Center, Baiyin, Gansu, China

This paper proposes a large-scale electromechanical equipment fault diagnosis framework that integrates continuous wavelet transform (CWT), Swin transformer, and cross attention. Firstly, CWT uses optimized complex Morlet wavelets (impact sharpness ratio factor 0.82, frequency focus ratio factor 0.67) to map vibration signals to high-resolution time-frequency images. Secondly, the four level Swin Transformer layer extracts multi-scale features (56 × 56 × 96 to 7 × 7 × 768), while stride convolution aligns shallow features for cross attention fusion, where shallow features serve as queries and deep features serve as keys/values, dynamically weighting and integrating cross stage information. The experimental results show that the classification accuracy is 98.7% (macro F1 is 98.5%), which can perfectly identify the motor rotor eccentricity (MRE). T-SNE visualization validates the enhanced intra class compactness and inter class separability, particularly between MRE and gear wear (GTW). The model maintains robustness under noise, with an accuracy increase from 82.3% (signal-to-noise ratio = 10 dB) to 98.7% (signal-to-noise ratio = 90 dB), verifying its effectiveness and reliability in intelligent maintenance in complex industrial environments.

KEYWORDS

swin transformer, continuous wavelet transform, electric pumping, electromechanical equipment, fault diagnosis

1 Introduction

Large-scale electric pumping equipment, as the core equipment of agricultural irrigation and water conservancy projects, undertakes key tasks such as drought resistance and flood control, and cross-regional water resources allocation. Its stable operation directly affects agricultural production efficiency and regional water security (Xia and Hong, 2022; Zhang, 2024). However, under long-term high-load conditions, electromechanical equipment is prone to complex faults such as bearing wear (Dexian, 2025; Wang and Zhang, 2023), motor instability, and blade breakage, resulting in system shutdown or even significant losses (Vokhidov and Churakova, 2021; Hatsey and Birkie, 2021). Traditional fault diagnosis (Ventricci et al., 2024; Ghafouri Matanagh et al., 2024) methods mostly rely on FFT (Fast Fourier Transform), STFT (Short-Time Fourier Transform), or CNN (Convolutional Neural Network). However, FFT is difficult to characterize the mutation characteristics of non-stationary signals; STFT has limited time-frequency resolution; CNN has local receptive field constraints in multi-scale

feature extraction, which makes it difficult to adapt to the nonlinear and multi-scale characteristics of electromechanical equipment fault signals (Wu et al., 2025; Trejo-Chavez et al., 2023). Therefore, it is urgent to explore more efficient feature representations and intelligent diagnosis frameworks to improve fault identification accuracy and engineering applicability.

This paper proposes a fault diagnosis method that integrates CWT and Swin Transformer. First, the vibration signal is converted into time domain information through CWT, and the complex Morlet wavelet basis function is used to realize multi-band time-frequency localization analysis, breaking through the resolution bottleneck of traditional time-frequency tools. Furthermore, a hierarchical Swin Transformer architecture is designed, and a feature pyramid is constructed through a multi-stage sliding window self-attention mechanism to abstract fault semantic information step by step from the local to the global scale. In view of the difficulty of multi-scale feature fusion, the Cross-Attention mechanism is applied to dynamically quantify the association weights of shallow detail features and deep abstract features to achieve cross-modal feature interaction and information enhancement. Compared with existing methods, this model not only retains the time-frequency characteristics of the signal, but also strengthens the expression ability of key fault features through an adaptive attention mechanism, solves the problems of traditional CNN feature simplification and Transformer global modeling redundancy, and provides a new paradigm for fault diagnosis under complex working conditions.

This paper proposes a CWT-Swin Transformer-Cross-Attention fusion framework, integrating the Cross-Attention mechanism between CWT and Swin Transformer to construct a "wavelet-Transformer" cross-modal fusion paradigm. By dynamically weighting shallow features as queries and deep features as keys/values, this system addresses the challenges of spatial misalignment and insufficient interaction of multi-scale features, achieving refined representation and discriminative fusion of the time-frequency features of non-stationary signals. The main contributions of this paper include that: (1) the CWT-Swin Transformer-Cross-Attention fusion framework is proposed, and the hierarchical visual Transformer is applied to the fault diagnosis of electric pumping equipment, breaking through the limitations of traditional time-frequency analysis and shallow feature extraction; (2) the refined time-frequency characterization of non-stationary signals is achieved through CWT mapping, providing structured input for deep learning; (3) the wavelet-Transformer cross-modal fusion paradigm is created; the dynamic interaction between shallow detail features and deep abstract features is realized through Cross-Attention; the multi-scale feature association weights are quantified in a unified dimensional space, systematically solving the feature coupling problem caused by the insufficient time-frequency resolution of the STFT-CNN method. The t-SNE visualization results show that the application of Cross-Attention makes the sample points of each fault category show stronger intra-class aggregation and inter-class separability, and strengthens the expression of key features. In the feature splicing method without Cross-Attention, the category boundaries of MRE and GTW are blurred, and the cross-stage information interaction ability is weak. The model in this paper has an accuracy rate of 82.3% in identifying multiple types of faults under noisy conditions (SNR = 10 dB) and has excellent noise resistance. This study not only provides a high-precision solution for mechanical and electrical equipment fault diagnosis, but also expands the application boundaries of the Transformer architecture in the field of industrial signal processing, which is of great significance to promoting the development of intelligent operation and maintenance technology.

2 Related work

Early studies mostly used FFT (Karagiovanidis et al., 2023; Bousseksou et al., 2025) to extract frequency domain features. STFT (Ribeiro Junior et al., 2022) enhances the time-frequency resolution through sliding windows, but the fixed window length limits the extraction of multi-scale features. Although WPT (Wavelet Packet Transform) (Ahmad et al., 2023) can adaptively decompose signal frequency bands, it relies on manually selected basis functions and feature engineering, and its generalization ability is insufficient. Recently, methods based on deep learning (Benameur et al., 2024; Ullah et al., 2020) have gradually become a research hotspot: CNN extracts local features through local receptive fields, but has weak modeling capabilities for long-distance dependencies; recurrent neural networks and their variants (Mallak and Fathi, 2021) can capture the dynamic characteristics of time series, but there are bottlenecks in gradient vanishing and computational efficiency. Traditional electromechanical equipment fault diagnosis methods mainly use time-frequency analysis tools such as FFT and STFT combined with shallow machine learning models. Although some results have been achieved in steady-state signal analysis, it is difficult to adapt to the needs of transient feature extraction of non-stationary signals. The FFT lacks temporal localization capability; the fixed window length of the STFT limits multi-scale modeling; the WPT relies on artificial basis function selection and feature engineering, which results in limited generalization capability. Existing research generally faces two core problems: first, it is difficult to balance the signal time-frequency resolution and computational complexity; second, it lacks the ability to hierarchically model multi-scale fault features, resulting in large fluctuations in diagnostic accuracy under complex working conditions.

In response to the difficulty of multi-scale feature extraction, researchers have tried to combine time-frequency analysis with deep learning. The hybrid model based on wavelet transform (Wavelet-CNN) (Liu et al., 2024) generates time-frequency images through CWT and inputs them into CNN, using its translation invariance to enhance feature robustness, but the fixed receptive field of CNN limits the learning of cross-scale feature correlation. Another method uses a multi-scale parallel network (Inception module) (Xu et al., 2024) to capture local and global features simultaneously through different convolution kernels, but the parameter redundancy is high, and it is difficult to adapt to nonlinear changes in the signal. Recently, the Transformer (Wang R. et al., 2024; Li et al., 2024) architecture has been applied in the field of fault diagnosis due to its global attention mechanism: ViT (Vision Transformer) (Xie et al., 2024) divides the image into blocks and then performs self-attention calculations, but its computational complexity increases with the square of the input

length, making it difficult to process high-resolution time-frequency graphs; Swin Transformer (Xiaofeng, 2025; Du and Shaohui, 2023) reduces the amount of computation through a sliding window mechanism while retaining the ability to extract hierarchical features, and has achieved significant results in industrial image classification tasks. In recent years, studies have attempted to combine time-frequency analysis with deep learning. For example, Wavelet-CNN uses CWT time-frequency graphs to enhance the robustness of CNN, but the fixed receptive field limits cross-scale associations; the Inception module achieves multi-scale parallel modeling through multi-core convolution, but its efficiency is affected by parameter redundancy. The Transformer architecture has been applied into the field of fault diagnosis due to its global attention mechanism, but ViT has high computational complexity. Although Swin Transformer reduces complexity, it still faces the problem of multi-scale feature space alignment and cross-stage interaction. Existing methods have not yet systematically addressed the problems of dynamic weight allocation and deep semantic co-optimization in multi-scale fusion.

The development of multi-scale feature fusion technology has provided new ideas for fault diagnosis. Early studies used feature concatenation or element-by-element addition to achieve multi-layer feature integration, but lacked quantitative evaluation of feature importance. Attention mechanisms (Jier et al., 2022; Xiang et al., 2024) enhance key features by adjusting channel or spatial weights, but are limited to single-scale internal optimization. Cross-modal attention (Song et al., 2024; Cao et al., 2024) achieves heterogeneous feature association through Query-Key-Value interaction and performs well in areas such as image-text retrieval, but its application in industrial signal processing is still in the exploratory stage. For time-frequency feature fusion, some studies have tried to input CWT coefficients (Wang S. et al., 2024) as additional channels into the Transformer, or extract features separately through multi-scale parallel branches and then perform weighted averaging, but failed to solve the spatial misalignment problem of features at different scales. Recent studies have explored the combination of CWT and Swin Transformer, for example, in applications such as ECG arrhythmia detection (Chen et al., 2024) and structural damage identification (Xin et al., 2024). These works have confirmed the potential of the CWT-Swin Transformer architecture in processing complex signals. However, these methods mainly focus on the hierarchical feature extraction capabilities of the Swin Transformer itself, but pay insufficient attention to the deep fusion and dynamic interaction mechanisms of features at different stages (scales). Multi-scale feature fusion technology has developed from early feature concatenation to attention mechanisms, gradually realizing dynamic adjustment of feature weights, but is still limited to single-scale internal optimization. Cross-modal attention improves the correlation of heterogeneous features through Query-Key-Value interaction, but its application in industrial signal processing is limited. Hybrid frequency-adaptive neural networks have shown great potential in the field of non-stationary signal processing. Reference (Ravikumar et al., 2025) dynamically adjusts the network architecture of its receptive field or attention weight according to the signal frequency band. In contrast, reference (Kumar et al., 2024) explores deep learning models that combine time-frequency analysis with adaptive filter banks. Existing research has made progress in signal processing and fault diagnosis. For example, reference (Zhou et al., 2025) proposed a heuristic denoising method based on empirical Fourier-Bessel, which effectively improved the identifiability of gear fault features; reference (Fan et al., 2024) designed a variable-scale multilayer perceptron for abnormal detection and efficient recovery of vibration data in helicopter transmission systems, providing a new analytical approach for complex electromechanical systems. These methods provide a new paradigm for signal processing by implementing frequency perception within the model. However, these technologies are still in the exploratory stage, and their generalization ability, computational efficiency, and integration with traditional time-frequency analysis tools in complex multi-fault classification tasks need further verification. For the fusion of time-frequency features, existing methods either directly concatenate CWT coefficients or use multi-branch weighted averaging, neither of which solves the problem of spatial misalignment between scales. In addition, the robustness under noise interference and working condition changes is insufficient, and a fusion framework with multi-scale alignment, dynamic weighting, and noise resistance is urgently needed. The CWT-Swin Transformer-Cross-Attention architecture proposed in this paper systematically breaks through the above bottlenecks through hierarchical feature construction and cross-stage attention interaction, providing a new paradigm for the diagnosis of complex electromechanical systems.

