Dual-Branch Noise-Aware Representation Learning for Fault Diagnosis

Duerguli Maimaititusun[1], Gen Ge[2], HaiChuan Hu[2], Jianbin Jiao[3,*]
[1]School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China
[2]China Chemical Safety Association, Beijing, China
[3]School of Emergency Management Science and Engineering, University of Chinese Academy of Sciences, Beijing, China
duergulimaimaititusun22@mails.ucas.ac.cn, gegen@emah.com.cn, hosea@emah.com.cn, jiaojb@ucas.ac.cn

***Abstract*—Robust fault diagnosis under mixed noise and low signal-to-noise ratio (SNR) remains challenging because weak fault signatures can be submerged by non-stationary disturbances. Existing approaches typically operate on single feature domains and employ relatively simple fusion strategies, failing to fully exploit complementary information from multi-domain representations and dual-sensor deployments under low SNR conditions. To address these limitations, we propose a dual-branch noise-aware representation learning (DB-NAR) framework that jointly learns from two synchronized acquisition channels and extracts complementary representations in both time and time-frequency (TF) domains. First, differentiable denoising modules, including learnable wavelet soft-thresholding and adaptive Wiener filtering with frequency attention, are embedded for domain-specific noise suppression. Then, domain-specific feature extraction and hierarchical fusion mechanisms are proposed to capture and integrate complementary cues from dual-channel measurements. The time-domain branch employs a lightweight bidirectional depthwise convolution block that enables efficient 1D sequence modeling with long-range temporal context aggregation, while the TF-domain branch utilizes 2D convolutions to capture joint time-frequency patterns from spectrogram representations. Leveraging dual-channel spatial redundancy, a gated cross-channel fusion mechanism adaptively fuses dual-position features, with an explicit difference pathway for channel mismatch and propagation cue capture. At last, the time-domain and TF-domain representations are jointly fused for robust fault classification. Experiments on the two benchmark datasets for bearing fault diagnosis demonstrate the effectiveness of the proposed method under challenging noise conditions.**

***Index Terms*—Fault Diagnosis, Low SNR, Dual-branch Noise-aware Representation, Depthwise Convolution**

I. Introduction

Industrial condition monitoring frequently operates in harsh environments, where measured signals are corrupted by a mixture of disturbances such as broadband noise, impulsive interference, and colored noise induced by complex transmission paths [1], [2]. Under such settings, robust fault diagnosis becomes particularly difficult when the signal-to-noise ratio (SNR) is low: fault-related components may be weak, sparse, and non-stationary, and naive denoising or feature extraction can either fail to suppress the interference or inadvertently remove discriminative fault signatures [3].

Beyond noise corruption, practical monitoring systems often deploy multiple sensors or acquisition sources with different modalities and characteristics. Even when signals are synchronized, different acquisition channels can exhibit distinct statistics, frequency responses, and noise patterns, causing acquisition-dependent distribution shifts. This mismatch becomes more pronounced under mixed noise, where each channel may be dominated by different interference sources. Thus, methods relying on a single acquisition channel or a single feature domain, may suffer from unstable performance across SNR levels and operating conditions [4].

Recent deep learning approaches have attempted to address noise robustness in bearing fault diagnosis. Wide-kernel architectures such as WDCNN [5] employ large first-layer convolutions for coarse-scale temporal filtering. Multi-scale CNNs [6], [7] use parallel branches with different kernel sizes, yet their fixed-coefficient filtering lacks adaptability to varying noise characteristics. Self-attention based methods like SANet [8] introduce adaptive weighting to capture long-range dependencies, but degrade when fault signatures are severely masked by noise. MLSCA-CW [4] combines learnable soft-thresholding with multi-location fusion, achieving improved robustness under mixed-noise conditions.

