A Multimodal Lightweight Transformer for Bearing Fault Diagnosis Under High-Noise Industrial IoT Environments

Junjie Liu, Jiaxian Zhu, Weihua Bai, Huibing Zhang, Lianghai Wu, Teng Zhou, *Member, IEEE*, and Keqin Li, *Fellow, IEEE*

***Abstract*—Industrial IoT (IIoT) sensing nodes for bearing monitoring often operate in high-noise environments, where acoustic and mechanical interference masks weak fault signatures and undermines diagnostic reliability. To address this challenge, we propose a lightweight multimodal fusion framework for robust fault diagnosis under extreme noise. Our method first applies multiresolution decomposition with selective reconstruction and adaptive enhancement to preserve fault-related components while suppressing interference. The enhanced signals are transformed from 1-D time series into 2-D representations to jointly capture temporal dynamics and spectral characteristics. We then design a tri-branch multihead attention architecture that integrates a multiscale recurrence plot (MSRP) network, a Gramian-angular-field (GAF) network, and a lightweight residual network. Learnable attention weights enable adaptive fusion of complementary cross-modal features with low computational overhead. Extensive experiments on the CWRU benchmark show superior robustness from 0 to −6 dB signal-to-noise ratio (SNR), with a mean accuracy above 99.6% and consistent gains over eight state-of-the-art methods. Additional tasks on single-domain diagnosis, cross-condition fault type recognition, and fault degree discrimination (T1–T3) confirm strong generalization and multiscale adaptability, with average improvements of 2%–4% over the second-best baseline and stable variance under noise. The compact architecture and high noise immunity indicate practical suitability for IIoT sensing nodes and edge deployment in complex industrial scenarios.**

***Index Terms*—Attention mechanism, discrete wavelet transform (DWT), fault diagnosis, multimodal fusion, multiscale, rolling bearing.**

I. Introduction

ROLLING bearings constitute the linchpin of rotating machinery, governing both transmission precision and load-bearing capacity while directly dictating system reliability and operational lifespan [1]. Statistical analyses reveal that bearing failures precipitate 40%–90% of all mechanical breakdowns, emerging as the predominant culprit behind equipment downtime and safety incidents across critical industrial applications [2]. This vulnerability manifests most acutely in demanding scenarios such as wind turbine drivetrains, CNC machine tool spindles, and aerospace propulsion systems [3], [4]. The diagnostic challenge intensifies as bearing vibration signals are invariably corrupted by broadband noise stemming from electromagnetic interference, structural resonance, and stochastic impacts inherent to industrial environments [5], [6]. Consequently, the pursuit of high-fidelity bearing health assessment under extreme noise conditions (signal-to-noise ratio (SNR) $\leq$ 0 dB) and variable loading regimes has crystallized as a pivotal research frontier in predictive maintenance.

Current bearing fault diagnosis methodologies bifurcate along two principal paradigms. Traditional techniques leverage handcrafted feature engineering coupled with machine learning classifiers, exemplified by complete ensemble local mean decomposition with adaptive noise (CELMDAN) [7] and Wigner–Ville distributions [8]. While computationally efficient, these methods exhibit three fundamental limitations, i.e., susceptibility to noise contamination exceeding −4 dB SNR, dependence on expert-designed features, and performance degradation under nonstationary loads [9], [10]. Modern data-driven approaches employ hierarchical feature learning through architectures, such as physics-informed convolutional neural networks (CNNs) [11] and multiscale CNNs and long short-term memory (LSTM) hybrids [12]. Although achieving superior accuracy on benchmark datasets, these models frequently falter under data scarcity scenarios or when confronted with distribution shifts between training and deployment environments [13].

The persistent diagnostic challenges stem from three intrinsic complexities. First, multidimensional bearing states are compressed into univariate vibration time-series, obscuring

Received 11 November 2025; accepted 16 November 2025. Date of publication 19 November 2025; date of current version 8 January 2026. This work was supported in part by the National Natural Science Foundation under Grant 62462021, Grant 62267003, and Grant 62576115; in part by Guangxi Key Laboratory of Trusted Software under Grant KX202319; in part by the Special Fund for Guangdong Province University Key Field under Grant 2023ZDZX3041; in part by Guangdong Basic and Applied Basic Research Foundation, China, under Project 2024A1515010144 and Project 2025A1515010197; in part by the Philosophy and Social Sciences Planning Project of Zhejiang Province under Grant 25JCXK006YB; in part by the Hainan Provincial Natural Science Foundation under Grant 625RC716; in part by the Project of Guangdong Province University under Grant 2024KQNCX023; in part by the Zhaoqing University Innovation Research Team Grant Project; in part by the 2022 Guangdong Higher Education Quality Engineering and Teaching Reform Projects; in part by the Foundation of State Key Laboratory of Public Big Data [2022] under Grant 415; and in part by the Projects of PhDs' Start-up Research of GDUPT. *(Corresponding authors: Lianghai Wu; Teng Zhou.)*

Junjie Liu, Jiaxian Zhu, and Weihua Bai are with the School of Computer Science and Software, Zhaoqing University, Zhaoqing, Guangdong 526061, China (e-mail: liujunjie@zqu.edu.cn; zhujiaxian@zqu.edu.cn; baiweihua@zqu.edu.cn).

Huibing Zhang is with Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China (e-mail: zhanghuibing@guet.edu.cn).

Lianghai Wu is with the School of Computer Science, Guangdong University of Petrochemical Technology, Maoming, Guangdong 525000, China (e-mail: wlh@gdupt.edu.cn).

Teng Zhou is with the School of Cyberspace Security (School of Cryptology), Hainan University, Haikou 570228, China, and also with Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Quzhou 324003, China (e-mail: teng.zhou@hainanu.edu.cn).

Keqin Li is with the Department of Computer Science, State University of New York at New Paltz, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/JIOT.2025.3634730

fault signatures when noise power spectral density exceeds −6 dB/Hz [14]. Second, load fluctuations induce time-varying spectral characteristics that violate the stationarity assumptions underpinning conventional methods [15]. Third, existing frameworks inadequately fuse complementary temporal, spectral, and topological descriptors, compromising diagnostic robustness [16]. These limitations are particularly acute in single-sensor scenarios where fault signatures must be extracted from information-starved vibration channels.

To address these challenges, we propose an intelligent fault-diagnosis framework that innovates across three synergistic dimensions. First, we exploit the dyadic decomposition properties of discrete wavelet transforms (DWTs) to implement sub-band-specific noise suppression while preserving transient impulses through nonlinear thresholding in the wavelet domain. Second, through simultaneous projection of denoised signals into multiscale recurrence plots (MSRPs) and Gramian-angular-fields (GAFs), we construct complementary representations capturing both phase-space dynamics and time-delay embeddings. Third, a tri-branch architecture comprising CNN-MSRP, CNN-GAF, and lightweight ResNet (LRNet) modules enables cross-modal feature integration via learnable attention weights, dynamically rebalancing contributions based on SNRs.

Our work introduces three main innovations compared with existing studies. First, a theoretically grounded strategy for noise-robust representation learning is developed by integrating discrete wavelet decomposition with multimodal feature encoding. The DWT-based denoising mechanism effectively preserves informative frequency components while suppressing interference, enabling stable feature representation under extremely low SNRs. Second, a lightweight tri-branch transformer architecture is proposed for complementary extraction and adaptive fusion of temporal, recurrence, and energy–correlation features from MSRP, Gramian angular summation field (GASF), and time-domain modalities. Unlike previous frameworks that independently process time or image data, the proposed cross-modal attention mechanism dynamically aligns feature representations across domains with shared parameters and residual compression, achieving high efficiency with reduced computational cost. Third, extensive experiments on the Case Western Reserve University (CWRU) and PT datasets verify the superiority and generalization of the proposed framework. The method maintains accuracy above 99.6% across noise levels from 0 to −6 dB and outperforms eight state-of-the-art baselines by 3.2–15.7 percentage points. Ablation, attention map visualization, and t-SNE analysis further confirm the interpretability and robustness of the design. These contributions collectively distinguish the proposed approach from prior works by combining noise-resilient signal decomposition, efficient tri-modal feature fusion, and strong transferability across varying operating conditions.

II. Related Work

Rolling-element bearings (REBs) constitute the cornerstone of rotating machinery systems, with their operational status directly impacting industrial safety and productivity. The evolution of fault diagnosis methodologies has progressed through three distinct paradigms, i.e., traditional signal processing techniques, hybrid machine learning approaches, and contemporary deep learning architectures. Each paradigm reflects the technological advancements and theoretical breakthroughs of its respective era, while addressing the specific limitations of its predecessors.

A. Time–Frequency Analytical Methods

Traditional diagnostic approaches primarily rely on signal processing techniques to extract handcrafted features from vibration data. Time-domain methods such as peak detection, kurtosis analysis, and root mean square (rms) measurements provide preliminary fault indicators but lack frequency resolution [17]. Frequency-domain transformations, including fast Fourier transform (FFT) and envelope analysis, enable characteristic frequency identification, yet fail to capture transient features in nonstationary signals [18]. To address these limitations, time–frequency analysis techniques have emerged as the predominant solution, with wavelet transforms (continuous wavelet transform (CWT)/DWT/WPT) and empirical mode decomposition (EMD) demonstrating particular effectiveness in resolving nonlinear and nonstationary bearing vibrations [19]. However, these methods exhibit inherent constraints in automated decision-making, requiring expert knowledge for feature selection and threshold setting [20]. Recent advancements in entropy-based domain adaptation have sought to mitigate these limitations through information-theoretic optimization strategies. Jiao et al. [21] developed an entropy-oriented domain adaptation (EODA) framework combining entropy optimization with convolutional networks for improved generalization across operating conditions, while Ding et al. [22] proposed deep imbalanced domain adaptation (DIDA) to address label shift and class imbalance in cross-domain scenarios.

B. Hybrid Intelligent Diagnosis Systems

The integration of signal processing with machine learning classifiers represents a significant advancement in bearing fault diagnosis. Yu et al. [23] pioneered a hybrid approach combining K-SVD sparse representation with particle-swarm-optimized time-varying filtering (PSO-TVF-EMD), achieving enhanced sparsity and automated parameter tuning for early fault detection. Subsequent work [17] demonstrated the effectiveness of CWT coupled with k-nearest neighbors (kNNs) classifiers for robust fault categorization. More sophisticated frameworks have incorporated evolutionary algorithms, such as the life-cycle approach [24] integrating genetic-algorithm-based variational mode decomposition (GA-VMD) with improved gray-wolf-optimized support vector machines (IGWO-LSSVM), and beluga-whale optimizer [25] combined with maximum correlated kurtosis deconvolution (VME-MCKD) for weak fault detection. Despite their improved performance, these hybrid systems remain constrained by their dependence on manual feature engineering and empirical parameter tuning, limiting adaptability to variable operating conditions and complex noise environments.

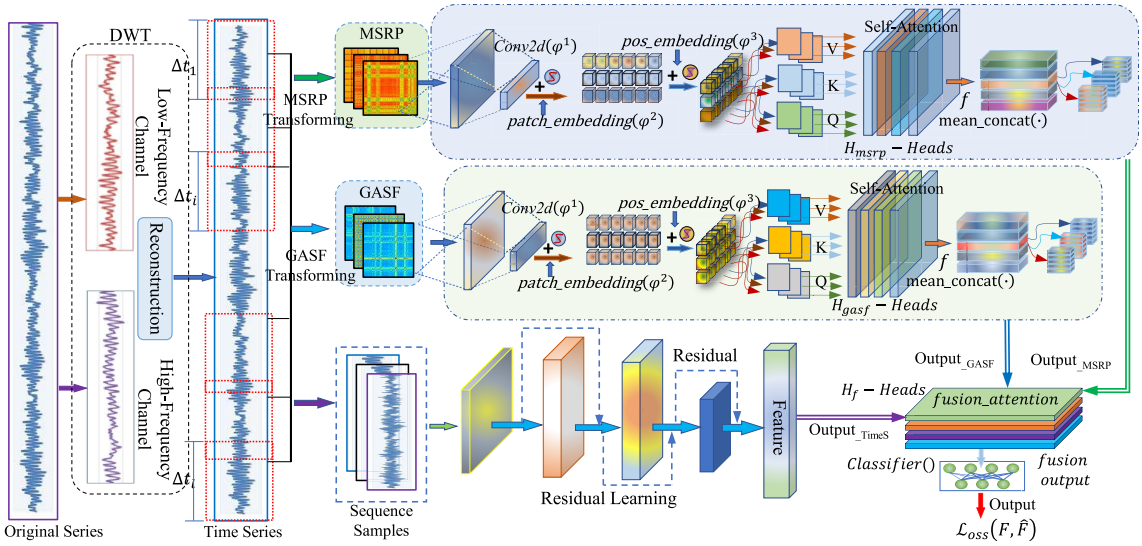


Fig. 1. Overall architecture of the proposed DAMMF-FD model.