3 Methods

3.1 Overall framework

The framework of this method is displayed in Figure 1.

This paper proposes a fault diagnosis method for large-scale electric pumping electromechanical equipment that integrates CWT and Swin Transformer. By constructing a two-dimensional time-frequency tensor and obtaining 224 × 224 × 3 through channel replication, high-precision classification is achieved by combining Swin Transformer hierarchical features and a cross-stage attention mechanism. First, the vibration signal is transformed by CWT to generate a time-frequency coefficient matrix. A hierarchical Swin Transformer architecture is designed to extract multi-scale features step by step through a four-stage sliding window self-attention mechanism (Stage1: 56 × 56 × 96→Stage4: 7 × 7 × 768). To solve the problem of multi-scale feature fusion, a feature alignment strategy based on stride convolution is proposed to unify the outputs of Stage1–3 into 7 × 7×C dimensions and input them into the Cross-Attention module together with the Stage4 features. The Query is constructed through shallow features, and the deep features are used as Key/Value. The attention weights are dynamically calculated to complete the weighted fusion of cross-stage features and output global features. Finally, the dimension is reduced through global pooling, and the fully connected layer and Softmax classifier output 8 types of fault probabilities (normal, inner/outer ring damage of bearing, rotor eccentricity, blade breakage, shaft misalignment, mechanical looseness, and gear wear).

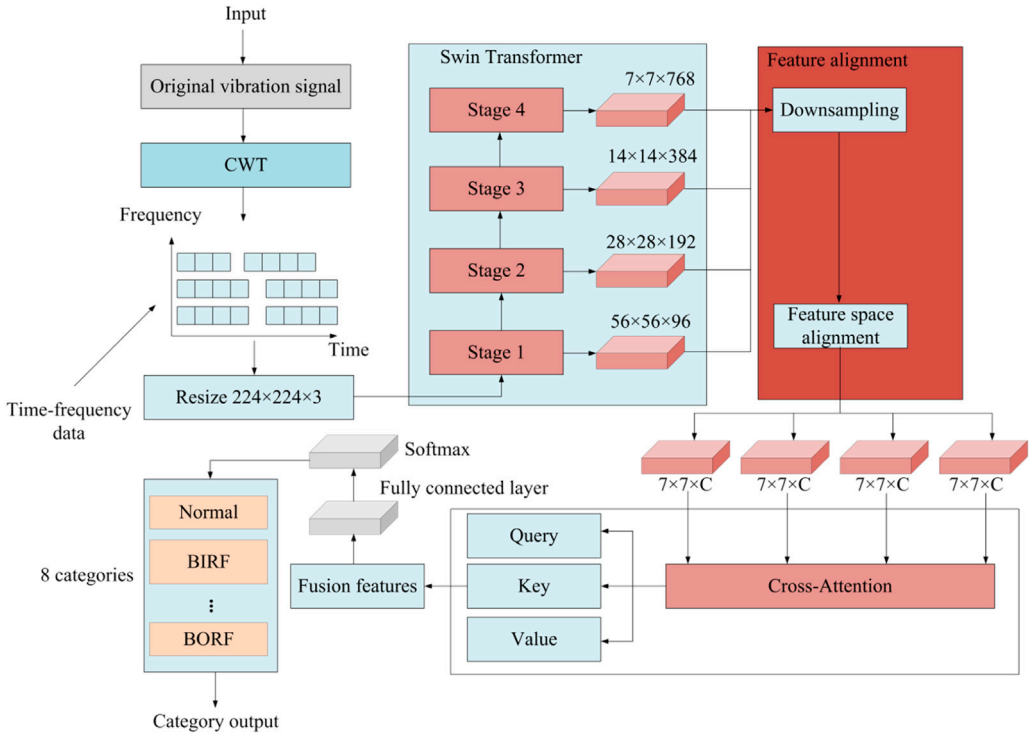

FIGURE 1
Framework of this method.

FIGURE 2
Fault simulation platform.

3.2 Data preprocessing and CWT conversion

3.2.1 Vibration signal acquisition

The fault diagnosis dataset of electric pumping electromechanical equipment constructed in this study is collected based on the fault simulation platform built in the laboratory, covering eight typical fault types and normal states: normal state, bearing inner ring damage, bearing outer ring damage, motor rotor eccentricity, blade breakage, shaft misalignment, mechanical looseness, and gear wear. Data acquisition is completed through a three-axis acceleration sensor. The sensor is installed at the motor drive end bearing seat and the pump body outlet flange. The sampling system is configured with 16-bit resolution, a sampling frequency of 10 kHz (satisfying the Nyquist sampling theorem of fault characteristic frequency <4 kHz), and a single sampling time of 204.8 m (covering at least 3 complete frequency cycles, the rated speed of the equipment is 3000 rpm corresponding to the base frequency of 50 Hz). The signal acquisition conditions cover 5 levels of load gradient from no load to full load (0%–100%), and 10 sets of repeated samples are collected at each load level to enhance the robustness of data distribution.

The fault simulation platform is shown in Figure 2.

The dataset is divided into training, validation, and test sets in a ratio of 8:1:1. The training set contains 8152 groups of samples, and the validation and test sets each contain 1019 groups. To eliminate the interference of signal amplitude changes with load, the original vibration data is normalized by Z-score, and the environmental noise is suppressed by wavelet threshold denoising (db4 mother wavelet, 3-layer decomposition, SureShrink threshold rule). Table 1 displays the sample distribution of various types of faults in the training, validation, and test sets. The data division strictly follows the principle of category balance to ensure the statistical validity of model training and evaluation.

3.2.2 Continuous wavelet transformation

This study uses CWT (Luczak, 2024; Djaballah et al., 2023) to map the vibration signal to the time-frequency domain, and its mathematical expression is:

$$W_x(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \quad (a>0) \tag{1}$$

TABLE 1 Sample distribution of nine operating states (eight fault types and normal state).

Type	Training set	Validation set	Test set	Total
Normal	1200	150	150	1500
BIRF (bearing inner race fault)	984	123	123	1230
BORF (bearing outer race fault)	896	112	112	1120
MRE	864	108	108	1080
IBB (impeller blade breakage)	1160	145	145	1450
SM (shaft misalignment)	1072	134	134	1340
ML (mechanical looseness)	984	123	123	1230
GTW	992	124	124	1240
Total	8152	1019	1019	10190

In Formula 1, $x(t)$ is the input vibration signal, and ψ is the mother wavelet basis function. $W_x(a,b)$ is the complex-valued wavelet coefficient, whose modulus reflects the energy intensity of the signal at scale a and time b.

The scale-frequency mapping formula is:

$$f = \frac{f_c}{a \cdot \Delta t} \tag{2}$$

In Formula 2, $f_c = 0.85$ is the wavelet center frequency, and Δt is the sampling interval. To meet the input format of the subsequent Swin Transformer, the data channel is copied three times to form three-dimensional data.

To optimize the time-frequency localization performance, this paper compares and analyzes the three wavelet basis functions of Gabor wavelet, Mexican hat wavelet, and complex Morlet (Silik et al., 2021; Zhang et al., 2022) in their ability to characterize fault characteristics.

The Gabor wavelet is:

$$\psi(t) = e^{-\frac{t^2}{2\sigma^2}} \cos(\omega_0 t) \tag{3}$$

In Formula 3, the parameter σ controls the window width, and ω_0 determines the oscillation frequency.

The Mexican hat wavelet is:

$$\psi(t) = (1-\beta t^2)e^{-\alpha t^2} \tag{4}$$

In the Formula 4, the Mexican hat wavelet adjusts the oscillation characteristics through parameters α and β, which is suitable for transient impulse signals, but the sidelobe suppression in the high-frequency band is insufficient.

The complex Morlet wavelet is as the Formula 5:

$$\psi(t) = \pi^{-1/4} e^{i\omega_0 t} e^{-t^2/2} \tag{5}$$

The complex Morlet wavelet has both real and imaginary parts, can separate signal phase information, and can flexibly balance time-frequency resolution, especially in non-stationary signal analysis.

To optimize the performance of complex Morlet wavelets in terms of impulse response sharpness and bandgap focusing, this study employs a grid search strategy to systematically evaluate different combinations of scaling factors ranging from 0.5 to 1.2 on a validation set. Evaluation metrics include impulse response sharpness (defined as the normalized value of the temporal impulse kurtosis) and bandgap focusing (defined as the normalized concentration of the frequency domain energy distribution). Experimental results show that an impulse response sharpness of 0.82 yields the best performance, while a bandgap focusing is highest at 0.67. Further sensitivity analysis reveals that the model accuracy changes by less than 1.5% when the scaling factor fluctuates within ±0.1, indicating strong robustness of the selected parameters. Ultimately, this study selects a scaling factor of 0.82 for impulse feature extraction and 0.67 for bandgap localization to achieve high-resolution time-frequency characterization of non-stationary signals.

The comparison of the prominence of different wavelet basis functions is displayed in Table 2.