However, existing methods predominantly assume single-type, low-intensity noise conditions and are typically validated on pure or weakly corrupted datasets [9], [10], which poorly reflect real industrial environments where bearings operate amidst multiple concurrent noise sources. Under mixed-noise conditions at low SNR levels (e.g., -9 to -3 dB), representative approaches, including multi-scale CNNs, wide-kernel networks, and attention-based architectures, suffer significant accuracy degradation due to their inability to adaptively disentangle weak fault features from strong, non-stationary disturbances. Furthermore, single-sensor strategies provide incomplete fault signatures that are easily masked by intense noise, further compromising diagnostic reliability.

To address these challenges, we propose a dual-branch noise-aware representation learning framework (DB-NAR) that jointly learns from two synchronized acquisition channels and extracts complementary representations in both time and time-frequency domains. The framework incorporates differentiable denoising modules tailored to each domain: a learnable wavelet soft-thresholding layer for adaptive time-domain

noise suppression [11], [12], and an adaptive Wiener filter with frequency attention for TF-domain enhancement. For feature extraction, the time-domain branch employs a lightweight Bidirectional Depthwise Convolution (Bi-DWConv) block that enables efficient 1D sequence modeling with long-range temporal context aggregation, while the TF-domain branch utilizes 2D convolutions to capture joint time-frequency patterns from spectrogram representations. To exploit the spatial redundancy inherent in dual-channel measurements, a gated cross-channel fusion mechanism adaptively combines features from both acquisition positions, with an explicit difference pathway to capture channel mismatch and propagation-related cues. The time-domain and TF-domain representations are then jointly fused for robust fault classification.

The proposed method is validated on two representative benchmark datasets under challenging noise conditions. The UofO [13] dataset provides heterogeneous dual-channel measurements, accelerometer vibration and encoder speed, offering complementary signal characteristics with different noise sensitivities. The CWRU [14] dataset contains homogeneous dual accelerometer channels (DE and FE) at different spatial locations. Experimental results demonstrate that the proposed method achieves superior performance on both datasets.

The main contributions of this paper are as follows:

- A **dual-branch noise-aware representation learning framework** (DB-NAR) is proposed that jointly models two synchronized channels across time and TF domains, with differentiable denoising modules embedded in each branch: learnable wavelet soft-thresholding for time-domain denoising and adaptive Wiener filtering with frequency attention for TF-domain enhancement.
- **Domain-specific feature extraction and hierarchical fusion** mechanisms are designed, including a lightweight Bi-DWConv block for efficient 1D temporal modeling, 2D convolutions for TF pattern extraction, and gated cross-channel fusion with an explicit difference pathway to capture complementary and mismatch cues from dual-channel measurements.
- **Comprehensive experiments** on UofO and CWRU datasets under mixed-noise conditions demonstrate the effectiveness of the proposed method. On UofO, DB-NAR achieves 94.81%, 98.52%, and 99.09% accuracy at $-9/-6/-3$ dB SNR, respectively. On CWRU, our model achieves 97.89%, 98.78%, and 100.00% accuracy at $-9/-6/-3$ dB SNR, respectively, consistently outperforming representative comparisons.

II. Methods

A. Problem Formulation and Overall Framework

The proposed DB-NAR model consists of two parallel branches: (i) a time-domain branch and (ii) a time-frequency (TF) branch, as illustrated in Fig. 1. Each branch encodes two channels independently (non-shared encoders) and then performs cross-channel fusion. The two channels in UofO dataset are accelerometer vibration and encoder speed, and DE and FE accelerometer signals for CWRU dataset.

Specifically, the time-domain branch (top of Fig. 1) incorporates a learnable wavelet denoising module for adaptive noise suppression, along with a lightweight Bidirectional Depthwise Convolution (Bi-DWConv) block to enable efficient 1D sequence modeling and long-range temporal context aggregation. The TF-domain branch (bottom of Fig. 1) adopts an adaptive Wiener denoising module for domain-specific enhancement, and uses 2D convolutions to capture joint time-frequency patterns from spectrograms. Finally, the representations from the time-domain and TF-domain branches are jointly fused to achieve robust fault classification.