C. Deep Learning Paradigms

Deep neural networks have revolutionized bearing fault diagnosis through end-to-end feature learning and classification. CNNs have demonstrated exceptional capability in extracting spatial patterns from vibration signals, with architectures like LeNet-5 and customized CNNs achieving superior performance in raw signal processing [26]. The integration of recurrent structures, particularly LSTM networks, has further enhanced temporal feature extraction, leading to hybrid models such as CNN-LSTM (CNN-L) [27] and CNN-transformer (CNN-T) [28] that simultaneously capture spatial and temporal dependencies.

Recent innovations have focused on attention mechanisms and specialized architectures, including dual-channel attention (DCA) [29] with bidirectional GRU for handling scarce labels, quadratic CNN (QCNN) [30] for noise-robust feature extraction, and QNN-Bi-LSTM [31] combining quadratic neurons with bidirectional LSTM for rapid diagnosis. In addition, a triple domain adversarial neural network (TDANN) with a multiscale feature extractor [32], [33], triple classifier, and adaptive back-propagation coefficient has been proposed to enhance bearing fault diagnosis under varying operating conditions [34].

While these deep learning approaches have significantly advanced diagnostic accuracy, three fundamental challenges persist: 1) inadequate multiscale feature extraction across time–frequency domains; 2) insufficient noise robustness in feature extraction; and 3) incomplete integration of complementary feature representations. Our proposed framework addresses these limitations through a novel combination of frequency-domain decoupling, 2-D signal re-encoding, and attention-based multimodal fusion.

III. Methodology

A. Problem Formulation

In a noisy, univariate bearing-vibration dataset, the measured signal $\mathbf{y} = \{y_1, y_2, \ldots, y_T\} \in \mathbb{R}^T$ is contaminated by random noise ε_t induced by electromagnetic coupling, structural resonance, and sensor errors, where each observation at time t can be expressed as $y_t = x_t + \varepsilon_t$ for $t = 1, 2, \ldots, T$. Here, $x_t \in \mathbb{R}$ represents the true vibration component while ε_t denotes a zero-mean disturbance with bounded variance. Given a labeled vibration database $\mathcal{D} = \{(\mathbf{y}^{(i)}, c_i)\}_{i=1}^N$ with $c_i \in \mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ indicating fault classes, our objective is to establish a discriminative mapping under noise contamination through the composite function

$$\hat{c} = G\left(F\left(\mathcal{D}\left(\mathbf{y}\right)\right)\right) \tag{1}$$

where $\mathcal{D}(\cdot)$ denotes a denoising operator that suppresses ε_t while restoring salient fault patterns, $F(\cdot)$ performs multiscale feature extraction across time–frequency domains, and $G(\cdot)$ constitutes the classification decision function yielding predicted fault category $\hat{c}$. This formulation addresses three critical aspects of bearing fault diagnosis in noisy environments: 1) temporal dependencies through dynamic evolution modeling; 2) frequency-domain fingerprints characterized by energy surges at fault frequencies and harmonics where noise induces spectral leakage; and 3) outlier detection for abrupt changes. The proposed solution follows a three-stage pipeline comprising signal denoising via DWTs, multiscale feature extraction in time–frequency domains, and robust classification through attention-based fusion, as detailed in Section III-H.

B. Overall Framework

Fig. 1 presents the overall architecture of the proposed DAMMF-FD model. The framework is composed of two major components: a noise-mitigation module and a lightweight multiscale multimodal fusion network. The first module applies the DWT to decompose each vibration signal into low- and high-frequency channels, which enhances robustness by isolating informative frequency bands from noise-dominant components. The second module, referred to as AMMF, integrates three cooperative feature-extraction branches to capture complementary information

across temporal and spatial representations. The MSRP branch encodes temporal recurrence patterns through fast-DTW-driven MSRPs, while the GASF branch preserves signal polarity and energy correlations through GASFs. The third branch learns temporal dependencies directly from the reconstructed time-domain sequences using a residual-enhanced transformer block. Features from the three branches are then adaptively fused through a cross-modal attention module, forming a unified representation for classification.

This tri-branch design differs from prior frameworks that simply combine separate image or temporal encoders. The proposed structure enables simultaneous modeling of local temporal cues and global dependency structures under high noise conditions. Although multiple transformations are involved, parameter sharing across transformer heads and lightweight residual operations substantially reduces computational cost, achieving an effective balance between model complexity and diagnostic accuracy.

C. DWT Denoising

The DWT provides simultaneous time–frequency analysis through a variable resolution window. For a square-integrable mother wavelet $\psi(t) \in L^2(\mathbb{R})$ satisfying the admissibility condition $C_\Psi = \int_{-\infty}^{+\infty} |\Psi(\omega)|^2/|\omega| d\omega < +\infty$, its scaled and translated versions are defined as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right), \quad a > 0, b \in \mathbb{R} \tag{2}$$

where a and b represent scale and translation parameters, respectively. The continuous wavelet transform of signal $f(t) \in L^2(\mathbb{R})$ is

$$W_f^\psi(a,b) = \int_{-\infty}^{+\infty} f(t) \overline{\psi_{a,b}(t)} dt. \tag{3}$$

Discretizing via $a = a_0^j$ and $b = k a_0^j b_0$ yields the discrete wavelet basis

$$\psi_{j,k}(t) = a_0^{-j/2} \psi\left(a_0^{-j} t - k b_0\right). \tag{4}$$

Setting $b_0 = 1$ produces an orthonormal basis $\{\psi_{j,k}(t)\}$ with reconstruction formula

$$f(t) = \sum_{j,k} W_f^\psi(j,k) \psi_{j,k}(t). \tag{5}$$

Our implementation employs Daubechies 1 and Haar wavelets for bearing vibration analysis, leveraging their compact support and vanishing moment properties for effective noise suppression while preserving transient fault impulses.

D. MSRPs Feature Enhancement

The MSRP enhancement addresses a critical challenge in bearing fault diagnosis, i.e., extracting discriminative features from noisy vibration signals where fault signatures are often obscured. Traditional single-scale approaches fail to capture the complex temporal dynamics of bearing faults, which manifest across different time scales. Our innovation lies in transforming the 1-D vibration signal into a multiscale 2-D representation that simultaneously preserves amplitude-phase correlations while enhancing fault-related features through scale-specific analysis.

Algorithm 1 MSRP Generation Algorithm

Input: time-series matrix $x(t) \in \mathbb{R}^{N \times P}$; scale set $S = \{s_1, s_2, \ldots, s_k\}$
Output: combined multi-scale recurrence plot $\mathbf{C}_{\text{msrp}}$
Initialize $Dataset \leftarrow x(t)$; $N \leftarrow \text{GetSampleCount}(x(t))$
for each sample **x in** $Dataset$ **do**
  $FeatureData \leftarrow \text{GetFeatureData}(\mathbf{x})$
  $NormalizedData \leftarrow \text{Normalize}(FeatureData)$
  $MultiScaleR \leftarrow \emptyset$// list to hold resized RPs
  for each $s_i \in S$ **do**
    $R[i,j] \leftarrow \text{Sim}(X_{s_i}, X_{s_j})$// Eq. (6)
    $R \leftarrow \text{NormalizeMatrix}(R)$
    $R_{\text{resized}} \leftarrow \text{ResizeMatrix}(R,(W,H))$
    $MultiScaleR.\text{append}(R_{\text{resized}})$
  end for
  $CombinedR \leftarrow \text{CombImages}(MultiScaleR)$
  SaveImage($CombinedR$)
end for
return $\mathbf{C}_{\text{msrp}}$

Given a time series $X = (x_1, x_2, \ldots, x_p)$, we first normalize the data and then construct multiscale subsequences $\text{Sub}_{s_i} = \{(n_j, n_{j+s_i}) | j = 0, 1, \ldots, p - s_i\}$ for each scale s_i in $S = \{s_1, \ldots, s_k\}$. For each scale, the pairwise similarity between subsequences is quantified by computing their Euclidean distance in the phase space. This formulation is a multiscale generalization of the traditional recurrence plot definition, where the distance matrix characterizes the local temporal recurrence of the system states. The similarity matrix at each scale is computed as

$$M_{j,m}^{(s_i)} = \sqrt{\sum_{l=0}^{s_i-1} \left(n_l^j - n_l^m\right)^2}, \quad j,\ m = 1, 2, \ldots, N_{s_i} \tag{6}$$

where $N_{s_i} = p - s_i + 1$. After normalization and resizing to target dimensions $H \times W$ using

$$\mathbf{R}_{\text{rsize}}^{(s_i)} = \text{Zoom}\left(\mathbf{R}_N^{(s_i)}, \left(\frac{H}{N_{s_i}}, \frac{W}{s_i}\right)\right). \tag{7}$$

The final MSRP representation combines all scales

$$\mathbf{C}_{\text{msrp}} = \left[\mathbf{R}_{\text{rsize}}^{(s_1)}, \ldots, \mathbf{R}_{\text{rsize}}^{(s_k)}\right]^T. \tag{8}$$

This multiscale fusion captures both local fault transients and global vibration patterns, overcoming the limitations of conventional single-scale methods. The approach effectively suppresses noise while preserving critical fault signatures across different temporal resolutions, as demonstrated by the framework's superior performance in high-noise environments.

The pseudocode for our MSRP generation algorithm is presented as Algorithm 1.

E. Gramian Signal Feature Enhancement

Industrial vibration signals often suffer from severe noise contamination and nonstationary characteristics, making it

challenging to extract discriminative fault features directly from raw time-series data. To overcome this limitation, we design a Gramian-based feature enhancement method that simultaneously preserves amplitude-phase correlations while amplifying fault-related patterns through polar coordinate transformation. Our method addresses the critical need for noise-robust representations in bearing fault diagnosis under harsh industrial conditions.

Given a denoised time series $X = (x_1, x_2, \ldots, x_p)$, our method first normalizes the signal to $[-1, 1]$ range and converts each point to polar coordinates through a nonlinear mapping

$$u_i = \cos(\theta_i) = n_i \tag{9}$$

$$v_i = \sin(\theta_i) = \sqrt{1 - n_i^2}. \tag{10}$$

The key innovation lies in constructing complementary Gramian matrices that capture distinct aspects of fault signatures. The GASF preserves global phase relationships through cosine summation

$$G_{ij}^G = n_i n_j - \sqrt{1 - n_i^2}\sqrt{1 - n_j^2}. \tag{11}$$

Conversely, the Gramian angular difference field (GADF) emphasizes local variations via sine differences

$$G_{ij}^D = n_j\sqrt{1 - n_i^2} - n_i\sqrt{1 - n_j^2}. \tag{12}$$

The fusion of these orthogonal representations through learnable weights η creates a comprehensive feature space that is more robust to noise than either individual representation

$$G_{ij}^F = \eta G_{ij}^G + (1 - \eta) G_{ij}^D, \quad 0 \le \eta \le 1. \tag{13}$$

This adaptive fusion mechanism automatically balances the contributions of global phase coherence (GASF) and local dynamic variations (GADF) based on signal characteristics to improve feature discriminability in low SNR conditions.

1) GAF Generation: The fused matrix $\mathbf{G}^F$ undergoes normalization and resizing to prescribed dimensions $H \times W$ following the scaling rule in (7), yielding the GAF representation. This transformation preserves the intrinsic amplitude-phase coupling through polar coordinate mapping, significantly enriching the signal feature space's descriptive capacity. The synergistic fusion of GASF and GADF components selectively amplifies discriminative patterns while suppressing noise artifacts, resulting in enhanced SNR for more reliable fault detection. The adaptive weighting parameter η dynamically balances the contributions of GASF and GADF representations according to varying fault conditions, demonstrating robust performance against nonlinear signal variations and dynamic operational environments. The resulting GAF images provide a comprehensive encoding of temporal dynamics and spectral characteristics that prove particularly effective for bearing fault diagnosis in high-noise scenarios. The multiscale GAF algorithm for a given time series is illustrated in Algorithm 2.

Algorithm 2 GAF Generation Algorithm

Input: data matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$, fusion weight η
Output: Gramian Angular Field plot C_{gaf}
$Dataset \leftarrow \mathbf{X}$; $N \leftarrow$ GetSampleCount($\mathbf{X}$)
for each sample $\mathbf{x}$ **in** $Dataset$ **do**
  $\mathbf{n} \leftarrow$ Normalize($\mathbf{x}$)
  for $i \leftarrow 1$ **to** P **do**
    $\theta_i \leftarrow \arccos(n_i)$
    $u_i \leftarrow n_i$
    $v_i \leftarrow \sqrt{1 - n_i^2}$
  end for
  for $i \leftarrow 1$ **to** P **do**
    for $j \leftarrow 1$ **to** P **do**
      $G_{ij}^G \leftarrow u_i u_j - v_i v_j$// Eq. (14)
      $G_{ij}^D \leftarrow v_i u_j - u_i v_j$// Eq. (15)
    end for
  end for
  $G_{ij}^F \leftarrow \eta G_{ij}^G + (1 - \eta) G_{ij}^D$// Eq. (16)
  $\mathbf{G}^F \leftarrow$ NormalizeMatrix($\mathbf{G}^F$)
  $C_{\text{gaf}} \leftarrow$ ResizeMatrix($\mathbf{G}^F, (W, H)$)
  SaveImage(C_{gaf})
end for
return C_{gaf}

F. Feature Extraction Network

As shown in Fig. 1, our DAMMF-FD framework encodes the MSRP and the GAF with a CNN-T backbone. Our feature extraction network integrates convolutional layers with transformer attention to embed each 2-D image into a compact latent space. For the MSRP, the encoder analyzes the spatial-temporal signatures of distinct fault modes, which are mapped onto the k-dimensional scale axis and observed under multiple delay settings. The GAF models the intricate amplitude-phase couplings characteristic of the same faults.