Table 2 shows that the complex Morlet wavelet is superior to Gabor and Mexican hat in terms of impact response sharpness (0.82) and frequency band focus (0.67), indicating that it has more advantages in capturing transient impulse features and frequency resolution in non-stationary signals. The complex Morlet can adaptively match the time-frequency localization characteristics of the fault signal through flexible adjustment. Especially in the diagnosis of bearing damage and blade fracture faults, its high sharpness and high focus can effectively separate periodic impulses and broadband noise. Although the computational complexity of the complex Morlet (1.7×10^6 FLOPs) is higher than that of Gabor (1.2×10^6 FLOPs) and Mexican hat (1.5×10^6 FLOPs), its feature enhancement effect is significant, which meets the high-precision requirements of industrial scenarios. Therefore, the complex Morlet wavelet base achieves the best balance between feature extraction performance and engineering applicability, and becomes the preferred solution for CWT transformation in this paper.

The comparison of the time-frequency transformation effects of different wavelet bases is depicted in Figure 3.

In Figure 3, the horizontal axis is time, and the vertical axis is frequency. The color represents energy intensity. By comparing and analyzing the time-frequency transformation effects of the three wavelet basis functions, the complex Morlet wavelet shows a significant advantage in time-frequency localization. Compared with the frequency band diffusion of the Gabor wavelet and the

TABLE 2 Comparison of the prominence of different wavelet basis functions.

Wavelet basis	Impact response sharpness	Frequency band focus	Computational complexity (FLOPs, floating point operations)
Gabor	0.68	0.52	1.2×10^6
Mexican hat	0.73	0.61	1.5×10^6
Complex morlet	0.82	0.67	1.7×10^6

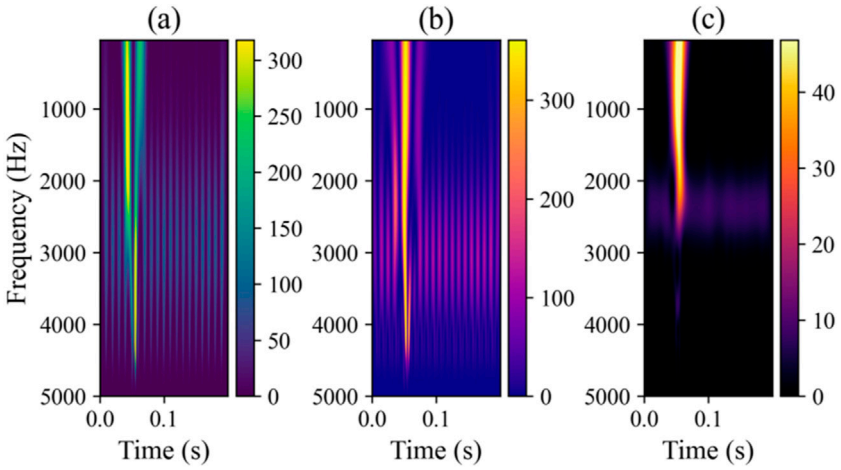

FIGURE 3
Time-frequency transformation effects of different wavelet bases. **(a)** Gabor wavelet base. **(b)** Mexican hat wavelet base. **(c)** Complex Morlet wavelet base.

fuzzy impulse response of the Mexican hat wavelet, the transient impulse feature of the complex Morlet wavelet at 0.05 s presents the highest sharpness while maintaining excellent frequency band focus ability. Its complex exponential characteristics effectively separate phase information, achieve the optimal balance of time-frequency resolution in non-stationary signal analysis, and can precisely capture the periodic impulse characteristics of faults such as bearing damage, significantly suppressing broadband noise interference. CWT transformation may introduce artifacts such as edge effects and scale leakage. This study effectively suppresses these effects by selecting a complex Morlet wavelet and optimizing scale parameters, combined with signal preprocessing (denoising and normalization).

3.3 Swin Transformer feature extraction

3.3.1 Network structure design

Swin Transformer (Zeng et al., 2024; Fu et al., 2024) is a hierarchical visual Transformer architecture that realizes local-global feature interaction through a sliding window mechanism, significantly reducing computational complexity. The input of Swin Transformer is a time-frequency graph generated by continuous wavelet transform, which is adapted to the input format of Swin Transformer through a three-channel replication strategy.

The input data is divided into non-overlapping image blocks (Patch Size = 4), each of which is 4 × 4 in size, and linearly projected along the channel dimension to the embedding dimension to generate the initial feature representation, as shown in Formula 6:

$$E_1 = \text{Linear}(\text{PatchPartition}(X)) \tag{6}$$

Stage 1: local feature extraction is performed by stacking two Swin Transformer Blocks (including window multi-head self-attention modules and feedforward networks) to output feature $F_1 \in R^{56\times56\times96}$. The features of this stage mainly capture the local texture and edge information in the time-frequency graph, which is suitable for characterizing the local impact features of faults such as bearing damage or blade fracture.

Based on Stage 1, Stage 2 uses the Patch Merging operation to halve the feature spatial resolution (downsample) and double the number of channels to 192. Each group of adjacent 2 × 2 image blocks is merged into a super block and linearly projected along the channel dimension, as shown in Formula 7:

$$P_2 = \text{PatchMerge}(F_1) \in R^{28\times28\times192} \tag{7}$$

Then, two Swin Transformer Blocks are used to further extract multi-scale features $F_2 \in R^{28\times28\times192}$. The features at this stage combine local details with a wider time-frequency region, which can enhance the characterization of global faults such as shaft misalignment or mechanical looseness.

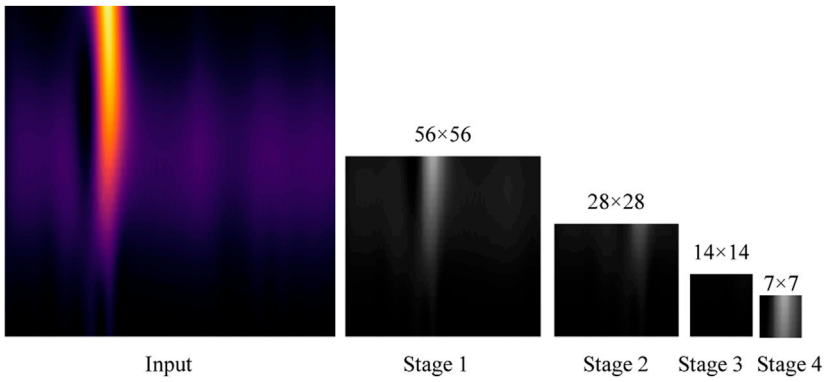

FIGURE 4
Feature visualization at different stages.

Stage 3 continues to use the Patch Merging operation to downsample the features to 14 × 14, expand the number of channels to 384, and stack 6 Swin Transformer Blocks to deepen the feature abstraction capability. The feature $F_3 \in R^{14\times14\times384}$ at this stage has strong semantic expression capabilities.

The final stage (Stage 4) uses the last Patch Merging operation to reduce the feature resolution to 7 × 7, expand the number of channels to 768, and stack 2 Swin Transformer Blocks to generate global features $F_4 \in R^{7\times7\times768}$.

The visualization of features at different stages is shown in Figure 4.

From Stage 1 to Stage 4, Swin Transformer abstracts fault features step by step through a hierarchical architecture, with feature granularity increasing from fine to coarse and the degree of abstraction gradually increasing. Shallow features focus on local detail characterization, while deep features focus on global semantic expression. Stage 1 captures local impulse responses and edge details in the time-frequency diagram, and Stage 2 enhances local region correlation through downsampling, revealing mid-level patterns such as low-frequency energy diffusion caused by shaft misalignment. The deep features of Stage 3 and Stage 4 focus on global spectrum distribution. This process constructs a multi-scale feature pyramid, and the coordinated optimization of shallow details and deep semantics realizes the separability of fault modes under complex working conditions.

3.3.2 Window multi-head attention mechanism

The core innovation of Swin Transformer lies in the application of a sliding window mechanism to calculate self-attention within a local window, significantly reducing the computational complexity. The traditional self-attention mechanism has significant limitations. The formula for the global self-attention mechanism is:

$$\mathrm{MSA}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{QK}^T}{\sqrt{d_k}}\right)\mathrm{V} \tag{8}$$

In Formula 8, $\mathrm{Q} = \mathrm{XW}_Q$, $\mathrm{K} = \mathrm{XW}_K$, and $\mathrm{V} = \mathrm{XW}_V$ are the query, key, and value matrices, respectively, and d_k is the dimension of the key. For high-resolution time-frequency graphs, this mechanism leads to excessive consumption of computing resources, limiting its application in real-time fault diagnosis.

Swin Transformer reduces computational complexity by dividing the input features into non-overlapping local windows and independently calculating self-attention in each window.

$$\text{W-MSA}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \bigoplus_w \mathrm{Softmax}\left(\frac{\mathrm{Q}_w\mathrm{K}_w^T}{\sqrt{d_k}} + B_w\right)\mathrm{V}_w \tag{9}$$

In Formula 9, $\oplus$ is the window feature concatenation operation. B_w is the relative position encoding matrix, which is used to model the relative spatial relationship of pixels in the window:

$$B_w(i,j) = \log\left(\frac{\| \Delta x_{ij}\| + 1}{\| \Delta x_{ij} \|}\right) \tag{10}$$

In Formula 10, Δx_{ij} is the coordinate difference between positions i and j. The application of relative position encoding enhances the model's sensitivity to the local time-frequency features of fault signals.

To compensate for the problem of cross-window information fragmentation caused by local windows, Swin Transformer applies a shift window mechanism. The original window division is maintained in the even layers. The window is shifted by one unit in the odd layers, and the attention is regrouped and calculated. The mask mechanism is used to avoid cross-window information leakage, and the formula is:

$$\mathrm{Attention}_{\mathrm{Shifted}} = \mathrm{MaskedSoftmax}\left(\frac{\mathrm{QK}^T}{\sqrt{d_k}} + B + M\right)\mathrm{V} \tag{11}$$

In Formula 11, M is a mask matrix, which is used to suppress information interaction between irrelevant windows. This design enables the model to achieve global feature interaction across windows while maintaining low computational complexity.