Given a pair of synchronized 1D segments acquired from two channels, $x_{\mathrm{a}}, x_{\mathrm{b}} \in \mathbb{R}^{1\times L}$, the goal is to predict the fault label $y \in \{1, \ldots, N_c\}$ under mixed noise and low SNR. We learn a mapping $f_\Theta : (x_{\mathrm{a}}, x_{\mathrm{b}}) \mapsto \hat{\mathbf{y}}$ parameterized by Θ. Overall, the time-domain ($\mathbf{z}^{time}$) and TF-domain ($\mathbf{z}^{tf}$) representations are generated and concatenated for fault diagnosis:

$$\mathbf{z}^{time} = \text{time-domain branch}(x_{\mathrm{a}}, x_{\mathrm{b}}), \quad (1)$$

$$\mathbf{z}^{tf} = \text{TF-domain branch}(x_{\mathrm{a}}, x_{\mathrm{b}}), \quad (2)$$

$$\mathbf{z} = [\mathbf{z}^{time}; \mathbf{z}^{tf}], \quad (3)$$

$$\hat{\mathbf{y}} = \mathrm{MLP}(\mathbf{z}), \quad (4)$$

where $\mathbf{z}$ is the final joint feature representation, and $\hat{\mathbf{y}}$ is the final fault diagnosis result predicted by an MLP classifier.

B. Time-Domain Branch

The time-domain branch processes raw waveform inputs $\mathbf{x}_a^{time}, \mathbf{x}_b^{time} \in \mathbb{R}^{1\times L}$ from two measurement locations (e.g., drive-end and fan-end). For brevity, we describe the processing pipeline for location a; location b is processed identically.

Multi-scale convolutional frontend. To capture temporal patterns at different receptive fields, we first apply a multi-scale Conv1D frontend with different kernel sizes:

$$\mathbf{T}_a = \mathrm{MultiScaleConv1D}(\mathbf{x}_a^{time}) \in \mathbb{R}^{D\times L}, \quad (5)$$

where D denotes the number of feature channels. The multi-scale features from different kernel sizes are concatenated and refined through channel attention weighting followed by a fusion convolution layer.

Learnable wavelet denoising. To achieve noise-aware representation learning, we insert a differentiable *Wavelet Denoising Layer* that performs wavelet-like shrinkage in an end-to-end trainable manner. Different from a standard discrete wavelet transform (DWT) with fixed analysis/synthesis filters [11], [15], our layer implements a *learnable analysis–synthesis filterbank* using depthwise grouped convolutions.

Let $\mathbf{h}^{(0)} = \mathbf{T}_a$. Each decomposition level $\ell \in \{1, \ldots, K\}$ applies a grouped 1D convolution with stride 2:

$$\mathbf{y}^{(\ell)} = \mathrm{Conv1D}_{\mathrm{dw}}^{(\ell)}(\mathbf{h}^{(\ell-1)}) \in \mathbb{R}^{2D\times L_\ell}, \quad (6)$$

where $\mathrm{Conv1D}_{\mathrm{dw}}$ denotes a depthwise grouped convolution ($\texttt{groups}= D$) with kernel size 4 that learns an analysis filter pair for each channel, and $L_\ell = L_{\ell-1}/2$ is the downsampled