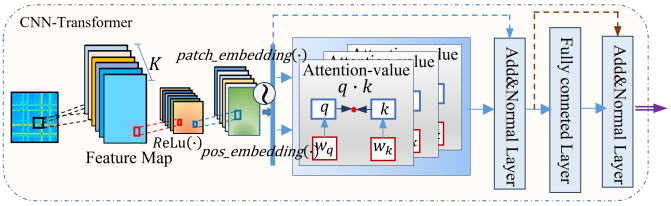


Fig. 2. Feature extraction network.

Fig. 2 illustrates the feature extraction network employed in DAMMF-FD. The architecture is a serial cascade of two units: 1) a patch embedding layer that extracts latent local patterns hidden in the 2-D representation of the original waveform; and 2) a transformer block equipped with self-attention and residual learning to capture global contextual information.

The network outputs a knowledge-level encoding of each MSRP/GAF, which is subsequently fed to the modality-fusion layer.

1) Patch Embedding Layer: As shown in Fig. 2, the patch embedding layer proceeds in three steps. First, a K-channel convolution extracts high-dimensional local spatial features from the input image. Then, a ReLU(·) activation introduces nonlinearity and improves the learning of local patterns. Finally, the original image is partitioned into patches and projected into an embedding space.

For an input GAF/MSRP image $X \in \mathbb{R}^{C \times H \times W}$, we apply K convolutional filters W to extract features. Each filter has

a size of (C, K_h, K_w), where C is the number of channels. H and W denote the image height and width, respectively.

a) Convolutional operation layer: The output tensor $Y \in \mathbb{R}^{N\times K\times H'\times W'}$ of the convolution is computed as

$$Y\left(n,k,h',w'\right) = \sum_{c=0}^{C-1}\sum_{i=0}^{K_h-1}\sum_{j=0}^{K_w-1} X \times \left(n, c,\ h'\cdot S + i - P, w'\cdot S + j - P\right) \times W\left(k,c,i,j\right) + b_k \tag{14}$$

where $Y(n,k,h,w')$ denotes the value at sample n, output channel k, and spatial location (h', w'). $X(\cdot)$ is the input tensor after stride S and padding P. $W(k,c,i,j)$ and b_k are the weight and bias of the kth kernel, respectively. K_h and K_w are the kernel height and width. We use a nonlinear transformation $\sigma(\cdot) = \mathrm{ReLU}(\cdot)$ to prevent vanishing gradients and enforce sparsity.

b) High-dimensional patch embedding: A second convolution with kernel size equal to the patch size embeds $Y^{(1)}$ into a higher dimensional space

$$Y^{(2)}\left(b,p,d\right) = \mathrm{Conv2d}\left(Y^{(1)}\right) \tag{15}$$

where $Y^{(2)} \in \mathbb{R}^{N\times K'\times H''\times W''}$, $b \in \{0,\ldots,\text{batch\_size}-1\}$, $p \in \{0,\ldots,n_{\text{patches}}-1\}$, $d \in \{0,\ldots,\text{embedding\_dim}-1\}$, and $n_{\text{patches}} = \lfloor H/\text{patch\_size}\rfloor \times \lfloor W/\text{patch\_size}\rfloor$.

c) Positional encoding: We add a learnable positional code to retain spatial ordering for the final embedding

$$Y^{(3)}\left(b,p,d\right) = Y^{(2)}\left(b,p,d\right) + \text{pos\_embedding}_1\left(0,p,d\right) \tag{16}$$

where $Y^{(3)} \in \mathbb{R}^{N\times K'\times H''\times W''}$, and $\text{pos\_embedding}_1(0,p,d)$ supplies the position information of the pth patch in the dth dimension to better capture spatial dependencies.

2) Transformer Layer: After high-dimensional feature extraction and embedding, we feed the data into a transformer layer composed of multiple self-attention modules and feed-forward networks to capture global contextual information.

a) Self-attention mechanism: For each patch in $Y^{(3)} \in \mathbb{R}^{N\times K'\times H''\times W''}$, the query, key, and value are obtained by

$$Q = W_Q Y^{(3)}, \quad K = W_K Y^{(3)}, \quad V = W_V Y^{(3)} \tag{17}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{K'\times K_{\text{att}}}$ project the embedded features into a lower dimensional space $\mathbb{R}^{K'\times K_{\text{att}}}$.

The self-attention allows each patch to weigh all others dynamically. For head i in a multihead setting

$$\text{head}_i = \text{Attention}\left(Q_i, K_i, V_i\right) = \text{softmax}\left(Q_i K_i^T \Big/ \sqrt{d_k}\right) V_i \tag{18}$$

and the combined output is

$$\text{MultiHead}\left(Q,K,V\right) = W_0\,\text{Concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_H\right) \tag{19}$$

where d_k is the key dimension, H is the number of heads, and W_0 is a learnable projection matrix.

b) Residual connection and layer normalization: Residual links and normalization improve stability

$$Z^{(1)} = \text{LayerNorm}\left(Y^{(3)} + \text{MultiHead}\left(Q,K,V\right)\right) \tag{20}$$

$$Z^{(2)} = \text{LayerNorm}\left(Z^{(1)} + \text{FeedForward}\left(Z^{(1)}\right)\right) \tag{21}$$

$$\text{FeedForward}\left(Z^{(1)}\right) = \text{ReLU}\left(Z^{(1)}W_1 + b_1\right)W_2 + b_2 \tag{22}$$

where $W_1 \in \mathbb{R}^{d_{\text{model}}\times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff}\times d_{\text{model}}}$, and b_1 and b_2 are learnable parameters.

c) Positional encoding: We add a positional code to preserve spatial relations since the transformer lacks intrinsic order awareness

$$Z^{(3)}\left(b,p,d\right) = Z^{(2)}\left(b,p,d\right) + \text{pos\_embedding}_1\left(0,p,d\right). \tag{23}$$

To summarize, the feature extract network first extracts local patterns from the MSRP and GAF images and then captures their global dependencies. The resulting embeddings are $\text{Output}_{\text{msrp}} = Z_{\text{msrp}}$ and $\text{Output}_{\text{gaf}} = Z_{\text{gaf}}$.

G. Lightweight Convolutional Residual Network

Conventional 1-D processing may overlook salient characteristics embedded in vibration signals. To address this limitation, we propose a LRNet. The LRNet consists of two core operations, i.e., convolution-based feature extraction and feature embedding.

Let the reconstructed time-series signal after the DWT be $X \in \mathbb{R}^{B\times P}$, where B is the batch size and P is the sequence length. We reshape X for convolutional processing as

$$X' = \text{TF}\left(X\left[b,p\right], K, 1\right), \quad X' \in \mathbb{R}^{B\times K\times P\times 1} \tag{24}$$

where $\text{TF}(\cdot)$ first performs dimensional expansion using `unsqueeze`(X,1), producing a tensor of shape $\mathbb{R}^{B\times 1\times P\times 1}$, and then replicates the channel to obtain $\mathbb{R}^{B\times K\times P\times 1}$. The parameter K denotes the expanded number of channels.

1) Convolutional Feature Extraction: The LRNet employs residual blocks composed of a convolution and identity mapping. The output of a block is

$$F\left(x\right) = \text{conv}\left(x,W\right) + x = \sigma\left(W * x + b\right) + x \tag{25}$$

where $x \in \mathbb{R}^{H\times W\times C_{\text{in}}}$ is the input feature map, $W \in \mathbb{R}^{\text{Kernel}\times\text{Kernel}\times C_{\text{in}}\times C_{\text{out}}}$ is the kernel, $\sigma(\cdot) = \text{ReLU}(\cdot)$, and $b \in \mathbb{R}^{C_{\text{out}}}$ is the bias.

2) Feature Mapping: Global average pooling compresses the high-dimensional map into a fixed-length vector, which is then linearly projected

$$Z = W_{\text{LCR}}\,G\left(F\left(X'\right)\right) + b_{\text{LCR}} \tag{26}$$

where W_{LCR} and b_{LCR} are the weights and bias of the output embedding layer.

The pooling operator is channel-wise as

$$G\left(F\right) = \left[\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F\left[h,w\right]\right] \tag{27}$$

where $H \times W$ is the spatial size of the feature map.

H. Multimodal Attention Fusion Layer

The multimodal attention fusion layer (MAFAL) integrates three heterogeneous feature sources-MSRP, GAF, and TimeS-by means of self-attention. The layer analyses, aligns, and fuses complementary cues from the separate streams to enhance the decision reliability.

1) Input Concatenation: The branch outputs are denoted as $X^{(1)} = \text{Output}_{\text{msrp}}$, $X^{(2)} = \text{Output}_{\text{gaf}}$, and $X^{(3)} = \text{Output}_{\text{Times}}$. They are concatenated to form

$$X_{\text{in}} = \left[\text{Output}_{\text{msrp}}; \text{Output}_{\text{gaf}}; \text{Output}_{\text{Times}}\right] \in \mathbb{R}^{B \times D_{\text{MAFAL}}} \quad (28)$$

where B is the batch size and $D_{\text{MAFAL}} = d_{\text{msrp}} + d_{\text{gaf}} + d_{\text{Times}}$.

2) Dynamically Weighted Multihead Attention: For each head $h \in \{1, \ldots, H_f\}$, the attention map $\text{Att}_f^{(h)}(Q_f^{(h)}, K_f^{(h)}, V_f^{(h)})$ is computed as in (18) and (19). A learnable weight matrix W_f projects the head output to a scalar importance

$$\beta_h = \text{softmax}\left(W_f\left[\text{Att}_f^{(h)}\left(Q_f^{(h)}, K_f^{(h)}, V_f^{(h)}\right)\right]^\top\right) \quad (29)$$

with $\beta_h \in \mathbb{R}^{B\times 1}$ and $\text{Att}_f^{(h)} \in \mathbb{R}^{B\times D_f \times D_v}$, where D_f is the sequence length and D_v the embedding dimension.

3) Fusion Output: The final fusion is a weighted sum over all heads as

$$Y^{(M\_\text{out})} = \sum_{h=1}^{H_f} \beta_h \,\text{Att}_f^{(h)}\left(Q_f^{(h)}, K_f^{(h)}, V_f^{(h)}\right) \quad (30)$$

with $Y^{(M\_\text{out})} \in \mathbb{R}^{B\times D}$ and $D = D_f D_v$. The coefficients β_h adaptively modulate the contribution of each head to tailor the fusion to the current input.

I. Fault-Diagnosis Classifier and Loss Function

The DAMMF-FD employs a two-layer MLP to map the fused feature vector $Z = Y^{(M\_\text{out})} \in \mathbb{R}^{B\times D}$ to class probabilities. The forward computation is

$$\hat{y}_i = \text{Softmax}\left(\text{Dropout}\left(\text{ReLU}\left(ZW_1 + b_1\right), p\right)W_2 + b_2\right) \quad (31)$$

where $W_1 \in \mathbb{R}^{D\times H}$ and $b_1 \in \mathbb{R}^H$ are the weights and bias of the first fully connected layer with H hidden units. $W_2 \in \mathbb{R}^{H\times C}$ and $b_2 \in \mathbb{R}^C$ correspond to the second layer, and C is the number of fault classes. The output $\hat{y}_i \in \mathbb{R}^C$ is the probability vector for the ith sample, and the batch output is $\hat{y} \in \mathbb{R}^{B\times C}$.

We train the DAMMF-FD by cross-entropy loss

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} y_i[j]\log\left(\hat{y}_i[j]\right) \quad (32)$$

where $y_i[j] \in \{0, 1\}$ is the ground-truth indicator of class j for sample i.

IV. Experiments

A. Dataset Description

We employ the publicly available rolling-bearing dataset from CWRU, USA, as the benchmark to evaluate the performance of our DAMMF-FD [35]. The experimental platform is shown in Fig. 3.

From the data-acquisition platform, the CWRU dataset records single-point defects with diameters of 7, 14, and 21 mm on the bearing outer race, inner race, and rolling elements. All vibration signals are sampled at 12 kHz under four load levels, i.e., 0, 1, 2, and 3 HP. The experiments diagnose ten fault categories, as listed in Table I.