Swin Transformer adopts a multi-head attention mechanism to parallelly calculate the concatenated outputs of multiple heads and integrate multi-scale information through linear projection, as shown in Formula 12:

$$\mathrm{MultiHead} = \mathrm{Concat}(\text{W-MSA}_1,\ldots,\text{W-MSA}_\mathrm{h})\mathrm{W}_O \tag{12}$$

The feedforward network consists of two layers of multi-layer perceptrons, and the middle layer is expanded to 4 times the dimension to enhance the nonlinear expression ability, as shown in Formula 13:

$$\mathrm{FFN}(\mathrm{X}) = \mathrm{W}_2 \cdot \mathrm{GELU}(\mathrm{W}_1\mathrm{X} + \mathrm{b}_1) + \mathrm{b}_2 \tag{13}$$

3.4 Multi-scale feature fusion

3.4.1 Feature alignment

To achieve cross-stage feature interaction, the multi-scale features of the four-stage output of Swin Transformer are mapped to a unified dimensional space. The feature maps of all stages are aligned to 7 × 7×C (C is the uniform number of channels, which is set to 768 in this paper). The feature maps of Stages 1–3 are gradually downsampled to 7 × 7 using stride convolution.

Stage 1 → 7 × 7: two consecutive 3 × 3 convolutional layers (stride = 2) are used to downsample 56 × 56 × 96 to 14 × 14 × 96 and 7 × 7 × 96, denoted as $F_1^{'} \in \mathrm{R}^{7\times7\times96}$.
Stage 2 → 7 × 7: 28 × 28 × 192 is downsampled to 7 × 7 × 192 through a single 3 × 3 convolution layer (stride = 2), denoted as $F_2^{'} \in \mathrm{R}^{7\times7\times192}$.
Stage 3 → 7 × 7: a 3 × 3 convolution layer (stride = 2) is directly used to downsample 14 × 14 × 384 to 7 × 7 × 384, denoted as $F_3^{'} \in \mathrm{R}^{7\times7\times384}$.
Stage 4: no downsampling is required, and $F_4 \in \mathrm{R}^{7\times7\times768}$ is directly retained.

To unify the number of channels, 1 × 1 convolution is applied to the downsampled features of Stages 1–3, as shown in the Formula 14:

$$F_i^{''} = Conv\,1\times1\left(F_i^{'}\right) \in \mathrm{R}^{7\times7\times768}, i = 1,2,3 \tag{14}$$

3.4.2 Cross-Attention weighted fusion mechanism

The Cross-Attention mechanism dynamically quantifies the association weights between shallow detail features and deep abstract features through cross-stage Query-Key-Value interactions, and realizes adaptive fusion of multi-scale features. Query is generated by shallow features (Stages 1–3) to capture local detail information, and Key/Value is generated by deep features (Stage 4) to represent global semantic information.

The shallow feature $F^{'}{}_i^{'}$ and the deep feature F_4 are applied with learnable weight matrices to generate Query, Key, and Value, as shown in Formulas 15-17:

$$Q_i = F^{'}{}_i^{'} W_Q \tag{15}$$

$$K = F_4 W_K \tag{16}$$

$$V = F_4 W_V \tag{17}$$

Query, Key, and Value are split into 8 heads, and the attention weights are calculated respectively:

$$\mathrm{Attention_h}\left(\mathrm{Q}_i^{\mathrm{h}}, \mathrm{K}^{\mathrm{h}}, \mathrm{V}^{\mathrm{h}}\right) = \mathrm{Softmax}\left(\frac{\mathrm{Q}_i^{\mathrm{h}}\left(\mathrm{K}^{\mathrm{h}}\right)^T}{\sqrt{d_k}}\right)\mathrm{V}^{\mathrm{h}} \tag{18}$$

According to the Formula 18, the attention weight of each shallow feature $F_i^{''}$ and the deep feature F_4 is calculated, and weighted features are generated, as shown in the Formula 19:

$$\mathrm{A}_i = \mathrm{MultiHead}(\mathrm{Q}_i, \mathrm{K}, \mathrm{V}) \in \mathrm{R}^{7\times7\times768} \tag{19}$$

The features of the four stages are residually connected and normalized with the weighted features:

$$\mathrm{F_{fused}} = \mathrm{LayerNorm}\left(\mathrm{F}_4 + \sum_{i=1}^{3} \alpha_i \mathrm{A}_i\right) \tag{20}$$

In Formula 20, α_i is the learnable weight parameter, which is optimized by back propagation to dynamically allocate the contribution ratio of the shallow features.

In this study, the Cross-Attention mechanism was designed as an information query and retrieval process. Its core concept is that deep features (Stage 4) serve as high-level semantic understanding of global fault patterns, while shallow features (Stages 1–3) serve as raw observations of local signal details. Cross-Attention allows the query to search the key/value pairs for the most relevant "global explanation" that best explains the current local phenomenon, and then integrates these explanations based on weighted relevance.

3.5 Classification module

The global feature vector after multi-scale fusion is mapped to the probability distribution of the fault category, and the model parameters are optimized by supervised learning. Global pooling aims to compress high-dimensional feature maps into feature vectors of fixed length while retaining key feature information. This paper adopts global average pooling, and the formula is:

$$\mathrm{v}_c = \frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W} \mathrm{F_{fused}}(i, j, c) \tag{21}$$

In Formula 21, $\mathrm{F_{fused}}(i, j, c)$ is the value of the cth channel of the fused feature map at position (i, j). v_c represents the global average of the cth channel. The local noise is suppressed by the averaging operation, and the spatial statistical characteristics of the feature map are preserved.

The feature vector after global pooling may contain redundant information and needs further dimension reduction and enhancement. This paper designs a submodule including normalization, activation function, and fully connected layer. The feature vector is standardized to accelerate training and improve stability:

$$\hat{v}_c = \frac{\mathbf{v}_c - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} \tag{22}$$

In Formula 22, μ_c and σ_c^2 are the mean and variance of channel c, and ϵ is a small constant to prevent division by zero.

Nonlinear transformation is applied to enhance the model expression ability, as shown in the Formula 23:

$$\tilde{v}_c = max(0, \hat{v}_c) \tag{23}$$

TABLE 3 Comparison model information.

Model	Core content	Parameter quantity (M)	Applicable scenarios
ViT	Divide the image into blocks and input them into transformer to model global dependencies	86M	High-resolution time-frequency graph modeling
FocalNet	Dynamically focus on distant areas in the local receptive field	89M	Multi-scale feature extraction and non-uniform signal analysis
ConvNeXt-V2	Use self-supervised pre-training to combine the locality of convolution with the efficient training strategy of transformer	50M	Self-supervised learning and industrial deployment optimization
ContextCluster	Generate "context groups" through clustering, and then apply transformer blocks to each group	91M	Dynamic grouping modeling of non-uniformly distributed signals
MaxViT	Combining local block attention with global attention	78M	Multi-scale modeling and small sample fault classification
Swin transformer + CrossViT	CrossViT for cross-stage interaction	110M	Cross-stage feature fusion and multi-modal interaction
MobileViT	Combining lightweight convolution and transformer modules to improve mobile inference efficiency	28M	Edge fault diagnosis and low-power deployment

The feature vector is mapped to the fault category space, and the number of categories is K (K = 8 in this paper). The fully connected layer is defined as:

$$z = W_{fc}\tilde{v} + b_{fc} \in R^{K} \tag{24}$$

In Formula 24, W_{fc} is the learnable weight matrix, and b_{fc} is the bias vector.

The probability distribution of the fault category is calculated by the Softmax classifier, as shown in the Formula 25:

$$p(y = k|\mathbf{z}) = \frac{e^{z_k}}{\sum_{k'=1}^{K} e^{z_{k'}}}, k = 1, 2, \ldots, K \tag{25}$$

To supervise the model training, the cross entropy loss function is used to measure the difference between the predicted probability distribution and the true label. The true label is a one-hot encoded vector, and the loss function is defined as the Formula 26:

$$L_{CE} = -\sum_{k=1}^{K} y_{true,k} \log(p_k) \tag{26}$$

The model parameters are optimized by back-propagation to minimize the loss:

$$\theta^{*} = \arg\min_{\theta} \left(L_{CE} + \lambda \parallel \theta \parallel_2^2\right) \tag{27}$$

In Formula 27, λ is the L2 regularization coefficient, which is used to prevent overfitting.

4 Experimental design

4.1 Experimental setup

All experiments in this study are conducted in a unified experimental environment and training configuration to ensure the comparability and reproducibility of the results. The experimental hardware configuration is: NVIDIA A100 (4 cards in parallel, 80 GB video memory per card), Intel Xeon Gold 6338 (2.00 GHz × 2), and 512 GB memory to ensure efficient reading and caching of large-scale datasets. The software environment is based on the PyTorch 2.0 framework; the optimizer uses AdamW (weight decay coefficient is 10^{-4}); the learning rate scheduling uses the cosine annealing strategy (initial learning rate is 10^{-4}); the batch size is set to 64 to balance video memory usage and training efficiency; the number of training rounds is set to 50; the early stopping mechanism is used to monitor the validation set loss (training is terminated if there is no improvement after 10 consecutive rounds).

To systematically verify the effectiveness of the CWT-Swin Transformer-Cross-Attention model proposed in this paper in the fault diagnosis task of electric pumping electromechanical equipment, 7 advanced models in the current field of vision and signal processing are selected as comparison objects. These models cover mainstream technical directions such as multi-scale modeling, lightweight design, and cross-stage interaction, including: ViT, FocalNet (Focal Modulation Network), ConvNeXt-V2, ContextCluster, MaxViT (Multi-Axis Vision Transformer), Swin Transformer + CrossViT combined model, and MobileViT (Mobile Vision Transformer). All comparison models accept a unified 224 × 224 × 3 time-frequency graph input (generated by CWT and replicated three-channel adaptation), and are trained and tested under the same data partitioning, optimization strategy, and evaluation indicators.

The comparison model information is described in Table 3.

4.2 Performance indicators

To comprehensively evaluate the performance of the proposed model in the fault diagnosis task of large-scale electric pumping electromechanical equipment, multi-dimensional evaluation indicators are used, including Accuracy, macro precision, macro recall, macro F1, ROC (Receiver Operating Characteristic), etc., to quantify the model performance from multiple perspectives such as overall classification ability, balance between categories, and sensitivity of model discrimination threshold.