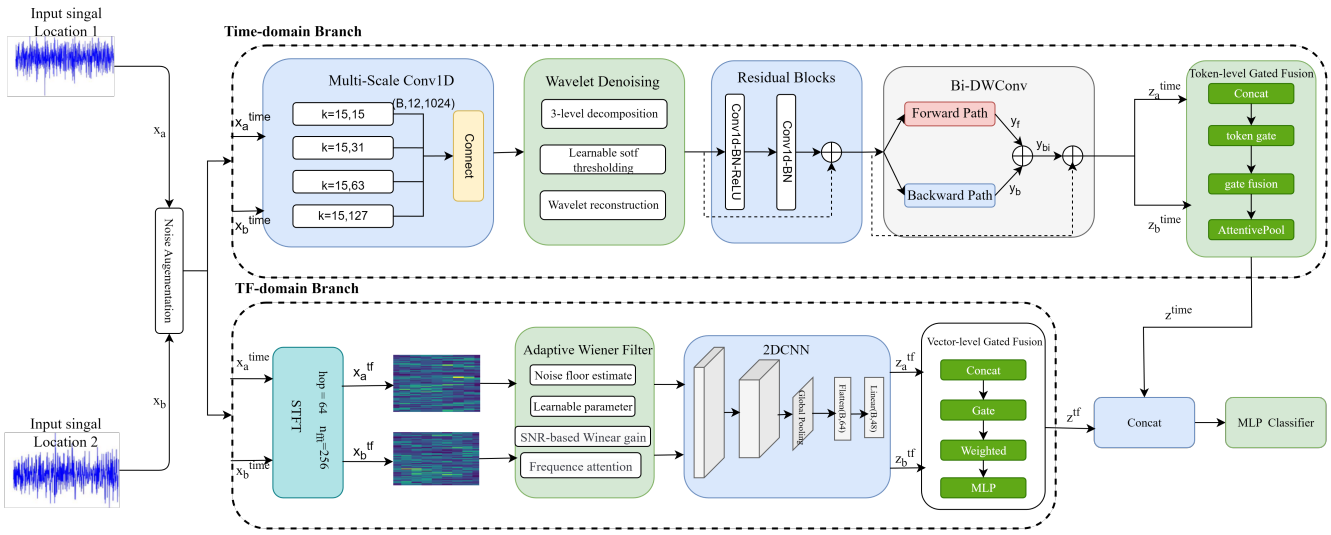


Fig. 1: Pipeline of the proposed dual-branch noise-aware representation learning (DB-NAR). DB-NAR receives two synchronized input channels that are processed in parallel through both the time-domain branch and the time-frequency (TF) branch.

length. The output is split along the channel dimension into low- and high-frequency subbands:

$$[\mathbf{c}_L^{(\ell)}, \mathbf{c}_H^{(\ell)}] = \mathrm{Split}(\mathbf{y}^{(\ell)}), \quad \mathbf{c}_L^{(\ell)}, \mathbf{c}_H^{(\ell)} \in \mathbb{R}^{D \times L_\ell}, \tag{7}$$

where the low-frequency component $\mathbf{c}_L^{(\ell)}$ is propagated to the next level, i.e., $\mathbf{h}^{(\ell)} = \mathbf{c}_L^{(\ell)}$.

At each level, we denoise the high-frequency coefficients via learnable soft-thresholding [11], [16]:

$$\tilde{\mathbf{c}}_H^{(\ell)} = \mathrm{sign}(\mathbf{c}_H^{(\ell)}) \odot \max\big(|\mathbf{c}_H^{(\ell)}| - \boldsymbol{\tau}_\ell, 0\big), \tag{8}$$

where the threshold $\boldsymbol{\tau}_\ell = \mathrm{softplus}(\boldsymbol{\theta}_\ell)$ with learnable parameter $\boldsymbol{\theta}_\ell \in \mathbb{R}^D$ is channel-wise and constrained to be positive. This design suppresses noise-dominated high-frequency responses while preserving task-relevant fault signatures.

Reconstruction is performed in a coarse-to-fine manner using grouped transposed convolutions:

$$\tilde{\mathbf{h}}^{(\ell-1)} = \mathrm{ConvT1D}_{\mathrm{dw}}^{(\ell)}\big([\mathbf{c}_L^{(\ell)}; \tilde{\mathbf{c}}_H^{(\ell)}]\big), \tag{9}$$

where $\mathrm{ConvT1D}_{\mathrm{dw}}$ denotes a depthwise transposed convolution with kernel size 4, stride 2, and $\texttt{groups}= D$. We enforce exact length consistency by linear interpolation at each level. The denoised output is $\tilde{\mathbf{T}}_a = \mathrm{WaveletDenoise}(\mathbf{T}_a) \in \mathbb{R}^{D \times L}$, and similarly $\tilde{\mathbf{T}}_b$ for location b.