Ten fault categories are shown in Table I. The mechanical period is approximately 0.0345 s with a motor speed varying from 1797 to 1730 r/min. Each experimental sample spans 412–1024 time steps, i.e., from 0.0345 to 0.0837 s.

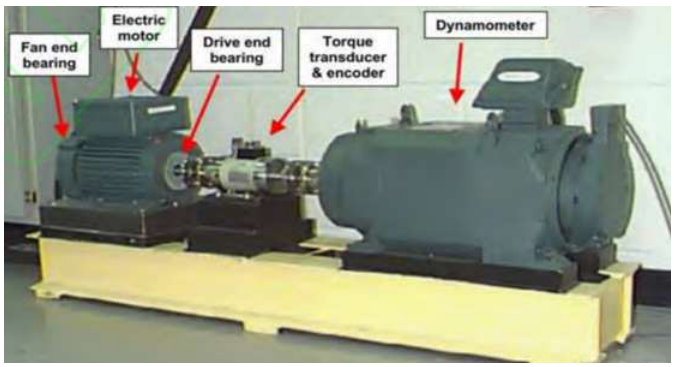


Fig. 3. Data acquisition platform for the CWRU dataset.

TABLE I
TEN FAULT CATEGORIES IN THE CWRU DATASET

Fault description	Class	Fault description	Class
Normal condition	0	Inner race, 14mm	5
Ball, 07mm	1	Inner race, 21mm	6
Ball, 14mm	2	Outer race, 07mm	7
Ball, 21mm	3	Outer race, 14mm	8
Inner race, 07mm	4	Outer race, 21mm	9

B. Noise-Injection Preprocessing

In industrial environments, bearing vibration signals are often contaminated by strong mechanical and acoustic interference. Such interference weakens or masks the weak impulsive components generated by bearing defects, causing severe distortion of both temporal and spectral features. When the SNR decreases, these fault-related impulses become indistinguishable from the surrounding noise, blurring the feature boundaries between different fault categories and reducing the stability of model training and classification. This phenomenon fundamentally explains why intelligent fault diagnosis in high-noise environments is much more difficult than in traditional, clean-signal conditions.

To quantitatively evaluate the noise robustness of the proposed DAMMF-FD, additive Gaussian white noise is injected into each signal segment. The SNR is defined as

$$\text{SNR} = 10 \cdot \log_{10}\left(\frac{\sum_i |x_i|^2}{\sum_i |N_i|^2}\right) \quad (33)$$

where the original signal is $x_i \in x(t) = \{x_1, x_2, \ldots, x_T\}$ and the noise term $N_i \sim \mathcal{N}(0, \sigma^2)$ is sampled from a zero-mean Gaussian distribution. After injection, each dataset sample becomes $\{x_{1_i} + N_{1_i}, x_{2_i} + N_{2_i}, \ldots, x_{T_i} + N_{T_i}\}$.

Three noise levels, SNR = $\{0, -2, -6$ dB$\}$, are imposed on the original signals. Fig. 4 illustrates the time-domain waveforms of a faulty sample under the 0 HP load before and after Gaussian white-noise superposition. The additive model is expressed as $X_{\text{noisy}}(t) = X_{\text{clean}}(t) + \text{Noisy}(t)$, where Noisy$(t)$ follows (33). A comparison among the three SNR levels shows that as the noise intensity increases, the distinctive periodic impulses caused by bearing defects gradually disappear, and at -6 dB they are almost completely buried in noise, making visual or statistical separation among fault types infeasible.

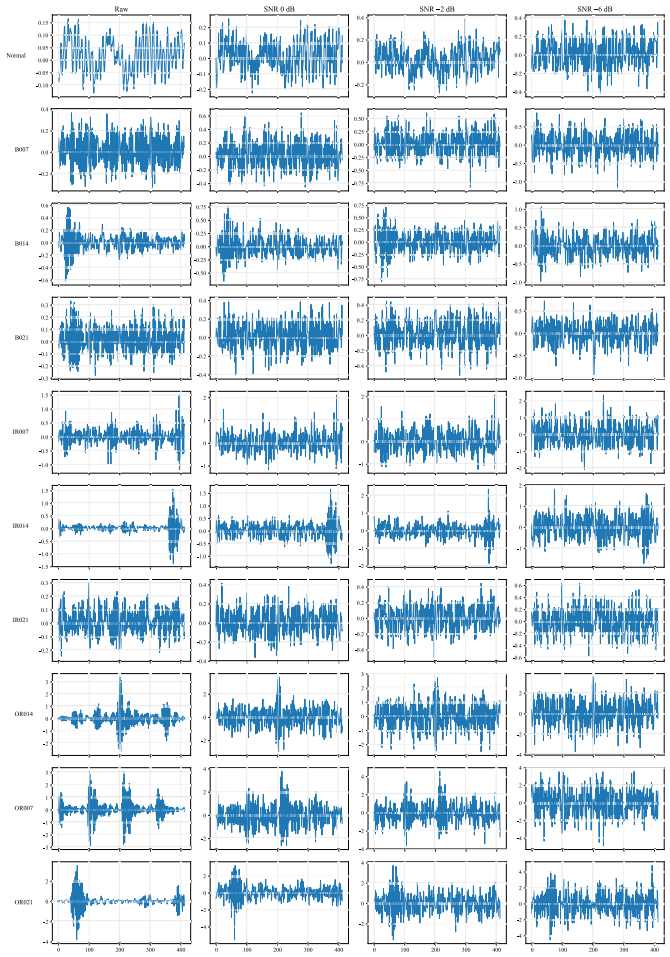


Fig. 4. Time-domain waveforms of the original vibration signal (0 HP load) and its noisy counterparts at different SNR levels.

This observation highlights the intrinsic difficulty of bearing fault diagnosis under strong noise and the necessity of noise-resilient diagnostic algorithms.

From a signal detection perspective, Gaussian white noise represents the maximum-entropy disturbance under equal power constraints and thus constitutes the most challenging random interference model [36], [37]. If a model maintains high diagnostic accuracy under –6 dB Gaussian noise, its robustness boundary against any stationary noise can be reasonably inferred. Moreover, colored noises arising from structural harmonics or non-Gaussian impulsive interferences are either periodic and learnable by spectral-domain models or correspond to abnormal mechanical states that are not representative of routine diagnostic conditions. Consequently, the additive Gaussian noise assumption has been widely adopted in benchmark studies for bearing fault diagnosis, such as [31], [38], [39], [40], and [41], ensuring both methodological consistency and experimental comparability in the community.

C. Experimental Settings

The experiments run on an Xeon[1] Platinum 8255C CPU with an NVIDIA RTX V100 16 GB GPU on the Ubuntu operating system. The code is written in Python 3.8 with PyTorch 1.9.0 and Torch-Vision 0.10.0. We use CUDA 11.0 to accelerate the training. The main hyperparameter settings of the DAMMF-FD model are shown in Table II.

[1]Registered trademark.

TABLE II
MAIN HYPERPARAMETER SETTINGS OF THE DAMMF-FD MODEL

Parameter	Description	Value
S_{size}	MSRP scale set	$[10, 20, 50]$
$(H \times W \times C)$	Image size and channels	$227 \times 227 \times 3$
L_T	Sequence length	$[412, 1024]$
R^{emd}	Embedding dimension	1024
H_L	Heads in multi-head attention	$\{4, 8, 16\}$
H_f	Heads in fusion layer	8
K	Output channels	16
K_h, K_w	Kernel size	$3, 3$
hidden_units	Neurons in feed-forward layer	3072
dropout_rate	Dropout probability	0.1
l_r	Initial learning rate	0.001
Adam	Optimizer (β_1, β_2)	$0.9/0.999$

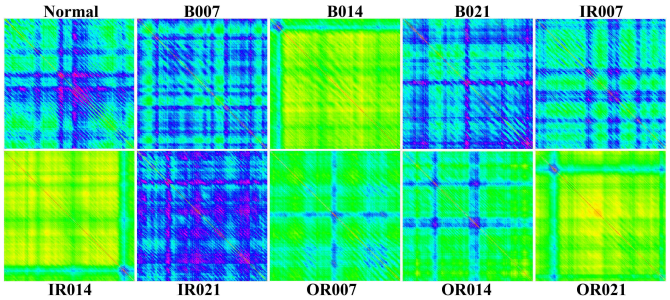


Fig. 5. MSRPs of vibration signals at SNR = −6 dB.

D. Evaluation Metrics

The model performance is evaluated by accuracy (ACC) and $F1$-score

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{34}$$

$$F1\text{-Score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{35}$$

where TP and TN denote the correctly predicted positive and negative samples, respectively. FP and FN are the corresponding mis-classifications. The $F1$-score is the harmonic mean of precision and recall, reflecting both accuracy and coverage of the predictions.

E. Results and Analysis

During DAMMF-FD diagnosis, the outlier-corrected sequences obtained by the DWT are further processed by two feature-enhancement schemes shown in Algorithms 1 and 2. Algorithm 1, `MSRPGeneration`$(x(t))$, builds multiscale 2-D recurrence plots. Algorithm 2, `GAFGeneration`$(x(t))$, constructs GAF maps. Figs. 5 and 6 illustrate the corresponding outputs for samples corrupted to SNR = −6dB.

1) Results on Noisy CWRU Benchmark: The CWRU dataset, contaminated with SNR = $\{0, -6$ dB$\}$, is evaluated under four load levels, i.e., 0, 1, 2, and 3 HP. For the ten fault categories listed in Table I, about 1000 samples are retained

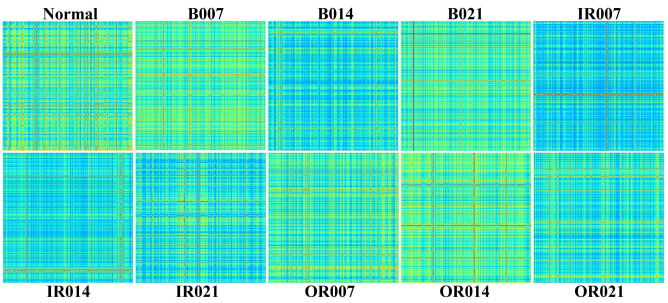


Fig. 6. GAF representations of vibration signals at SNR = −6 dB.

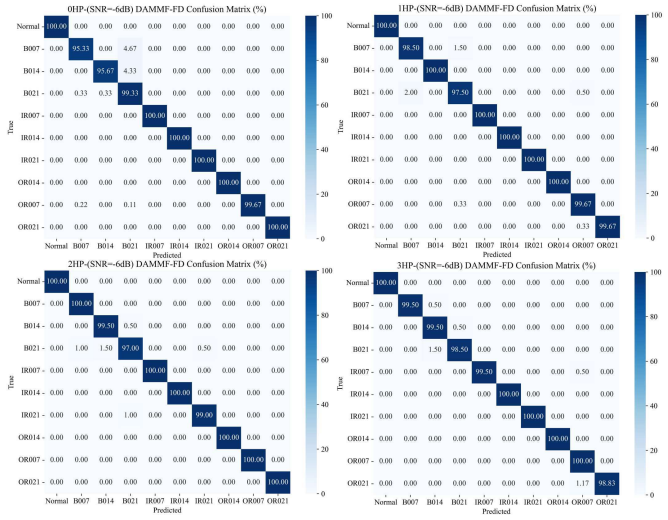


Fig. 7. Confusion matrices of DAMMF-FD on the test sets at SNR = −6 dB under four load levels.

per class. Each class is then split into training and test sets with an 8:2 ratio. The DAMMF-FD model is trained on these partitions and tested on every load-noise pair. The following analysis focuses on this case, since SNR = −6 dB introduces the strongest disturbance.

Fig. 7 reports the confusion matrices obtained at SNR = −6 dB for the four loads. In the 0 HP subset the overall accuracy (ACC) reaches 99.24%. Eight classes achieve a recall of ≥ 99%; the remaining two, B007 and B014, obtain 95.33% and 95.67%, respectively. Specifically, 4.67% of B007 samples are misclassified as B021, and 4.33% of B014 samples are likewise confused with B021. The three ball-fault classes (B007/B014/B021) thus exhibit the greatest feature overlap under no load, accounting for most errors.

The other subplots in Fig. 7 correspond to 1–3 HP, where the accuracies are 99.57%, 99.68%, and 99.54%, respectively.

Overall, even in the most challenging -6 dB scenario of the 0 HP condition, yielding the lowest accuracy, it still maintains ACC = 99.24%. This result underlines the model's strong noise immunity. As the load rises from 0 to 3 HP, the recognition rates for the normal class and for all inner-race (IR) and outer-race (OR) faults remain close to 99%, demonstrating the robustness of the DAMMF-FD to load variation.

2) t-SNE Visualization: For each dataset, the t-SNE plots generated by DAMMF-FD are shown in Figs. 8–11.