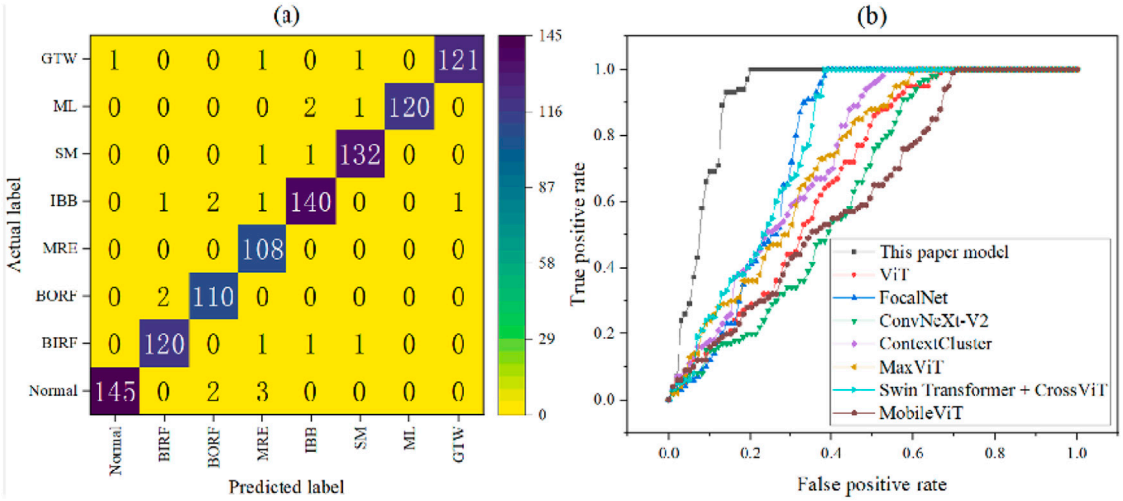

FIGURE 5
Binary classification performance of various types of faults and normal-fault. **(a)** Confusion matrix. **(b)** ROC curve.

Accuracy is the most intuitive performance indicator, and the formula is as the Formula 28:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{28}$$

Macro precision measures the consistency of the model's predictions for all categories. It is calculated by first calculating the precision of each category and then taking the average. The single-category precision is defined as:

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k} \tag{29}$$

In Formula 29, k represents the *k*th category of fault.

Macro precision is the average precision of all categories, as shown in the Formula 30:

$$\text{Macro Precision} = \frac{1}{K}\sum_{k=1}^{K}\text{Precision}_k \tag{30}$$

Macro recall is the average recall of all categories, as shown in the Formula 31:

$$\text{Macro Recall} = \frac{1}{K}\sum_{k=1}^{K}\text{Recall}_k \tag{31}$$

Macro F1-Score is the harmonic mean of macro precision and macro recall, which is used to balance the impact of the two. The single-class F1-Score is defined as the Formula 32:

$$\text{F1}_k = 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{32}$$

The macro F1-Score is the average of the F1 values of all classes, as shown in the Formula 33:

$$\text{Macro F1} = \frac{1}{K}\sum_{k=1}^{K}\text{F1}_k \tag{33}$$

5 Results

5.1 Fault classification accuracy

The confusion matrix reflects the classification performance of the proposed model for each type of sample, and the ROC curve is used to compare the normal-fault binary classification capabilities of different models, as depicted in Figure 5.

There are 150 normal samples in the test set, with 145 correctly classified samples, 2 incorrectly classified as BORF, and 3 misclassified as MRE. There are a total of 1019 samples in the test set, and 996 samples are completely correctly classified. Figure 5b compares the ability of the proposed model to distinguish between normal and faulty models. The area under the curve of the proposed model is the largest, and it is closest to the upper left point, which shows that in the normal-fault binary classification task, the proposed model has higher performance than the comparison model.

The ROC curve of the proposed model is closer to the upper left corner, indicating that the model has a higher true positive rate and a lower false positive rate at different thresholds, and has stronger robustness in discrimination. This advantage mainly comes from the fact that after CWT maps the original vibration signal into a two-dimensional time-frequency tensor, it significantly enhances the ability to express local mutation features of non-stationary signals. Combining the hierarchical feature pyramid of Swin Transformer and the dynamic weighted fusion mechanism of Cross-Attention, the model can achieve efficient information interaction between local details and global semantics, which not only retains the shallow transient impact response, but also enhances the discrimination ability of deep abstract features, thereby improving the sensitivity to weak fault signals under complex working conditions. In addition, through the attention mechanism in the unified dimensional space, the model effectively alleviates the redundant modeling problem of traditional Transformer in cross-stage feature transfer, so that multi-

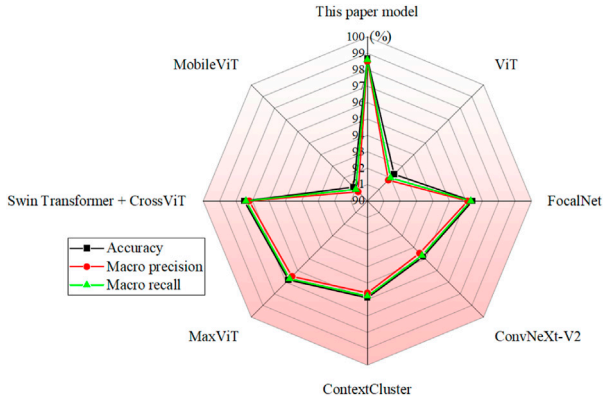

FIGURE 6
Overall classification performance of different models.

scale information is reasonably distributed in the decision-making process.

To more comprehensively analyze the fault classification performance of the proposed model in large-scale electric pumping electromechanical equipment, it is evaluated by accuracy, macro precision, and macro recall. Figure 6 illustrates the overall classification performance of different models.

The proposed model is significantly better than other comparison models in terms of accuracy, macro precision, and macro recall, reaching 98.7%, 98.5%, and 98.6%, respectively. In contrast, the second-best model is Swin Transformer + CrossViT, with an accuracy of 97.5%, while the worst model is the lightweight MobileViT, with an accuracy of only 91.2%.

The excellent performance of the proposed model is mainly due to the following key mechanisms:

1. CWT time-frequency feature extraction: the vibration signal is mapped to a time-frequency two-dimensional tensor through complex Morlet wavelet, which effectively captures the transient impact characteristics of non-stationary signals and provides high-quality structured input for subsequent deep learning.
2. Swin Transformer's hierarchical feature pyramid: through the sliding window attention mechanism, multi-scale features are gradually abstracted, which retains local detail information, constructs a global semantic representation, and realizes feature interaction from shallow to deep layers.
3. Cross-Attention multi-scale fusion: by dynamically weighting and fusing shallow and deep features, the problem of cross-stage information fragmentation in traditional Transformer is solved, and the expression ability of key fault features is enhanced.
4. Adaptive attention mechanism: by quantifying the importance of features of different scales through Cross-Attention, the limitations of single-scale features are avoided, and the robustness of the model to complex working conditions is improved.

The proposed model shows significantly better classification performance than the comparative model in fault diagnosis tasks. Its core advantages are reflected in multi-scale feature modeling and cross-stage information fusion capabilities. The proposed model is ahead of other models in all three indicators, indicating that it has obvious advantages in category identification consistency and comprehensive coverage. In contrast, ViT is limited by the global attention mechanism, and there are problems of redundant calculation and local detail loss when processing non-stationary signals, resulting in its accuracy rate of only 92.3%. Although FocalNet enhances the multi-scale modeling capability by focusing attention locally, it is still slightly insufficient in the scenario of category imbalance, with macro precision and macro recall rates of 96.1% and 96.3%, respectively. MaxViT and ContextCluster perform well with 96.8% and 95.9% accuracy, respectively, but their interaction modeling between shallow features and deep semantics is still not sufficient. The Swin Transformer + CrossViT combined model is closest to this paper in terms of structural design, with an accuracy of 97.5%, but still lower than the model in this paper. The main difference comes from the CWT time-frequency analysis applied in this paper, which further improves the local expression ability of the input signal, making it easier for the model to capture key fault features in periodic shocks and broadband noise. ConvNeXt-V2 and MobileViT are relatively inferior in modeling complex electromechanical signals because of their lightweight architecture. In summary, the model in this paper generates high-quality time-frequency graphs through CWT mapping, combines the hierarchical feature extraction of Swin Transformer with the dynamic weighted fusion mechanism of Cross-Attention, effectively solves the limitations of traditional methods in multi-scale information mining and cross-stage feature interaction, and thus achieves the best performance in fault identification tasks under complex working conditions of electric pumping equipment, reflecting stronger discrimination ability and engineering applicability.

5.2 Model convergence and stability

The convergence of the model is analyzed by analyzing the loss reduction and accuracy improvement curves of the model in this paper, and the stability analysis is performed through 10-fold cross-validation, as displayed in Figure 7.

The epoch in this paper is set to 50 times. When epoch is 1, the accuracy and loss values are 70.9% and 0.723, respectively. As the training level deepens, the accuracy shows an overall upward trend, while the loss value begins to decrease. The model in this paper shows good convergence. The performance is stable at the 41st epoch, with the accuracy increased to 98.7% and the loss value decreased to 0.017, indicating that the model can quickly learn key fault features and converge. This feature is due to the high-quality time-frequency input provided by CWT and the hierarchical attention mechanism of Swin Transformer, which enables the model to establish efficient multi-scale feature representation in fewer training rounds, improving optimization efficiency and stability.

From the 10-fold cross-validation results, the model in this paper shows high stability and consistency in various evaluation indicators. The Kappa coefficient is stable between 0.972 and 0.979, indicating that the model has excellent consistency and robustness

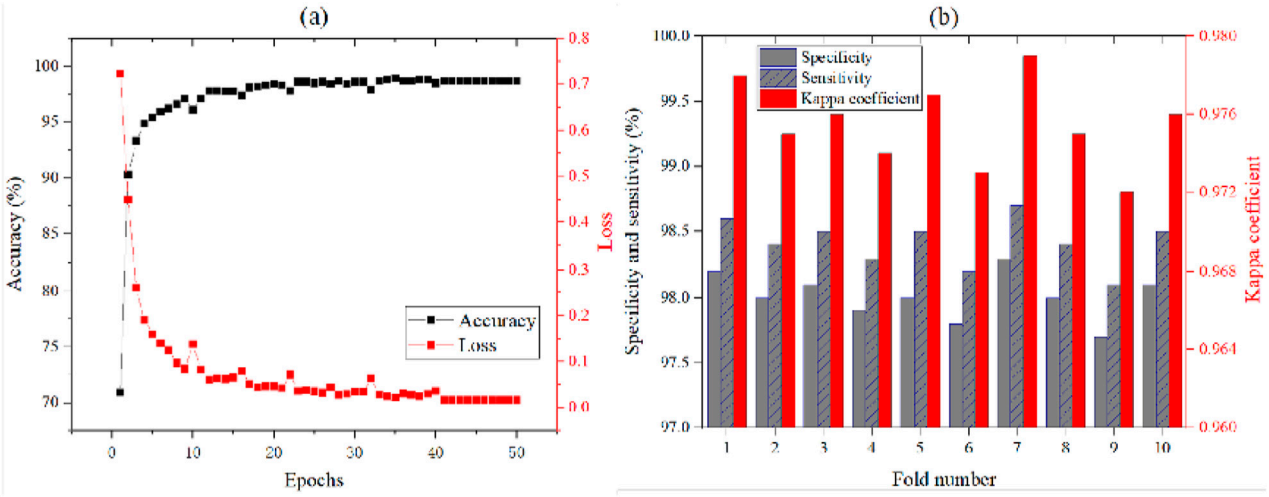

FIGURE 7
Model convergence and stability analysis. **(a)** Model convergence. **(b)** 10-fold cross-validation results.