Residual refinement. After wavelet denoising, residual blocks further refine local discriminative patterns:

$$\mathbf{R}_a = \mathrm{ResBlocks}(\tilde{\mathbf{T}}_a) \in \mathbb{R}^{D \times L}, \tag{10}$$

and similarly $\mathbf{R}_b$ for location b. Each residual block consists of two Conv1D layers with batch normalization and ReLU activation, followed by a skip connection. The outputs are then transposed to token sequences $\mathbf{R}_a^\top, \mathbf{R}_b^\top \in \mathbb{R}^{L \times D}$.

Sequence modeling with Bi-DWConv. To model long-range temporal dependencies efficiently, we adopt a Bidirectional Gated Depthwise Convolution (Bi-DWConv) block [17], [18] instead of self-attention. Given a token sequence $\mathbf{X} \in \mathbb{R}^{L \times D}$, the gated depthwise convolution is defined as:

$$\mathrm{DWConv}(\mathbf{X}) = \mathbf{U} \odot \sigma(\mathbf{V}), \tag{11}$$

where $[\mathbf{U}, \mathbf{V}] = \mathrm{PW}(\mathrm{DW}(\mathrm{LN}(\mathbf{X})))$. Here LN denotes layer normalization, DW is depthwise convolution with kernel size k [19], [20], PW is pointwise projection that doubles the channel dimension, and $\sigma(\cdot)$ is the sigmoid gating function.

To capture bidirectional context, we process both forward and reversed sequences:$\mathrm{Bi\text{-}DWConv}(\mathbf{X}) = \mathrm{LN}\big(\mathbf{X} + \mathrm{DWConv}(\mathbf{X}) + \mathrm{Rev}(\mathrm{DWConv}(\mathrm{Rev}(\mathbf{X}))\big)$, where $\mathrm{Rev}(\cdot)$ denotes sequence reversal along the temporal dimension. This yields $\mathcal{O}(Lk)$ complexity for kernel size k, avoiding the quadratic $\mathcal{O}(L^2)$ cost of self-attention.

The time-domain representations for both locations are:

$$\mathbf{z}_a^{time} = \text{Bi-DWConv}(\mathbf{R}_a^\top) \in \mathbb{R}^{L \times D}, \tag{12}$$

and similarly $\mathbf{z}_b^{time} \in \mathbb{R}^{L \times D}$ for location b.

C. Time-Frequency Branch

The TF branch computes the short-time Fourier transform (STFT) magnitude [21]. Given a time signal x, the STFT is defined as:

$$\mathrm{STFT}(x)[f,t] = \sum_m x[m]\, w[m - tH]\, e^{-j2\pi f m/N}, \tag{13}$$

where $w(\cdot)$ is a Hann window, H is the hop size, and N is the FFT size.

For measurement location a, we compute the time-frequency representation:

$$\mathbf{x}_a^{tf} = |\mathrm{STFT}(\mathbf{x}_a^{time})| \in \mathbb{R}^{1 \times F \times T}, \tag{14}$$

where $F = N/2 + 1$ is the number of frequency bins and $T = \lfloor (L-N)/H \rfloor + 1$ is the number of time frames. Similarly, $\mathbf{x}_b^{tf} = |\mathrm{STFT}(\mathbf{x}_b^{time})| \in \mathbb{R}^{1 \times F \times T}$ for location b.