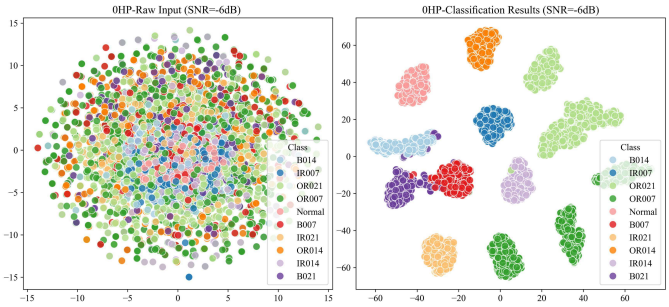


Fig. 8. t-SNE of the testing set at 0 HP, SNR = −6 dB, before/after classification.

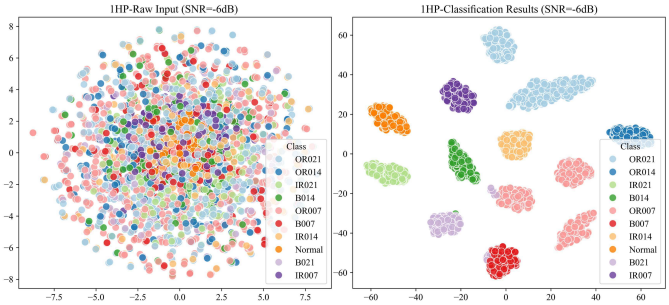


Fig. 9. t-SNE of the testing set at 1 HP, SNR = −6 dB, before/after classification.

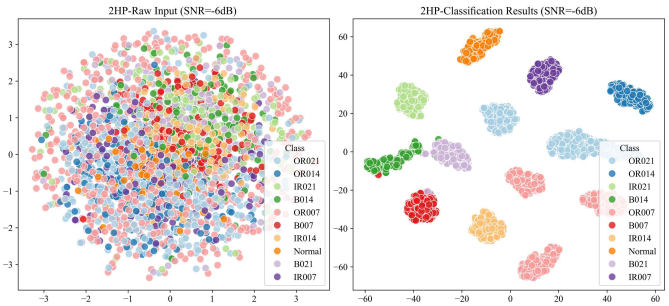


Fig. 10. t-SNE of the testing set at 2 HP, SNR = −6 dB, before/after classification.

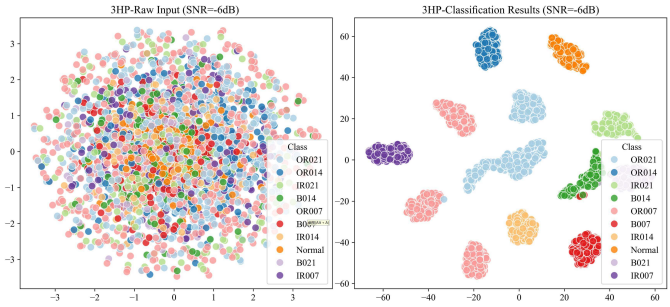


Fig. 11. t-SNE of the testing set at 3 HP, SNR = −6 dB, before/after classification.

Figs. 8–11 present the four-load datasets processed by DAMMF-FD. The model separates the fault classes effectively, comparing the raw distributions with the postclassification clusters. Greater intercluster distance signifies better separability, whereas a smaller distance implies potential confusion. To quantify this property, the Euclidean distances between class centroids are computed and logged. The resulting

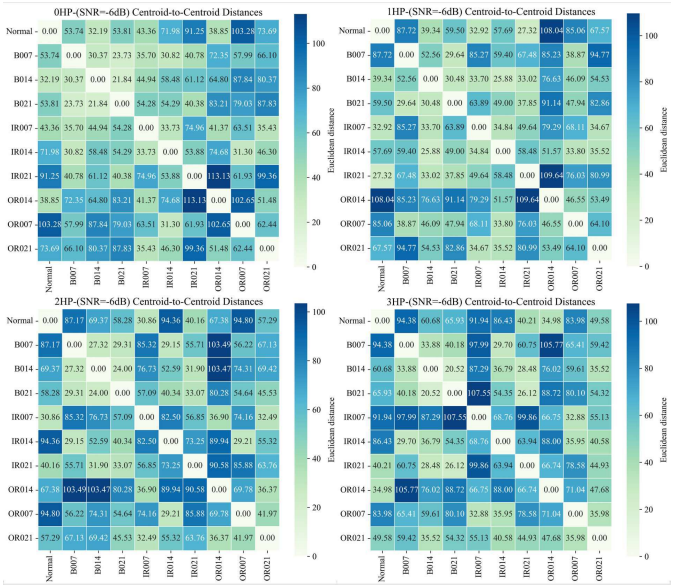


Fig. 12. Heat-maps of centroid Euclidean distances for the four loads at SNR = −6 dB.

centroid-distance heat-maps for the four loads at SNR = −6 dB are plotted in Fig. 12.

Fig. 12 reports the maximal and minimal Euclidean distances between the t-SNE centroids under the following four load levels.

1) 0 HP: max(IR021 ↔ OR014) = 113.13, min(B014 ↔ B021) = 21.84.
2) 1 HP: max(IR021 ↔ OR014) = 109.64, min(B014 ↔ IR014) = 25.88.
3) 2 HP: max(OR014 ↔ B007) = 103.49, min(B014 ↔ B021) = 24.00.
4) 3 HP: max(IR007 ↔ B021) = 107.55, min(B014 ↔ B021) = 20.52.

The distances delineate a three-tier structure as follows.

1) Normal versus all fault classes ($d > 30$).

2) The three major fault families—Ball, IR, and OR—($d > 50$).

3) Different severities within the same family ($d < 30$, most pronounced for the Ball group).

At SNR = −6 dB every load setting still yields ACC > 99%, confirming the model's high noise tolerance and load invariance. The 1 HP heat-map, for example, shows a noticeably reduced separation between the normal class and the two faults IR007 and B021, yet the accuracy remains 99.57%. Hence, DAMMF-FD does not rely solely on centroid spacing; it also leverages finer time–frequency cues and deeper nonlinear decision boundaries, strengthening its robustness.

Finally, the result by the DAMMF-FD is compared with recent diagnosis networks, i.e., CNN-L, CNN-T, FSCL, DCA-BiGRU, QCNN, and QNN-Bi-LSTM. The quantitative results across the four load conditions are summarized in Table III.

From Table III, it is clear that our DAMMF-FD outperforms all baselines. The ACC on all load-noise combinations exceeds 99.5%. Boldface in the table marks column maxima.

To quantify the margin over the baselines, we compute the relative improvement (%) of the DAMMF-FD against each

TABLE III

ACC (%) OF COMPETING FAULT-DIAGNOSIS MODELS ON NOISY CWRU SIGNALS

Dataset	Model	0 dB	−6 dB	Avg
CWRU 0 HP	SVM	66.35 ± 0.36	65.37 ± 0.36	65.87
	LSTM	79.26 ± 0.59	76.48 ± 0.64	77.87
	CNN-L	93.56 ± 1.28	81.88 ± 2.24	87.72
	CNN-T	96.37 ± 0.59	85.45 ± 0.72	90.91
	FSCL	95.26 ± 0.76	82.36 ± 0.52	88.81
	DCA-BiGRU	95.63 ± 2.36	84.89 ± 2.65	90.26
	QCNN	99.94 ± 0.02	97.71 ± 1.05	98.83
	QNN-Bi-LSTM	99.98 ± 0.01	98.10 ± 0.08	99.04
	DAMMF-FD	$\mathbf{99.98 \pm 0.01}$	$\mathbf{99.24 \pm 0.04}$	**99.61**
CWRU 1 HP	SVM	59.65 ± 5.34	53.21 ± 3.81	56.53
	LSTM	85.35 ± 0.29	82.36 ± 0.46	83.86
	CNN-L	94.15 ± 1.26	93.23 ± 0.96	93.69
	CNN-T	96.82 ± 0.81	94.75 ± 0.56	95.79
	FSCL	93.85 ± 0.74	92.05 ± 0.71	92.95
	DCA-BiGRU	84.43 ± 1.32	96.77 ± 2.13	90.06
	QCNN	99.70 ± 0.26	98.28 ± 1.62	98.99
	QNN-Bi-LSTM	99.91 ± 0.05	99.51 ± 0.21	99.71
	DAMMF-FD	$\mathbf{99.97 \pm 0.03}$	$\mathbf{99.58 \pm 0.06}$	**99.78**
CWRU 2 HP	SVM	66.67 ± 0.67	60.39 ± 0.38	63.53
	LSTM	86.24 ± 1.34	83.77 ± 1.42	85.01
	CNN-L	94.68 ± 2.11	91.59 ± 1.22	93.14
	CNN-T	97.33 ± 0.44	93.82 ± 0.25	95.58
	FSCL	94.54 ± 0.74	92.34 ± 0.21	93.44
	DCA-BiGRU	97.13 ± 0.61	93.85 ± 1.72	95.49
	QCNN	100.00 ± 0.00	99.08 ± 0.12	99.54
	QNN-Bi-LSTM	100.00 ± 0.00	99.21 ± 0.10	99.61
	DAMMF-FD	$\mathbf{100.00 \pm 0.00}$	$\mathbf{99.65 \pm 0.07}$	**99.82**
CWRU 3 HP	SVM	69.32 ± 0.63	62.37 ± 0.64	66.31
	LSTM	87.19 ± 0.77	85.38 ± 1.24	86.29
	CNN-L	96.11 ± 0.87	95.09 ± 0.82	95.60
	CNN-T	98.71 ± 2.13	96.66 ± 1.11	97.69
	FSCL	96.03 ± 0.45	94.15 ± 0.61	95.09
	DCA-BiGRU	99.05 ± 0.19	95.05 ± 1.98	97.05
	QCNN	100.00 ± 0.00	99.08 ± 0.12	99.54
	QNN-Bi-LSTM	100.00 ± 0.00	99.22 ± 0.11	99.61
	DAMMF-FD	$\mathbf{100.00 \pm 0.00}$	$\mathbf{99.53 \pm 0.04}$	**99.77**

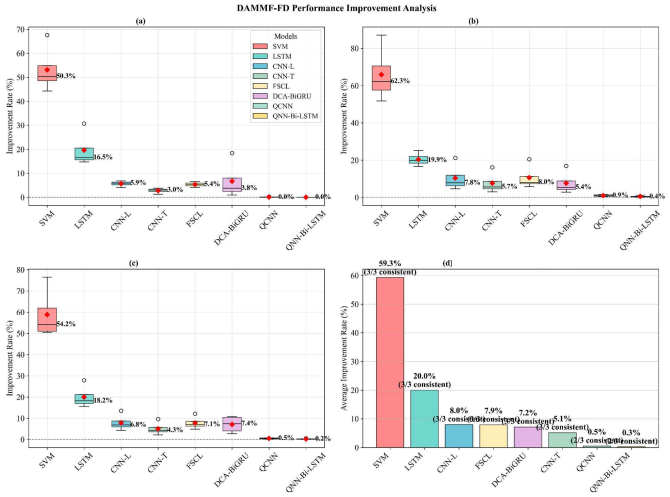


Fig. 13. Performance-improvement analysis of DAMMF-FD. (a) 0 dB SNR condition. (b) –6 dB SNR condition. (c) Average SNR condition. (d) Overall performance improvement.

model. The distribution is summarized by the box plots in Fig. 13.

The descriptive statistics of the four subplots in Fig. 13 reveal a hierarchical advantage as follows.

1) At SNR = 0 dB, the conventional machine-learning method SVM shows the greatest room for improvement with a mean of 53.1%. The LSTM gains are moderate (about 20%). The CNN variants improve by 5%–8%. The QCNN and the QNN-Bi-LSTM improve by less than 1%.

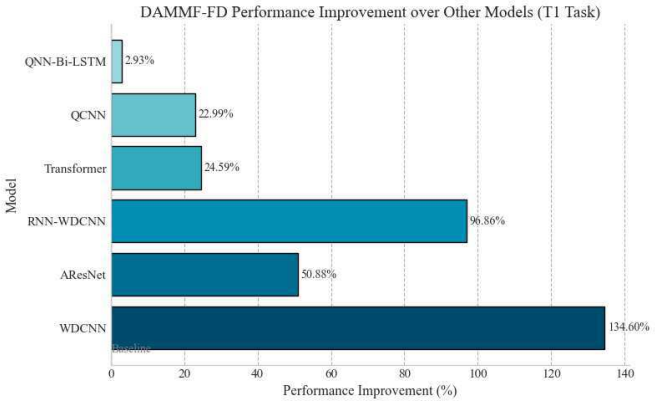


Fig. 14. DAMMF-FD performance improvement over other models (T1 task).

2) At SNR = −6 dB, the improvements are most pronounced, and all eight baselines reach their highest improvement. The SVM rises to 65.9%, underscoring the superior noise robustness of the DAMMF-FD.

3) The average statistics confirm the consistency of these improvements.

Across the four datasets, the improvement ranges from a significant increase of 59.3% over conventional SVM to a consistent and modest improvement over the latest QCNN. Thus, the DAMMF-FD achieves uniform improvements under all test conditions. Especially in low-SNR scenarios, its outstanding noise robustness and cross-architecture adaptability demonstrate comprehensive superiority and practical value.