TABLE 4 10-fold cross-validation performance statistics.

Evaluation metric	Minimum	Maximum	Average	Standard deviation
Accuracy	97.50%	98.90%	98.20%	0.38%
Macro-precision	97.30%	98.80%	98.00%	0.41%
Macro-recall	97.40%	98.70%	98.10%	0.35%
Macro-F1	97.20%	98.60%	97.90%	0.37%

under different data partitions; the specificity and sensitivity are maintained in the range of 97.7%–98.3% and 98.1%–98.7%, respectively, with very small fluctuations, indicating that the model has a balanced ability to identify normal and faulty samples without significant deviation. This stable performance is mainly due to the structured time-frequency input provided by CWT and the effective modeling and fusion mechanism of multi-scale features by the Swin Transformer-Cross-Attention architecture, which enables the model to adapt to signal changes under different working conditions and has good generalization ability and engineering implementation potential.

The 10-fold cross-validation performance statistics are shown in Table 4.

Across 10 different training-test splits, the standard deviation of each core metric (accuracy, macro precision, macro recall, and macro F1) was less than 0.5%. This demonstrates that the model's performance fluctuates minimally across different data subsets, demonstrating high stability and consistency. This demonstrates that the model's superior performance is not due to overfitting to a specific data split, but rather its strong generalization capabilities even with limited data. This result strongly addresses concerns about the balance between data volume and model complexity, validating the effectiveness of the high-quality time-frequency input provided by CWT and the Swin Transformer-Cross-Attention architecture, enabling the model to learn transferable fault features from limited samples efficiently.

5.3 Multi-scale feature fusion effect

t-SNE visualization is used to show the difference in feature expression ability with/without Cross-Attention processing. The t-SNE visualization results are shown in Figure 8.

Figure 8a shows the feature distribution extracted by the model in this paper (including Cross-Attention). The sample points of each fault category show stronger intra-class clustering and inter-class separability. The points of the same color are highly concentrated, indicating that Cross-Attention effectively enhances the expression of key features. In contrast, Figure 8b shows the feature splicing method without Cross-Attention. There is obvious overlap between its categories, especially the blurred boundaries between categories such as MRE and GTW, indicating that the cross-stage information interaction ability is weak. This comparison fully verifies the effectiveness of Cross-Attention in multi-scale feature fusion and enhances the model's ability to distinguish weak fault features under complex working conditions.

5.4 Robustness test under noise conditions

White noise is applied into the vibration signal to test the robustness under noise conditions. The results are shown in Figure 9.

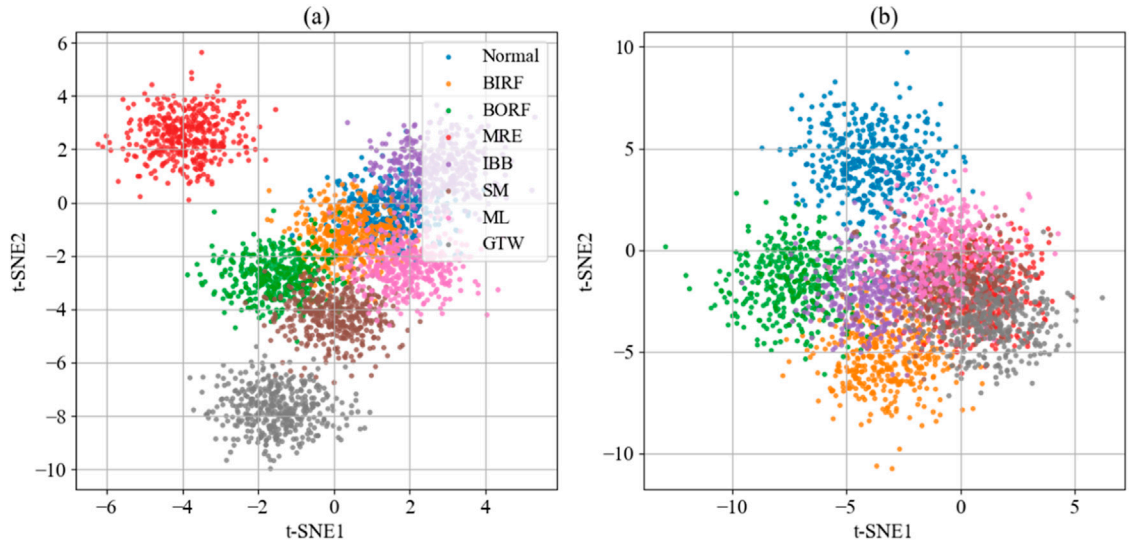

FIGURE 8
t-SNE visualization. **(a)** With cross-attention. **(b)** Without cross-attention (feature splicing method).

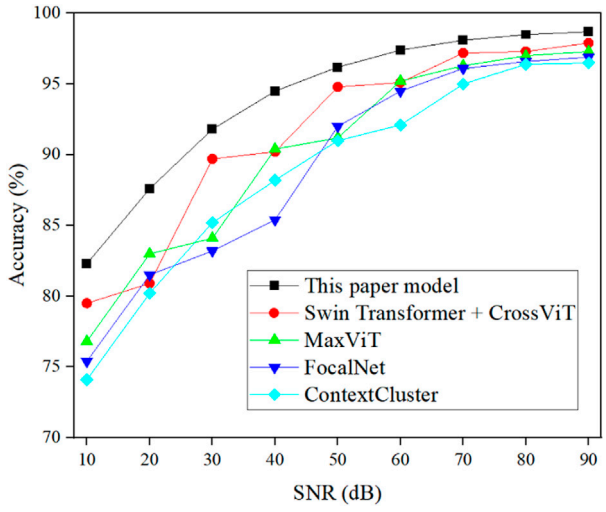

FIGURE 9
Robustness results under noise conditions.

From the change in accuracy under different SNRs, it can be learned that the proposed model shows significantly better robustness than the comparison model under noise interference. As the intensity of white noise decreases (that is, the SNR increases), the identification performance of all models improves, but the accuracy of the proposed model increases more steadily and always leads. When SNR = 10 dB, the proposed model still maintains an identification accuracy of 82.3%; Swin Transformer + CrossViT is 79.5%; the other models are all below 77%, indicating that the proposed model has a stronger adaptability to strong noise. This advantage is mainly attributed to the fact that CWT maps the vibration signal into a two-dimensional time-frequency tensor, which effectively enhances the localized feature expression ability of the signal, making it easier for the model to distinguish fault features from background noise; meanwhile, the Cross-Attention mechanism further improves the discrimination ability of key frequency band information by dynamically weighted fusion of multi-scale features, avoiding redundant interference of shallow features. In contrast, although other models such as FocalNet and MaxViT have certain noise resistance capabilities, they are limited in cross-stage feature interaction and non-stationary signal modeling. The experimental results fully verify that the proposed method has excellent generalization ability and engineering applicability under complex working conditions.

In order to comprehensively evaluate the robustness of the proposed method in a real industrial complex electromagnetic environment, this study superimposed three typical noises on the original vibration signal for simulation testing: Gaussian white noise, impulse noise, and sinusoidal interference. The test was carried out under strong interference conditions with a signal-to-noise ratio of 10 dB. Each noise type was tested independently 10 times and the average value was taken. The results are shown in Table 5.

Under strong interference of SNR = 10dB, the accuracy of the proposed model under Gaussian white noise, impulse noise and sinusoidal interference reached 82.3%, 79.8% and 81.5% respectively. The performance difference is due to the noise characteristics: impulse noise has the largest interference due to its similarity to the fault impact morphology; while the multi-resolution analysis of CWT can effectively disperse the energy of white noise, and the sliding window mechanism of Swin Transformer can suppress the frequency domain spikes of sinusoidal interference. Cross-Attention effectively distinguishes real faults with periodic patterns from random noise through deep semantic guidance and dynamic weighting of shallow features, ensuring the high-precision diagnosis ability of the model in complex electromagnetic environments.

TABLE 5 Comparison of model performance under various noise interferences.

Noise type	Accuracy	Macro-precision	Macro-recall	Macro-F1
Gaussian white noise	82.30%	81.90%	82.10%	81.70%
Impulse noise	79.80%	78.50%	79.20%	78.10%
Sine interference	81.50%	80.70%	81.00%	80.30%
Clean signal	98.70%	98.50%	98.60%	98.50%

TABLE 6 Ablation experiment results.

Modules	M1	M2	M3	M4
	This paper model	No cross-attention (feature concatenation)	CWT + transformer	Transformer
CWT	√	√	√	×
Swin transformer	√	√	×	×
Cross-attention	√	×	×	×
Accuracy (%)	98.7	97.7	96.4	95.6
Macro F1 (%)	98.5	96.2	95.7	94.9
Gaussian white noise (accuracy)	82.3	81.2	79.4	78.5
Impulse noise (accuracy)	79.8	78.5	76.8	74.5
Sine interference (accuracy)	81.5	80.9	78.8	77.6

The strong robustness of the model in this paper comes from the synergy of CWT, Swin Transformer and Cross-Attention. First, CWT disperses the broadband Gaussian noise energy to different scales through multi-resolution analysis, and focuses the transient pulse energy of the fault, realizing the preliminary "denoising" in the time-frequency domain. Secondly, the hierarchical structure of Swin Transformer suppresses noise in feature abstraction: the shallow layer captures local details, and the deep layer filters out incoherent noise responses by learning the spatiotemporal consistency of fault features. Finally, Cross-Attention, as a decision optimizer, gives higher weights to reliable shallow features (Query) with high correlation through global semantic guidance of deep features (Key/Value), while suppressing false features contaminated by noise, ensuring the accuracy of the final classification decision.