To suppress noise while preserving fault-related features, we apply an adaptive Wiener-like filter. For location a, the per-bin SNR and Wiener gain are:

$$\mathrm{SNR}_a[f,t] = \frac{(\mathbf{x}_a^{tf}[f,t])^2}{\widehat{\mathbf{N}}_a[f]^2 + \epsilon}, \quad \mathbf{G}_a[f,t] = \frac{\mathrm{SNR}_a[f,t]}{1+\mathrm{SNR}_a[f,t]}, \tag{15}$$

where $\widehat{\mathbf{N}}_a$ is a noise estimate computed as the 10th percentile over time. The filtered spectrograms are:

$$\tilde{\mathbf{x}}_a^{tf} = \mathbf{G}_a \odot \mathbf{x}_a^{tf}, \quad \tilde{\mathbf{x}}_b^{tf} = \mathbf{G}_b \odot \mathbf{x}_b^{tf}. \tag{16}$$

The filtered spectrograms are log-compressed and encoded by a 2D CNN with global average pooling:

$$\mathbf{z}_a^{tf} = \mathrm{GAP}\big(\mathrm{Conv2D}(\log(1+\tilde{\mathbf{x}}_a^{tf}))\big) \in \mathbb{R}^D, \tag{17}$$

and similarly $\mathbf{z}_b^{tf} \in \mathbb{R}^D$ for location b.

D. Cross-Channel Fusion

Token-level gated fusion (Time branch). The two channel representations $\mathbf{z}_a^{time}, \mathbf{z}_b^{time} \in \mathbb{R}^{L\times D}$ are fused via a learnable gating mechanism that adaptively weights the contribution of each channel at each token position. Let $\mathbf{z}_{a,i}^{time}, \mathbf{z}_{b,i}^{time} \in \mathbb{R}^D$ denote the i-th token of each channel:

$$\mathbf{g}_i = \sigma\big(\phi([\mathbf{z}_{a,i}^{time}; \mathbf{z}_{b,i}^{time}])\big) \in \mathbb{R}^D, \tag{18}$$

$$\mathbf{f}_i = \mathbf{g}_i \odot \mathbf{z}_{a,i}^{time} + (1-\mathbf{g}_i) \odot \mathbf{z}_{b,i}^{time}, \tag{19}$$

where $[\cdot;\cdot]$ denotes concatenation, $\phi : \mathbb{R}^{2D} \to \mathbb{R}^D$ is a two-layer MLP, $\sigma(\cdot)$ is the sigmoid function, and $i \in \{1,\dots,L\}$ indexes the token position. This token-level gating allows the model to dynamically select informative features from either channel based on local signal characteristics.

Then the fused token sequence is aggregated into a fixed-dimensional representation via attentive pooling:

$$\mathbf{z}^{time} = \sum_{i=1}^{L} \alpha_i \mathbf{f}_i, \quad \alpha_i = \frac{\exp(\mathbf{w}^\top \mathbf{f}_i)}{\sum_{j=1}^{L}\exp(\mathbf{w}^\top \mathbf{f}_j)}, \tag{20}$$

where $\mathbf{w} \in \mathbb{R}^D$ is a learnable attention vector. The attention weights α_i are computed via softmax over the token-wise similarity scores, enabling the model to emphasize discriminative temporal positions while suppressing less informative ones. The resulting $\mathbf{z}^{time} \in \mathbb{R}^D$ is the final time branch output.

Vector-level gated fusion (TF branch). Given TF embeddings $\mathbf{z}_a^{tf}, \mathbf{z}_b^{tf} \in \mathbb{R}^D$, we use gated fusion plus an explicit difference pathway:

$$\mathbf{g} = \sigma\big(\psi([\mathbf{z}_a^{tf}; \mathbf{z}_b^{tf}])\big) \in \mathbb{R}^D, \tag{21}$$

$$\mathbf{z}^{tf} = \mathbf{g} \odot \mathbf{z}_a^{tf} + (1-\mathbf{g}) \odot \mathbf{z}_b^{tf} + \rho(\mathbf{z}_a^{tf} - \mathbf{z}_b^{tf}), \tag{22}$$

where $\psi : \mathbb{R}^{2D} \to \mathbb{R}^D$ is a two-layer MLP that computes the gate, and $\rho : \mathbb{R}^D \to \mathbb{R}^D$ is a linear projection that encodes the difference between the two locations. The output $\mathbf{z}^{tf} \in \mathbb{R}^D$ is the fused TF-domain representation.