F. Expanded Experimental

To evaluate the proposed DAMMF-FD in terms of generalization capability, robustness, and multiscale adaptability under varying temporal resolutions and cross-working conditions, three experimental tasks (T1–T3) were designed. These tasks encompass three levels of evaluation: single-domain fault diagnosis (SFD), fault type diagnosis (FTD), and cross-domain recognition with fault degree diagnosis (FDD). Together, they establish a comprehensive framework for validating the diagnostic performance and assessing whether the dataset introduces overfitting.

1) T1 (Single-Domain Fault Diagnosis): This task aims to perform fundamental fault identification using time-series samples of length 2048. Under individual operating conditions with varying rotational speeds and loads, four health states are classified: normal condition (NC), inner race fault (IF), outer race fault (OF), and ball fault (BF).

T1 is designed to evaluate the model's feature extraction and stable classification capability at a longer temporal resolution. It serves as the baseline for evaluating the framework's core diagnostic performance and provides a reference for subsequent cross-domain validation. The result of Task T1 is shown in Table IV. The improvement compared with other models is shown in Fig. 14.

From Table IV and Fig. 14, it shows that DAMMF-FD achieves the highest diagnostic accuracy across all fault types under noisy (−6 dB SNR) and variable load conditions,

TABLE IV
AVERAGE DIAGNOSTIC ACCURACY (%) FOR EACH FAULT TYPE UNDER −6 DB SNR ACROSS 0–3 HP LOAD CONDITIONS

Model	BF	IF	OF	Normal
WDCNN [42]	34.40	20.53	67.33	70.40
AResNet [43]	80.50	37.60	72.00	100.00
RNN-WDCNN [44]	71.20	35.20	39.30	100.00
Transformer	85.28	65.37	79.56	100.00
QCNN	89.60	67.87	75.73	100.00
QNN-Bi-LSTM	95.21	91.15	92.30	100.00
DAMMF-FD	**97.19**	**94.32**	**95.31**	**100.00**

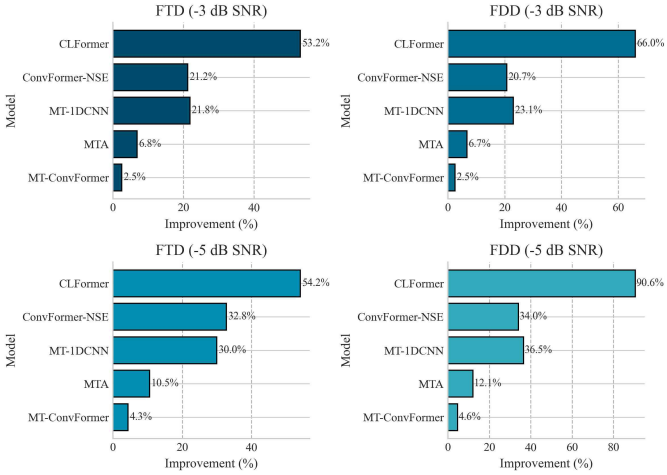


Fig. 15. DAMMF-FD performance improvement across SNR levels and diagnosis tasks (T2 and T3).

outperforming all benchmark models. Compared with advanced baselines such as QNN-Bi-LSTM and QCNN, DAMMF-FD exhibits a further 2%–25% improvement, demonstrating superior robustness and generalization in the T1 task.

2) T2 (Fault Type Diagnosis): Task T2 also focuses on classifying the same four fault categories (NC, IF, OF, and BF) but utilizes the C_1-C_3 cross-domain dataset configuration shown in Table V. Here, different speed–load combinations (1797–1730 rpm, 0–3 HP) are defined as source and target domains to assess cross-condition recognition stability.

3) T3 (Fault Degree Diagnosis): Task T3 adopts the same C_1-C_3 structure shown in Table V, but focuses on distinguishing fault severities represented by three crack sizes (0.007, 0.014, and 0.021 inch). By defining source and target domains under different speeds and loads, the experiment simulates realistic multiscale damage patterns and working condition variations.

The experimental results for tasks T2 and T3 are presented in Tables VI and VII, respectively.

The overall performance improvements of tasks T2 and T3 over other models under different noise levels are illustrated in Fig. 15.

According to the results in Fig. 15 and Table VI, the proposed DAMMF-FD demonstrates remarkable robustness and cross-domain generalization under noisy conditions. For task T2 (FTD) and task T3 (FDD), DAMMF-FD achieves

TABLE V
DESCRIPTION OF DIAGNOSIS TASKS AND DOMAIN SETTINGS IN THE CWRU DATASET

Speed	Load	Samples	T_2: FSD	T_3: CDV	Source and Target Domain (C_1–C_3)
1797 rpm	0 HP	1900	NC	NC	C1: Source: 1797 rpm, 0 HP → Target: 1772 rpm, 1 HP; 1750 rpm, 2 HP
1772 rpm	1 HP	1900	IF	0.007 inch	C2: Source: 1772 rpm, 1 HP → Target: 1797 rpm, 0 HP; 1750 rpm, 2 HP
1750 rpm	2 HP	1900	OF	0.014 inch	C3: Source: 1750 rpm, 2 HP → Target: 1797 rpm, 0 HP; 1772 rpm, 1 HP
1730 rpm	3 HP	1900	BF	0.021 inch	–

Note: A sliding window of size 1024 was used to consecutively extract samples from each health state under every operational condition, yielding a total of 7600 samples. For both of Tasks T_2 and T_3, the training/test split was 4:1, both source- and target-domain samples followed the dataset combinations (C_1–C_3) and were divided at ratios of 2:1 and 1:2, respectively, to assess the generalization and robustness of the proposed method across different operating conditions and fault sizes.

TABLE VI
AVERAGE ACC (%) OF MODELS FOR T2 AND T3 UNDER DIFFERENT SNRS

Model	-3 dB SNR		-5 dB SNR	
	FTD	FDD	FTD	FDD
MT-ConvFormer [45]	95.03±0.64	92.16±0.42	88.36±0.99	86.53±0.54
MTA [46]	91.15±1.55	88.52±2.22	83.38±1.40	80.74±2.68
MT-1DCNN [47]	79.93±1.79	76.73±2.19	70.90±1.98	66.32±2.13
ConvFormer-NSE [48]	80.32±1.37	78.21±0.74	69.42±2.95	67.53±3.07
CLFormer [49]	63.56±6.97	56.87±5.85	59.79±5.33	47.49±2.50
DAMMF-FD	**97.36±0.29**	**94.43±0.37**	**92.17±0.33**	**90.52±1.27**

TABLE VII
FAULT DIAGNOSIS MODEL ACC RESULTS (%) FOR DIFFERENT DATASETS UNDER VARIOUS SNR CONDITIONS IN T2 AND T3

Datasets	Methods	-3 dB SNR		-5 dB SNR	
		FTD	FDD	FTD	FDD
C1	MT-ConvFormer	91.42±1.91	91.18±1.62	86.10±1.54	86.94±1.28
	MTA	87.72±1.03	83.88±3.11	77.04±1.90	75.98±3.16
	MT-1DCNN	68.50±2.74	53.92±4.33	57.38±2.21	44.50±1.14
	ConvFormer-NSE	90.58±2.80	71.08±2.90	83.24±1.77	62.12±7.44
	CLFormer	82.96±4.75	55.82±4.42	64.91±16.64	39.47±4.57
	DAMMF-FD	**93.41±0.81**	**93.12±0.63**	**91.22±0.82**	**90.15±0.73**
C2	MT-ConvFormer	92.84±0.63	91.14±2.05	87.76±2.11	85.94±1.41
	MTA	77.44±1.70	81.66±0.46	73.28±2.34	74.88±1.74
	MT-1DCNN	64.06±3.24	58.24±5.13	59.56±1.80	54.58±1.15
	ConvFormer-NSE	91.32±2.00	75.40±8.69	85.34±2.08	61.82±6.62
	CLFormer	82.14±2.83	51.22±1.78	60.42±9.90	43.82±1.34
	DAMMF-FD	**94.29±0.39**	**93.64±0.46**	**90.73±0.51**	**90.07±1.13**
C3	MT-ConvFormer	92.32±1.00	91.60±0.43	86.64±1.38	85.88±2.74
	MTA	83.06±2.24	83.96±3.18	71.98±0.52	73.32±1.46
	MT-1DCNN	62.42±1.40	58.70±4.33	59.24±0.77	51.52±3.03
	ConvFormer-NSE	92.10±0.98	77.06±2.98	85.72±1.75	69.26±1.35
	CLFormer	81.44±2.60	54.58±7.82	69.96±5.25	41.74±3.03
	DAMMF-FD	**94.08±0.22**	**93.61±0.66**	**90.55±0.48**	**90.11±1.32**

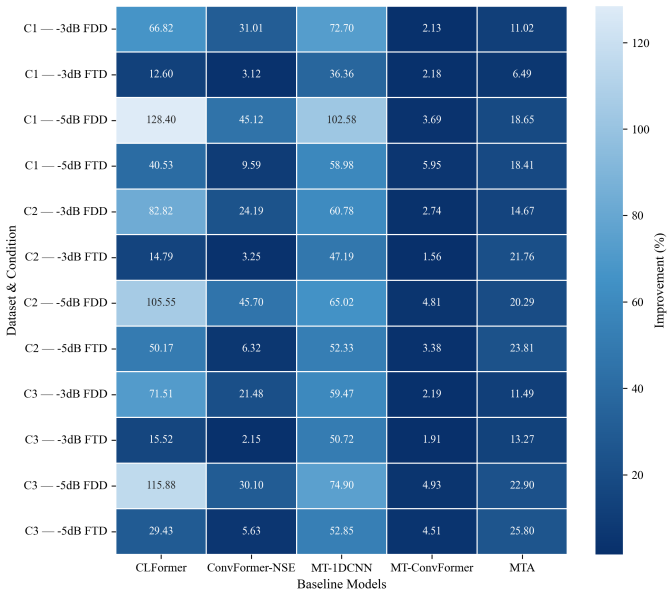


Fig. 16. DAMMF-FD performance improvement across datasets, noise, and tasks (T2 and T3).

significant improvements over MT-ConvFormer, yielding gains of 2.5% and 2.5% at −3 dB SNR, and 4.3% and 4.6% at −5 dB SNR, respectively. The corresponding average accuracies reach 97.36%/94.43% (FTD/FDD), surpassing all compared models. In contrast to CLFormer and MT-1DCNN, whose performances degrade sharply under high noise levels, DAMMF-FD attains improvements of up to 54.2%–90.6%, confirming its superior robustness, feature transferability, and multiscale fault sensitivity across varying working conditions.

Fig. 16 presents a heatmap illustrating the performance improvement of DAMMF-FD over other models for tasks T2 and T3, across different datasets (C_1, C_2, and C_3) and noise levels.

From Fig. 16 and Table VII, in Tasks T2 and T3, the proposed DAMMF-FD consistently achieves the highest accuracy and stability across all three cross-domain configurations (C_1, C_2, and C_3) and noise levels (−3 and −5 dB). Compared with other baseline models, DAMMF-FD exhibits substantial improvements, maintaining an accuracy above 90% even under severe noise degradation, which demonstrates its superior robustness and domain generalization capability.

Under the C_1 configuration, DAMMF-FD outperforms the second-best model by approximately 2%–5.95% in both FTD and FDD tasks, while achieving the lowest standard deviations, indicating enhanced stability in noisy industrial environments. For C_2 and C_3, the accuracy gains become more prominent, especially in FDD tasks, where DAMMF-FD outperforms conventional models by 10%–20% under both −3 and −5 dB conditions, confirming its effective feature transfer and domain-invariant representation. The improvement visualization further reveals that performance gains in FDD tasks are consistently greater than in FTD, highlighting the model's superior capability for multiscale feature adaptation and fine-grained damage discrimination.

Based on the results summarized in Tables VI and VII, DAMMF-FD demonstrates outstanding performance consistency and reliability across different SNRs (−3 and −5 dB),

TABLE VIII
ACC/F1 (%) OF THE ABLATION VARIANTS ON CWRU SIGNALS. THE BEST VALUE IN EACH COLUMN IS BOLD AND SHADED

Dataset	Model	0 dB		−6 dB	
		ACC	F1	ACC	F1
CWRU 0 HP	RP	96.26	96.39	93.88	93.69
	MSRP	89.67	90.04	83.41	83.40
	GAF	91.78	92.12	84.29	84.29
	-DWT	86.95	87.14	79.78	78.66
	+DWT	90.46	90.39	81.23	81.24
	DAMMF-FD	**99.98**	**99.98**	**99.24**	**99.26**
CWRU 1 HP	RP	96.51	97.64	94.31	94.33
	MSRP	91.87	92.07	85.19	85.19
	GAF	92.08	92.09	86.25	86.71
	-DWT	87.28	87.35	82.11	82.07
	+DWT	90.83	90.85	81.76	82.01
	DAMMF-FD	**99.97**	**99.98**	**99.58**	**99.62**

cross-domain datasets (C_1, C_2, and C_3), and two diagnostic tasks T2 (FTD), and T3 (FDD).