5.5 Ablation experiment results

To explore the role of important modules of the proposed model, an ablation experiment is set up, and the findings are displayed in Table 6.

Table 6 shows the performance changes of the proposed model when different modules are missing, verifying the key role of CWT, Swin Transformer, and Cross-Attention modules in fault diagnosis performance. M1 is a complete model, and its accuracy and Macro F1 reach 98.7% and 98.5%, respectively, which are significantly better than the other ablation models, indicating that the proposed method has excellent feature expression and classification capabilities. After removing the Cross-Attention module, M2 uses feature concatenation to fuse multi-scale features. Its accuracy drops to 97.7%, and Macro F1 drops to 96.2%, indicating that the Cross-Attention mechanism plays a key role in the interaction of multi-scale features by dynamically weighted fusion of shallow details and deep semantic information, effectively improving the model's discrimination ability. M3 further removes the Swin Transformer structure and only retains CWT and ordinary Transformer. Its accuracy is 96.4%, and Macro F1 is 95.7%. The performance continues to decline, reflecting the importance of Swin Transformer in hierarchical feature extraction and local-global modeling. Its sliding window attention mechanism significantly enhances the model's adaptability to non-stationary signals under complex working conditions. As the simplest model, M4 only uses the standard Transformer, with an accuracy of only 95.6% and Macro F1 of 94.9%, indicating that the lack of CWT's time-frequency localization analysis and hierarchical architecture design can significantly weaken the model performance. In summary, CWT provides high-quality structured input; Swin Transformer builds an efficient feature pyramid; Cross-Attention realizes the effective fusion of cross-stage features. The three work together to improve the overall performance of the model in the fault diagnosis task of electric pumping electromechanical equipment.

The performance of M1, M2, M3, and M4 in distinguishing normal/faulty is shown in Figure 10.

From the (Normal/Fault) confusion matrix, it can be learned that M1 has the highest number of correct classifications of normal

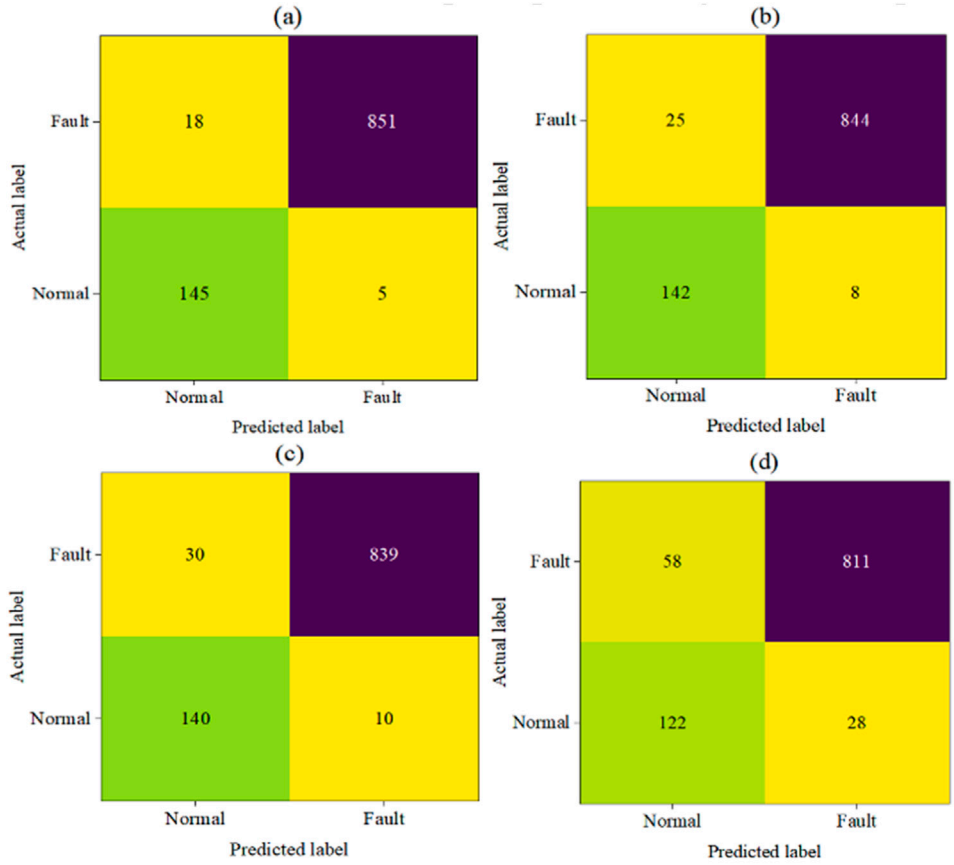

FIGURE 10
(Normal/Fault) Confusion matrix. **(a)** M1. **(b)** M2. **(c)** M3. **(d)** M4.

and faulty, with 145 correct samples for normal classification and 851 correct samples for fault classification. The correct samples for normal classification of M2, M3, and M4 are 142, 140, and 122, respectively; the correct samples for fault classification are 844, 839, and 811, respectively. The fault classification advantage of M1 mainly comes from the effective fusion capability of the Cross-Attention mechanism, which enables the model to precisely capture key fault features and suppress redundant information; at the same time, the hierarchical feature extraction of Swin Transformer and the time-frequency localization analysis provided by CWT further enhance the structural expression of the input signal and improve the sensitivity to weak fault signals. In contrast, M2 (without Cross-Attention), M3 (without Swin Transformer), and M4 (only Transformer) have more misjudgments in normal and fault classification due to the lack of multi-scale dynamic fusion or hierarchical modeling capabilities, verifying the rationality and necessity of the model module design in this paper.

5.6 Computational efficiency

Evaluating computational efficiency is crucial to achieving lightweight deployment. Low latency and high throughput are prerequisites for real-time fault diagnosis. By analyzing FLOPs and inference time, models suitable for edge devices can be screened to ensure that diagnostic results can be fed back instantly, meet the strict requirements of industrial sites for response speed, and promote intelligent operation and maintenance from theory to practical application. The comparison of computational efficiency is shown in Table 7.

This paper model achieves a good balance between accuracy and efficiency. Its FLOPs are 4.3G and its inference time is 15.2 m, outperforming global attention models such as ViT and FocalNet. This is due to the sliding window mechanism of the Swin Transformer. Although the lightweight MobileViT is the most efficient, its accuracy is significantly lower than our model. Our model has a computational overhead approximately 3–4 times that of MobileViT, yet it still meets real-time requirements on modern edge GPUs. This demonstrates that our model not only leads in accuracy but also has the potential for deployment on edge devices, providing a viable solution for high-precision real-time diagnosis.

5.7 Model calibration evaluation

To systematically evaluate the reliability of the model output probability, especially for fault types with similar time-frequency

TABLE 7 Computational efficiency comparison results.

Model	Parameters (M)	FLOPs (G)	Single-sample inference time (ms)
This paper model	92.1	4.3	15.2
ViT	86	15.7	48.5
FocalNet	89	14.2	42.3
ConvNeXt-V2	50	8.9	26.8
ContextCluster	91	16.1	50.1
MaxViT	78	12.5	38.7
Swin transformer + CrossViT	110	18.3	55.6
MobileViT	28	0.56	3.8

TABLE 8 Model calibration evaluation results.

Evaluation metrics	This paper model	ViT	FocalNet
Expected calibration error	2.8%	6.5%	5.1%
Percentage of high-confidence samples ($p > 0.85$)	92.4%	85.1%	88.3%
Accuracy of high-confidence samples ($p > 0.85$)	97.1%	93.2%	94.8%
MRE high-confidence accuracy ($p > 0.85$)	98.3%	91.7%	93.5%
GTW high-confidence accuracy ($p > 0.85$)	97.9%	92.1%	94.2%

characteristics such as MRE and GTW, this study uses expected calibration error and high confidence accuracy for evaluation. The results are shown in Table 8.

The predicted probability was divided into 10 intervals (0–0.1, 0.1–0.2, ., 0.9–1.0), and the actual accuracy of samples within each interval was calculated. This study conducted an in-depth evaluation of the model's calibration performance through expected calibration error and high-confidence analysis. The results showed that the expected calibration error of our model was only 2.8%, lower than ViT (6.5%) and FocalNet (5.1%), indicating that its predicted probability and true accuracy were highly consistent. A key finding was that 92.4% of samples received high-confidence predictions ($p > 0.85$), and the accuracy of these samples reached 97.1%, demonstrating that the model can provide reliable judgments in the vast majority of cases. In particular, for the easily confused MRE and GTW faults, when the model predicted with high confidence ($p > 0.85$), the accuracy reached 98.3% and 97.9%, respectively. This is due to the Cross-Attention mechanism's dynamic weighting of key discriminant features (such as the low-frequency energy diffusion of MRE and the harmonic pattern of GTW), which enables the model to not only correctly classify, but also make accurate assessments of the credibility of the classification results, providing a solid basis for intelligent decision-making in industrial scenarios.