TABLE I: Dataset Configurations for UofO and CWRU.

Item	UofO	CWRU
Channels	Ch1/Ch2	DE/FE
Modalities	Vib. + Speed	Vib. + Vib.
Sampling rate	∼20 kHz	12 kHz
Window length L	1024	1024
Window stride	1024	1024
Number of classes	3	4

E. Training Objective

The whole network is trained using the cross-entropy loss:

$$\mathcal{L}_{\mathrm{ce}} = -\sum_{c=1}^{N_c} \mathbf{1}[y=c] \log p_\Theta(y=c \mid \mathbf{x}_a^{time}, \mathbf{x}_b^{time}), \tag{23}$$

where p_Θ is the softmax output of the classifier.

To enhance generalization to mixed-noise distributions, we apply noise augmentation during training. Let s denote a clean segment and n a sampled noise segment. We generate noisy input x by scaling noise to a target SNR:

$$\alpha = \sqrt{\frac{P_s/10^{\mathrm{SNR}/10}}{P_n}}, \tag{24}$$

$$x = s + \alpha n, \tag{25}$$

where $P_s = \frac{1}{L}\sum_{i=1}^{L} s_i^2$ and $P_n = \frac{1}{L}\sum_{i=1}^{L} n_i^2$ denote the signal and noise power, respectively. We consider four noise types for evaluation: Gaussian, Laplacian, Brownian, and Violet. For mixed noise, we combine these sources with equal weights and scale to match the target SNR. During training, with probability 0.5, each input is corrupted by additive noise at a random SNR uniformly sampled from $[-15, 10]$ dB.

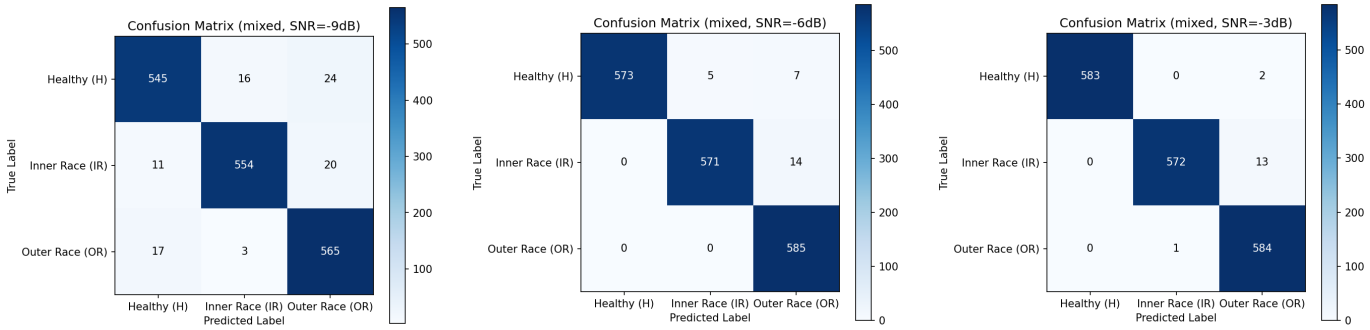
III. Experiments

A. Datasets and Preprocessing

We evaluate on two datasets to demonstrate the effectiveness of our method under mixed noise and validate generalization, as shown in Table I.

UofO Dataset. Each sample contains two synchronized channels:accelerometer vibration and encoder speed. The raw sampling rate is approximately 200 kHz and we downsample by a factor of 10 to approximately 20 kHz [13]. We segment raw signals into non-overlapping windows with size 10240 and stride 10240; after downsampling, each segment has length $L = 1024$. The label space contains $C = 3$ categories: {H, I, O} representing healthy, inner race fault, and outer race fault conditions. We adopt a trial-level split protocol to prevent leakage: samples from the same trial appear in only one split.