In the overall averages, DAMMF-FD achieves 97.36% and 94.43% accuracy for FTD and FDD under −3 dB conditions and maintains 92.17% and 90.52% under −5 dB. Compared with the second-best model (MT-ConvFormer), it shows an average improvement of 2%–4% with minimal fluctuation (±1%), indicating excellent model stability under noisy conditions.

The three tasks (T1–T3) are systematically structured to evaluate the model's single-condition baseline accuracy, its domain generalization ability (C_1, C_2, and C_3), and its multiscale adaptability to fault evolution. This hierarchical design rigorously mitigates concerns of overfitting on the CWRU dataset and extends the experimental validity to varied domain scenarios.

G. Ablation Study

1) Ablation Study of DWT and 2-D Branches: To evaluate the necessity of the DWT and the two image-conversion branches in DAMMF-FD, we conduct an ablation study on the 0 and 1 HP datasets under SNR = 0 and -6 dB. The following variants are compared.

1) *RP:* The plain recurrence-plot branch only.
2) *MSRP:* The multiscale recurrence-plot branch only.
3) *GAF:* The GAF branch only.
4) *−DWT:* The DWT module removed.
5) *+DWT:* The DWT is retained, but both image branches are removed.

The averaged ACC and $F1$-scores are listed in Table VIII. From Table VIII, we have the following observations.

1) The single RP branch maintains about 96% and 93% ACC at 0 and −6 dB, respectively, about 3–5 pp below the full model.

2) When only MSRP or GAF is used, ACC drops to 89%–92% at 0 dB and to 83%–84% at −6 dB, confirming that neither a single scale nor a single encoding can cover the diverse spatiotemporal patterns of bearing faults.

3) *−DWT* causes the steepest decline. ACC falls to 86.9% at 0 dB and to 79.8% at −6 dB, a loss of 13–20 pp, the largest among all settings. This highlights that multiresolution analysis in the joint time–frequency domain is crucial for discriminative feature extraction.

4) *+DWT* restores accuracy to only about 90%, still nearly 10 pp lower than the full model. Only time–frequency spectra are therefore insufficient. Multiscale RP and GAF must capture global geometric relations in the 2-D space to provide complementary cues.

5) Comparing 0 with −6 dB shows that the performance gap between each reduced model and the complete network widens as noise increases (mean gap enlarges from 9.2 to 13.7 pp), reaffirming the importance of the multibranch fusion for noise suppression and robustness enhancement.

Table VIII shows that the complete DAMMF-FD reaches near-perfect recognition at both operating conditions (0 dB: ACC/$F1$ = 99.98%/99.98%; −6 dB: 99.24%/99.26%). Any removal of a module or branch causes a marked decline, which becomes more severe at −6 dB, indicating that every subcomponent is crucial for noise robustness.

The ablation results confirm the necessity and complementarity of the two-level design. The DWT supplies high-/low-frequency decomposition along the temporal axis, highlighting impulse features and serving as the backbone of the framework. The recurrence-plot branches (including the multiscale extension) and the GAF branch project the 1-D vibration signal into 2-D topological and phase spaces, greatly enhancing perception of dynamic patterns and global dependencies.

Overall, the superiority of DAMMF-FD does not stem from any single block. It arises from the synergistic pipeline of "frequency-domain decomposition → 2-D re-encoding → feature-level fusion."

2) Hyperparameter Tuning: To identify a suitable configuration for fault diagnosis, we investigate two key settings, the convolutional kernel size $K_h \times K_w$ and the number of self-attention heads H_L. We perform five independent runs on the 0 and 1 HP subsets under SNR = −6 dB. The averages are reported in Table IX. We have the following three findings.

1) Replacing the baseline 3×3 kernel with 1×1 or 5×5 significantly degrades performance. The 1×1 kernel lowers ACC by 5.66 and 5.32 pp on 0 and 1 HP, respectively. The 5×5 kernel performs even worse. A large receptive field dilutes fine-grained fault cues, whereas an overly small one

TABLE IX
AVERAGE ACC/F1 (%) UNDER DIFFERENT HYPER-PARAMETER SETTINGS

Dataset	Setting	ACC	F1
CWRU 0 HP, −6 dB	$K_h \times K_w = 1 \times 1$	93.58	93.61
	$K_h \times K_w = 5 \times 5$	92.87	93.15
	$H_L = 4$	90.69	90.69
	$H_L = 8$	96.84	96.84
	DAMMF-FD $(3 \times 3, 16)$	**99.24**	**99.26**
CWRU 1 HP, −6 dB	$K_h \times K_w = 1 \times 1$	94.26	94.28
	$K_h \times K_w = 5 \times 5$	93.73	93.73
	$H_L = 4$	91.44	91.47
	$H_L = 8$	97.73	97.81
	DAMMF-FD $(3 \times 3, 16)$	**99.58**	**99.62**

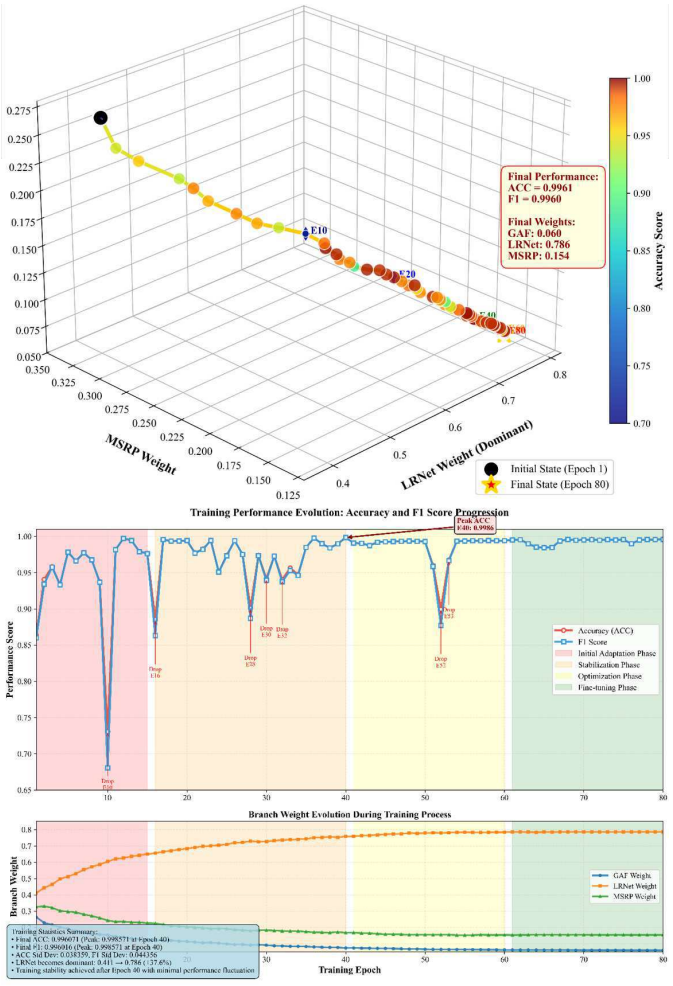


Fig. 17. Branch-weight evolution and ACC/$F1$ variation during training.

weakens contextual linkage. Thus, a moderate 3×3 kernel, combined with multiscale frequency fusion, is preferable for extracting local time–frequency motifs.

2) Adequate parallel subspaces help recover correlations buried by noise. Reducing H_L to 8 yields only a slight loss, but $H_L = 4$ introduces a clear bottleneck, confirming that insufficient heads hamper channel complementarity.

3) All hyper-parameter changes exhibit almost identical relative trends on 0 and 1 HP, indicating stable generalization. Close ACC-$F1$ agreement further suggests balanced class distributions and minimal prediction bias.

In summary, the DAMMF-FD attains its robustness and precision by maintaining a moderate receptive field and a sufficient number of attention heads, thereby realizing the synergy among kernel size, attention parallelism, and dynamic frequency-domain fusion.

3) Branch-Weight Analysis: Fig. 17 illustrates the evolution of the three branch weights of the DAMMF-FD during one complete training run together with the corresponding ACC/$F1$ curves for the 1 HP subset under SNR = −6 dB. The upper panel in Fig. 17 is the 3-D training trajectory of the DAMMF-FD weight evolution in 3-D space. The lower panel plots the branch weights against the ACC and $F1$-scores.

In the 3-D trajectory, the axes correspond to the weights of the LRNet, MSRP, and GAF branches, respectively. The path moves monotonically from an almost balanced start point $(0.41, 0.33, 0.26)$ to an endpoint that is strongly biased toward the CNN-T branch ($\approx 0.79, 0.15, 0.06$). Point color and size encode instantaneous ACC and $F1$. Both increase steadily along the path, indicating a strong positive correlation between performance improvement and the rising LRNet weight. Milestones (E10, E20, E40, E60, and E80) reveal three dynamic stages. 1) The CNN-T weight rises sharply from epochs 1 to 15. 2) MSRP declines slowly and GAF decays markedly from epochs 16 to 60. 3) All three weights oscillate slightly and converge after epoch 60.

The lower panel quantifies these trends. The LRNet weight grows by about 93%. The GAF weight declines by roughly 77%, and the MSRP weight stabilizes near 0.15, further confirming the stage-wise roles of the branches.

Combining spatial and temporal visualization, the LRNet branch contributes most to the final discriminative power and is the main driver of performance gain. The MSRP keeps a moderate weight in the middle and late phases, providing auxiliary robustness. The GAF contribution gradually weakens, yet the ablation study shows that this branch still exerts a significant supportive influence.

In practice, the model runs at the edge: sensors acquire vibration/acoustic signals, and on-node multiresolution enhancement and tri-branch fusion produce fault labels and confidence scores with low overhead. When needed, compact features or summaries are uploaded to a gateway/cloud for batch diagnosis and fleet-level supervision, reducing bandwidth while enabling centralized updates.

V. Conclusion

This work presents a two-stage framework (DAMMF-FD) for robust bearing fault diagnosis in high-noise industrial IoT (IIoT) environments. The method targets three core challenges in vibration-based diagnosis: 1) noise suppression via subband decomposition, selective reconstruction, and adaptive enhancement; 2) comprehensive representation using MSRPs and GAFs to capture temporal and spectral patterns; and 3) adaptive cross-modal fusion through a tri-branch attention architecture.

On the CWRU benchmark, DAMMF-FD attains a mean accuracy of 99.61% from 0 to −6 dB and under varying loads, outperforming eight state-of-the-art baselines by 3.2–15.7 percentage points. Ablation results verify the contribution of each component. t-SNE visualizations show clear class separation in the latent space at −6 dB, supporting the model's discriminative capacity. Unlike prior single-condition or single-scale evaluations, the multidomain and multiscale configuration (T1–T3) explicitly assesses domain shift and noise perturbation. The results provide concrete evidence of cross-condition robustness and enhanced generalizability. The lightweight design and strong noise immunity indicate suitability for IIoT sensing nodes and edge deployment.

Future work will focus on: 1) on-device optimization and quantization-aware training for stricter resource budgets; 2) self-supervised pretraining to reduce labeled data demands;

and 3) reliability assessment under sensor aging and nonstationary noise.

REFERENCES

[1] D. Wang, Y. Li, L. Jia, Y. Song, and T. Wen, "Attention-based bilinear feature fusion method for bearing fault diagnosis," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 3, pp. 1695–1705, Jun. 2023.

[2] Y. Zhang, Z. Ren, S. Zhou, K. Feng, K. Yu, and Z. Liu, "Supervised contrastive learning-based domain adaptation network for intelligent unsupervised fault diagnosis of rolling bearing," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 6, pp. 5371–5380, Dec. 2022.

[3] A. A. Soomro et al., "Insights into modern machine learning approaches for bearing fault classification: A systematic literature review," *Results Eng.*, vol. 23, Sep. 2024, Art. no. 102700.

[4] B. Peng, Y. Bi, B. Xue, M. Zhang, and S. Wan, "A survey on fault diagnosis of rolling bearings," *Algorithms*, vol. 15, no. 10, p. 347, Sep. 2022.

[5] T. Lin, Y. Zhu, Z. Ren, K. Huang, and D. Gao, "CCFT: The convolution and cross-fusion transformer for fault diagnosis of bearings," *IEEE/ASME Trans. Mechatronics*, vol. 29, no. 3, pp. 2161–2172, Jun. 2024.

[6] J. Chen et al., "Adversarial-based super feature reconstruction meta-transfer network for weak feature enhancement and fault diagnosis of harmonic drive," *IEEE/ASME Trans. Mechatronics*, early access, Dec. 11, 2024, doi: 10.1109/TMECH.2024.3506746.