6 Conclusion

This study presents a high-precision fault diagnosis framework for large-scale electromechanical equipment by integrating CWT, Swin Transformer, and Cross-Attention, achieving 98.7% accuracy and superior noise robustness. The method leverages optimized complex Morlet wavelet for refined time-frequency representation, hierarchical Swin Transformer for multi-scale feature extraction, and Cross-Attention for dynamic fusion of shallow and deep features, effectively enhancing feature discriminability and model reliability. Experimental results validate its effectiveness and potential for intelligent operation and maintenance in complex industrial environments.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JC: Formal Analysis, Data curation, Writing – original draft, Writing – review and editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmad, Z., Kim, J. Y., and Kim, J. M. (2023). A technique for centrifugal pump fault detection and identification based on a novel fault-specific mann–whitney test. *Sensors* 23 (22), 9090. doi:10.3390/s23229090

Benameur, R., Dahane, A., Kechar, B., and Benyamina, A. (2024). An innovative smart and sustainable low-cost irrigation system for anomaly detection using deep learning. *Sensors* 24 (4), 1162. doi:10.3390/s24041162

Bousseksou, R., Bessous, N., Elzein, I. M., Mahmoud, M., Ma'arif, A., Touti, E., et al. (2025). Utilizing short-time fourier transform for the diagnosis of rotor bar faults in induction motors under direct torque control. *Int. J. Robotics Control Syst.* 5 (2), 1441–1457. doi:10.31763/ijrcs.v5i2.1886

Cao, X., Hu, L., Peng, W., Wu, X., Xiang, J., Ding, W., et al. (2024). Coal foreign body detection method based on cross-modal attention fusion. *J. Mine Automation* 50 (1), 57. doi:10.13272/j.issn.1671-251x.2023110035

Chen, S., Wang, H., Zhang, H., Peng, C., Li, Y., Wang, B., et al. (2024). A novel method of swin transformer with time-frequency characteristics for ECG-based arrhythmia detection. *Front. Cardiovasc. Med.* 11, 1401143. doi:10.3389/fcvm.2024.1401143

Dexian, H. (2025). Analysis of mechanical and electrical equipment fault diagnosis and safety maintenance technology. *Hydroelectr. Sci. and Tecnology* 8 (2), 83–86. doi:10.33142/hst.v8i2.15571

Djaballah, S., Meftah, K., Khelil, K., and Sayadi, M. (2023). Deep transfer learning for bearing fault diagnosis using CWT time–frequency images and convolutional neural networks. *J. Fail. Analysis Prev.* 23 (3), 1046–1058. doi:10.1007/s11668-023-01645-4

Du, K., and Shaohui, N. (2023). Research on rolling bearing fault diagnosis based on ADASYN and swin transformer. *Mach. Tool and Hydraulics* 51 (15), 209. doi:10.3969/j.issn.1001-3881.2023.15.034

Fan, C., Peng, Y., Shen, Y., Guo, Y., Zhao, S., Zhou, J., et al. (2024). Variable scale multilayer perceptron for helicopter transmission system vibration data abnormity beyond efficient recovery. *Eng. Appl. Artif. Intell.* 133, 108184. doi:10.1016/j.engappai.2024.108184

Fu, W., Zheng, B., Li, S., Liao, W., Huang, Y., Chen, X., et al. (2024). Batch channel normalized-CWGAN with swin transformer for imbalanced data fault diagnosis of rotating machinery. *Meas. Sci. Technol.* 36 (1), 016207. doi:10.1088/1361-6501/ad8673

Ghafouri Matanagh, A., Ozturk, S. B., Goktas, T., and Hegazy, O. (2024). Classifying the percentage of broken magnets in permanent magnet synchronous motors using combined short-time fourier transform and a pre-trained convolutional neural network. *Energies* 17 (2), 368. doi:10.3390/en17020368

Hatsey, N. H., and Birkie, S. E. (2021). Total cost optimization of submersible irrigation pump maintenance using simulation. *J. Qual. Maintenance Eng.* 27 (1), 187–202. doi:10.1108/jqme-08-2018-0064

Jier, Q., Hongfeng, T., Long, C., and Lingzhi, S. (2022). Bearing fault diagnosis based on self-attention mechanism assisted classification generative adversarial network. *Inf. Control* 51 (6), 753–762. doi:10.13976/j.cnki.xk.2022.2002

Karagiovanidis, M., Pantazi, X. E., Papamichail, D., and Fragos, V. (2023). Early detection of cavitation in centrifugal pumps using low-cost vibration and sound sensors. *Agriculture* 13 (8), 1544. doi:10.3390/agriculture13081544

Kumar, S. R., Ali, M. S. W., Pandian, C. K. A., and Muralidharan, V. (2024). Condition monitoring of electric vehicle motor testing machine's vital components using bagged trees and quadratic SVM: a comparative study. *Eng. Res. Express* 6 (2), 025531. doi:10.1088/2631-8695/ad476e

Li, X., Li, M., Liu, B., Lv, S., and Liu, C. (2024). A novel transformer network based on cross–spatial learning and deformable attention for composite fault diagnosis of agricultural machinery bearings. *Agriculture* 14 (8), 1397. doi:10.3390/agriculture14081397

Liu, X., Guo, C., Xunkai, W., Liu, Y., Wang, H., He, Z., et al. (2024). Early fault warning method for rolling bearings based on wavelet analysis and convolutional neural network. *J. Aerosp. Power* 39 (9), 20220622. doi:10.13224/j.cnki.jasp.20220622

Luczak, D. (2024). Machine fault diagnosis through vibration analysis: continuous wavelet transform with complex morlet wavelet and time–frequency RGB image recognition via convolutional neural network. *Electronics* 13 (2), 452. doi:10.3390/electronics13020452

Mallak, A., and Fathi, M. (2021). Sensor and component fault detection and diagnosis for hydraulic machinery integrating LSTM autoencoder detector and diagnostic classifiers. *Sensors* 21 (2), 433. doi:10.3390/s21020433

Ravikumar, S., Sharik, N., Syed, S., Muralidharan, V., and Kumar, P. (2025). A comparative analysis of fault diagnosis by vibration signals for critical gear components in electric vehicle motor testing machines using machine learning algorithms. *SAE Tech. Pap.* 1. doi:10.4271/2025-01-0040

Ribeiro Junior, R. F., dos Santos Areias, I. A., Campos, M. M., Teixeira, C., da Silva, L. E. B., and Gomes, G. F. (2022). Fault detection and diagnosis in electric motors using convolution neural network and short-time fourier transform. *J. Vib. Eng. and Technol.* 10 (7), 2531–2542. doi:10.1007/s42417-022-00501-3

Silik, A., Noori, M., Altabey, W. A., Ghiasi, R., and Wu, Z. (2021). Comparative analysis of wavelet transform for time-frequency analysis and transient localization in structural health monitoring. *Struct. Durab. and Health Monit.* 15 (1), 1. doi:10.32604/sdhm.2021.012751

Song, H., Yuan, L., and Shuangquan, G. (2024). Grey wolf optimization algorithm based on data enhancement and feature attention mechanism - optimized residual neural network transformer fault diagnosis method. *Mod. Electr. Power* 41 (2), 392–400. doi:10.19725/j.cnki.1007-2322.2022.0163

Trejo-Chavez, O., Cruz-Albarran, I. A., Resendiz-Ochoa, E., Salinas-Aguilar, A., Morales-Hernandez, L., Basurto-Hurtado, J. A., et al. (2023). A CNN-based methodology for identifying mechanical faults in induction motors using thermography. *Machines* 11 (7), 752. doi:10.3390/machines11070752

Ullah, I., Khan, R. U., Yang, F., and Wuttisittikulkij, L. (2020). Deep learning image-based defect detection in high voltage electrical equipment. *Energies* 13 (2), 392. doi:10.3390/en13020392

Ventricci, L., Ribeiro Junior, R. F., and Gomes, G. F. (2024). Motor fault classification using hybrid short-time fourier transform and wavelet transform with vibration signal and convolutional neural network. *J. Braz. Soc. Mech. Sci. Eng.* 46 (6), 337. doi:10.1007/s40430-024-04890-2

Vokhidov, A., and Churakova, E. (2021). Problems of ensuring the operation of load units of irrigation stations and large mechanical units. *J. Balkan Tribol. Assoc.* 27 (5), 838.

Wang, X., and Zhang, G. (2023). Installation and maintenance methods of mechanical and electrical equipment in water conservancy pumping station. *Hydroelectr. Sci. and Tecnology* 6 (3), 111–113. doi:10.33142/hst.v6i3.8545

Wang, R., Dong, E., Cheng, Z., Liu, Z., and Jia, X. (2024). Transformer-based intelligent fault diagnosis methods of mechanical equipment: a survey. *Open Phys.* 22 (1), 20240015. doi:10.1515/phys-2024-0015

Wang, S., Jia, Z., Li, Y., and Yang, Q. (2024). Research on fault diagnosis for three-phase rectifier device in direct current transmission system based on continuous wavelet transform and vision transformer. *Meas. Sci. Technol.* 35 (12), 126016. doi:10.1088/1361-6501/ad7e3f

Wu, X., He, R., and Chen, M. (2025). Fault diagnosis of agricultural sprinkler irrigation machinery equipment based on machine vision. *Nonlinear Eng.* 14 (1), 20240052. doi:10.1515/nleng-2024-0052

Xia, H., and Hong, X. (2022). Research on technical measures for installation and maintenance of electromechanical equipment in pumping stations in water conservancy projects. *Hydroelectr. Sci. and Tecnology* 5 (5), 60–62.

Xiang, C., Liu, Q., and Hu, J. (2024). Bearing fault diagnosis and life prediction based on attention mechanism and long short-term memory network under multi-source sensor data. *Inf. Control* 53 (2), 211–225. doi:10.13976/j.cnki.xk.2023.3056

Xiaofeng, L. I. U. (2025). Fault diagnosis and maintenance of electromechanical equipment of coal mine roadheader. *Coal Mine Mod.* 34 (1), 47–51. doi:10.13606/j.cnki.37-1205/td.2025.01.009

Xie, F., Wang, Y., Wang, G., Sun, E., Fan, Q., Song, M., et al. (2024). Fault diagnosis of rolling bearings in agricultural machines using SVD-EDS-GST and ResViT. *Agriculture* 14 (8), 1286. doi:10.3390/agriculture14081286

Xin, J., Tao, G., Tang, Q., Zou, F., and Xiang, C. (2024). Structural damage identification method based on swin transformer and continuous wavelet transform. *Intell. and Robotics* 4 (2), 200–215. doi:10.20517/ir.2024.13

Xu, L., Teoh, S. S., and Ibrahim, H. (2024). A deep learning approach for electric motor fault diagnosis based on modified InceptionV3. *Sci. Rep.* 14 (1), 12344. doi:10.1038/s41598-024-63086-9

Zeng, F., Ren, X., and Wu, Q. (2024). A fault diagnosis method for motor vibration signals incorporating swin transformer with locally sensitive hash attention. *Meas. Sci. Technol.* 35 (4), 046121. doi:10.1088/1361-6501/ad1cc4

Zhang, C. (2024). Application of information technology in water conservancy project pumping station. *Hydroelectr. Sci. and Tecnology* 7 (7), 28–30.

Zhang, R., Liu, X., Zheng, Y., Lv, H., Li, B., Yang, S., et al. (2022). The double synchroextracting and complex shifted morlet wavelet-based application for bearing faults diagnosis under varying speed. *J. Vib. Eng. and Technol.* 10 (1), 131–147. doi:10.1007/s42417-021-00368-w

Zhou, J., Peng, Y., Shao, H., Shen, Y., Bin, G., Zheng, J., et al. (2025). Empirical fourier-bessel heuristic denoising and its application to gear fault diagnosis. *ISA Transactions*. doi:10.1016/j.isatra.2025.10.003