[7] L. Wang, Z. Liu, Q. Miao, and X. Zhang, "Complete ensemble local mean decomposition with adaptive noise and its application to fault diagnosis for rolling bearings," *Mech. Syst. Signal Process.*, vol. 106, pp. 24–39, Jun. 2018.

[8] G. Dong and J. Chen, "Noise resistant time frequency analysis and application in fault diagnosis of rolling element bearings," *Mech. Syst. Signal Process.*, vol. 33, pp. 212–236, Nov. 2012.

[9] L. Li, W. Meng, X. Liu, and J. Fei, "Research on rolling bearing fault diagnosis based on variational modal decomposition parameter optimization and an improved support vector machine," *Electronics*, vol. 12, no. 6, p. 1290, Mar. 2023.

[10] W. Shen, M. Xiao, Z. Wang, and X. Song, "Rolling bearing fault diagnosis based on support vector machine optimized by improved grey wolf algorithm," *Sensors*, vol. 23, no. 14, p. 6645, Jul. 2023.

[11] D. Ruan, J. Wang, J. Yan, and C. Gühmann, "CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis," *Adv. Eng. Informat.*, vol. 55, Jan. 2023, Art. no. 101877.

[12] T. Huang, Q. Zhang, X. Tang, S. Zhao, and X. Lu, "A novel fault diagnosis method based on CNN and LSTM and its application in fault diagnosis for complex systems," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 1289–1315, Feb. 2022.

[13] L. Zhao, Y. He, D. Dai, X. Wang, H. Bai, and W. Huang, "A novel multi-task self-supervised transfer learning framework for cross-machine rolling bearing fault diagnosis," *Electronics*, vol. 13, no. 23, p. 4622, Nov. 2024.

[14] M. Ye, X. Yan, N. Chen, and M. Jia, "Intelligent fault diagnosis of rolling bearing using variational mode extraction and improved one-dimensional convolutional neural network," *Appl. Acoust.*, vol. 202, Jan. 2023, Art. no. 109143.

[15] Z. Feng, S. Wang, and M. Yu, "A fault diagnosis for rolling bearing based on multilevel denoising method and improved deep residual network," *Digit. Signal Process.*, vol. 140, Aug. 2023, Art. no. 104106.

[16] H. Wang, Z. Liu, D. Peng, and M. J. Zuo, "Interpretable convolutional neural network with multilayer wavelet for noise-robust machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 195, Jul. 2023, Art. no. 110314.

[17] H. S. Kumar and G. Upadhyaya, "Fault diagnosis of rolling element bearing using continuous wavelet transform and K-nearest neighbour," *Mater. Today, Proc.*, vol. 92, pp. 56–60, Jan. 2023.

[18] P. Shakya, A. K. Darpe, and M. S. Kulkarni, "Vibration-based fault diagnosis in rolling element bearings: Ranking of various time, frequency and time-frequency domain data-based damage identi cation parameters," *Int. J. Condition Monitor.*, vol. 3, no. 2, pp. 53–62, Oct. 2013.

[19] Z. K. Peng and F. L. Chu, "Application of the wavelet transform in machine condition monitoring and fault diagnostics: A review with bibliography," *Mech. Syst. Signal Process.*, vol. 18, no. 2, pp. 199–221, Mar. 2004.

[20] Q. Qian, B. Zhang, C. Li, Y. Mao, and Y. Qin, "Federated transfer learning for machinery fault diagnosis: A comprehensive review of technique and application," *Mech. Syst. Signal Process.*, vol. 223, Jan. 2025, Art. no. 111837.

[21] J. Jiao, H. Li, J. Lin, and H. Zhang, "Entropy-oriented domain adaptation for intelligent diagnosis of rotating machinery," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 2, pp. 1239–1249, Feb. 2024.

[22] Y. Ding, M. Jia, J. Zhuang, Y. Cao, X. Zhao, and C.-G. Lee, "Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions," *Rel. Eng. Syst. Saf.*, vol. 230, Feb. 2023, Art. no. 108890.

[23] Q. Yu, J. Li, Z. Li, and J. Zhang, "A clustering K-SVD-based sparse representation method for rolling bearing fault diagnosis," *Insight-Non-Destructive Test. Condition Monitor.*, vol. 63, no. 3, pp. 160–167, Mar. 2021.

[24] J. Li, W. Luo, M. Bai, and M. Song, "Fault diagnosis of high-speed rolling bearing in the whole life cycle based on improved grey wolf optimizer-least squares support vector machines," *Digit. Signal Process.*, vol. 145, Feb. 2024, Art. no. 104345.

[25] S. Wang, W. Wang, and S. Song, "A rolling bearing failure feature extraction approach based on IBWO-VME-MCKD," *J. Mech. Sci. Technol.*, vol. 38, no. 10, pp. 5255–5280, Oct. 2024.

[26] M. Xu, Q. Yu, S. Chen, and J. Lin, "Rolling bearing fault diagnosis based on CNN-LSTM with FFT and SVD," *Information*, vol. 15, no. 7, p. 399, Jul. 2024.

[27] A. Khorram, M. Khalooei, and M. Rezghi, "End-to-end CNN + LSTM deep learning approach for bearing fault diagnosis," *Appl. Intell.*, vol. 51, no. 2, pp. 736–751, Feb. 2021.

[28] Y. Hou, J. Wang, Z. Chen, J. Ma, and T. Li, "Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved transformer," *Eng. Appl. Artif. Intell.*, vol. 124, Sep. 2023, Art. no. 106507.

[29] X. Zhang, C. He, Y. Lu, B. Chen, L. Zhu, and L. Zhang, "Fault diagnosis for small samples based on attention mechanism," *Measurement*, vol. 187, Jan. 2022, Art. no. 110242.

[30] J.-X. Liao, H.-C. Dong, Z.-Q. Sun, J. Sun, S. Zhang, and F.-L. Fan, "Attention-embedded quadratic network (qttention) for effective and interpretable bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[31] Y. Keshun, W. Puzhou, and G. Yingkui, "Toward efficient and interpretative rolling bearing fault diagnosis via quadratic neural network with bi-LSTM," *IEEE Internet Things J.*, vol. 11, no. 13, pp. 23002–23019, Jul. 2024.

[32] C. Chen, Z. Wang, J. Shi, D. Yue, G. Shi, and C. Wang, "Noise-resilient bearing fault diagnosis via hybrid attention-based multiscale ScConv and quaternion transformer," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–18, 2025.

[33] C. Chen, X. Li, and J. Shi, "Hierarchical gray wolf optimizer-tuned flexible residual neural network with parallel attention module for bearing fault diagnosis," *IEEE Sensors J.*, vol. 24, no. 12, pp. 19626–19635, Jun. 2024.

[34] C. Chen et al., "A triple domain adversarial neural network for bearing fault diagnosis," *Mech. Syst. Signal Process.*, vol. 238, Sep. 2025, Art. no. 113202.

[35] J. Hendriks, P. Dumond, and D. A. Knox, "Towards better benchmarking using the CWRU bearing fault dataset," *Mech. Syst. Signal Process.*, vol. 169, Apr. 2022, Art. no. 108732.

[36] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[37] H. Shao, Y. Lai, H. Liu, J. Wang, and B. Liu, "LSFConvformer: A lightweight method for mechanical fault diagnosis under small samples and variable speeds with time-frequency fusion," *Mech. Syst. Signal Process.*, vol. 236, Aug. 2025, Art. no. 113016.

[38] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—A tutorial," *Mech. Syst. signal Process.*, vol. 25, no. 2, pp. 485–520, 2010.

[39] Y. Xiao, H. Shao, and B. Liu, "Evaluating calibration of deep fault diagnostic models under distribution shift," *Comput. Ind.*, vol. 171, Oct. 2025, Art. no. 104334.

[40] S. Han, S. Sun, Z. Zhao, Z. Luan, and P. Niu, "Deep residual multiscale convolutional neural network with attention mechanism for bearing fault diagnosis under strong noise environment," *IEEE Sensors J.*, vol. 24, no. 6, pp. 9073–9081, Mar. 2024.

[41] S. Li and X. Zhao, "A lightweight multi-feature fusion vision transformer bearing fault diagnosis method with strong local sensing ability in complex environments," *Meas. Sci. Technol.*, vol. 35, no. 6, Jun. 2024, Art. no. 065104.

[42] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, Feb. 2017.
[43] X. Zhong, F. Wang, and H. Ban, "Development of a plug-and-play anti-noise module for fault diagnosis of rotating machines in nuclear power plants," *Prog. Nucl. Energy*, vol. 151, Sep. 2022, Art. no. 104344.
[44] A. Shenfield and M. Howarth, "A novel deep learning model for the detection and identification of rolling element-bearing faults," *Sensors*, vol. 20, no. 18, p. 5112, Sep. 2020.
[45] Y. Han et al., "MT-ConvFormer: A multitask bearing fault diagnosis method using a combination of CNN and transformer," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025.
[46] Y.-L. Xu, X.-X. Li, D.-R. Chen, and H.-X. Li, "Learning rates of regularized regression with multiple Gaussian kernels for multi-task learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5408–5418, Nov. 2018.
[47] Z. Liu, H. Wang, J. Liu, Y. Qin, and D. Peng, "Multitask learning based on lightweight 1DCNN for fault diagnosis of wheelset bearings," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.
[48] K. Su, J. Liu, and H. Xiong, "Hierarchical diagnosis of bearing faults using branch convolutional neural network considering noise interference and variable working conditions," *Knowl.-Based Syst.*, vol. 230, Oct. 2021, Art. no. 107386.
[49] F. Wang, R. Liu, Q. Hu, and X. Chen, "Cascade convolutional neural network with progressive optimization for motor fault diagnosis under nonstationary conditions," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2511–2521, Apr. 2021.

Junjie Liu received the Ph.D. degree from Aberystwyth University, Aberystwyth, U.K., in 2014.

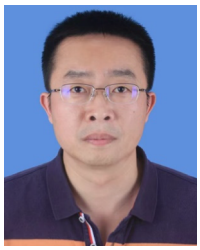
He is currently an Associate Professor with Zhaoqing University, Zhaoqing, China. His main research interests include deep learning and its applications (such as graph neural networks), 3-D image processing, machine vision, pattern recognition, and machine learning.

Jiaxian Zhu received the M.Sc. degree in computer science from Guangdong University of Technology, Guangzhou, China, in 2006.

She is currently an Associate Professor with the School of Computer Science and Software Engineering, Zhaoqing University, Zhaoqing, China. Her current research interests include scheduling optimization, large-scale resource scheduling, and deep learning and its applications.

Weihua Bai received the Ph.D. degree in computer science and engineering from the South China University of Technology, Guangzhou, China, in 2017.

He is currently a Professor with the School of Computer Science and Software Engineering, Zhaoqing University, Zhaoqing, China. His research interests include cloud computing, scheduling optimization, large-scale resource scheduling, and deep learning and its applications.



Huibing Zhang received the B.S. and M.S. degrees in computer science and technology from Guilin University of Electronic Technology, Guilin, China, in 2000 and 2005, respectively, and the Ph.D. degree in computer science and technology from Beijing University of Technology, Beijing, China, in 2012.

He is currently a Professor with Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology. He presided over and participated in a number of NSFC projects. His research interests include educational big data, the Internet of Things, and AI.

Lianghai Wu received the B.S. degree in Computer Science from South China Normal University, Guangzhou, in 2000, the M.S. degree in software engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2006, and the Ph.D. degree in computer technology and application from Macau University of Science and Technology, Macau, China, in 2023.

He is currently an Associate Professor with the School of Computer Science, Guangdong University of Petrochemical Technology, Maoming, China. His current research interests include computer vision, remote sensing data processing and analysis, and fault diagnosis.

Teng Zhou (Member, IEEE) received the M.Eng. degree from Sun Yat-sen University (SYSU), Guangzhou, China, in 2012, under the supervision of Prof. Xiaonan Luo, and the Ph.D. degree from the South China University of Technology (SCUT), Guangzhou, in 2017, under the supervision of Prof. Guoqiang Han.

He was a Post-Doctoral Research Fellow with the Center of Smart Health, The Hong Kong Polytechnic University, Hong Kong. He is currently a Full Professor with the School of Cyberspace Security (School of Cryptology), Hainan University, Haikou, China. His research interests include spatiotemporal data analysis, traffic flow theory, and graph convolutional networks.

Keqin Li (Fellow, IEEE) received the B.S. degree in computer science from Tsinghua University, Beijing, China, in 1985, and the Ph.D. degree in computer science from the University of Houston, Houston, TX, USA, in 1990.

He is currently a SUNY Distinguished Professor at the State University of New York at New Paltz, New Paltz, NY, USA, and a National Distinguished Professor with Hunan University, Hunan, China.

Dr. Li is a member of the SUNY Distinguished Academy. He is an AAAS Fellow, an AAIA Fellow, an ACIS Fellow, and an AIIA Fellow. He is a member of the European Academy of Sciences and Arts. He is a member of Academia Europaea (Academician of the Academy of Europe).