CWRU Dataset. We use the Case Western Reserve University bearing dataset [14] with dual-location inputs from DE and FE accelerometer channels. The sampling rate is 12 kHz. We use sliding windows with length $L = 1024$ and stride 1024. The four classes, healthy, inner race, outer race, and ball fault conditions, are represent as {H, IR, OR, BF}.



(a) SNR = -9 dB (b) SNR = -6 dB (c) SNR = -3 dB

Fig. 2: Confusion matrices on the UofO dataset under mixed-noise conditions at different SNR levels. Rows denote true labels and columns denote predicted labels. Classes are Healthy (H), Inner race fault (I), and Outer race fault (O).

TABLE II: Comparison Results Under Mixed-Noise Low SNR Conditions on UofO dataset (%).

Method	-9 **dB**	-6 **dB**	-3 **dB**
LR	37.50	38.23	43.35
SANet	53.84	55.03	58.11
WDCNN	68.09	70.48	73.89
MC-CNN	68.17	70.73	74.15
QCNN	72.70	74.06	75.34
MLSCA-CW (2-loc)	73.98	81.66	86.26
DB-NAR (Ours)	**94.81**	**98.52**	**99.09**
Improvement	+20.83	+16.86	+12.83

TABLE III: Results Under Mixed Noise Across Different SNRs on CWRU dataset (%).

Methods	-9 **dB**	-6 **dB**	-3 **dB**	**0 dB**	**3 dB**	**6 dB**	**9 dB**
LR	52.94	53.78	57.98	58.82	62.61	65.13	66.39
SANet	68.49	81.09	84.03	90.34	97.48	99.16	100.00
WDCNN	82.77	84.64	91.18	99.58	100.00	100.00	100.00
MC-CNN	82.77	86.22	99.56	100.00	100.00	100.00	100.00
QCNN	89.08	92.02	92.44	98.32	100.00	100.00	100.00
MLSCA-CW	95.38	97.90	99.58	100.00	100.00	100.00	100.00
DB-NAR	**97.89**	**98.78**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

B. Experiment Settings and Metrics

Noise Types and SNR Settings. We consider four noise types: Gaussian (white), Laplacian (impulsive), Brownian (low-frequency), and Violet (high-frequency). Mixed noise is constructed by equally combining the four noise types. We evaluate at SNR levels from -9 dB to 9 dB with a step size of 3 dB: $\{-9, -6, -3, 0, 3, 6, 9\}$ dB. We report classification accuracy on the test set under the SNR setting.

Implementation Details. All models are implemented in PyTorch and trained with the Adam optimizer. We use a learning rate of 1×10^{-3}, weight decay of 1×10^{-5}, batch size of 32, and train for up to 100 epochs with early stopping (patience = 15). The best checkpoint is selected based on validation accuracy. For the TF branch, we compute STFT magnitude with FFT size $N = 256$, window length $w = 256$, and hop size $H = 64$. Key hyperparameters of DB-NARare set as follows: wavelet decomposition levels = 3, feature dimension = 48, number of time residual blocks = 2, and kernel size of Bi-DWConv = 9.

C. Comparison Results Under Mixed Noise

We compare DB-NAR with the following representative approaches: (1) **LR**: Logistic Regression with hand-crafted features; (2) **SANet** [8]: CNN with self-attention for sequence modeling; (3) **WDCNN** [5]: wide first-layer kernel CNN; (4) **MC-CNN** [7]: multi-scale CNN with parallel branches; (5) **QCNN** [22]: quaternion-valued CNN; (6) **MLSCA-CW** [4]: multi-location soft-thresholding with channel-wise attention. All methods are evaluated on the same datasets with the same mixed-noise protocol for fair comparison.

Table II reports the comparison results on UofO under mixed-noise conditions at low SNR levels. Our model achieves 94.81%, 98.52%, and 99.09% accuracy at -9 dB, -6 dB, and -3 dB, respectively, consistently outperforming all baselines. Compared with the strongest comparison method MLSCA-CW (two locations), our method improves accuracy by 20.83, 16.86, and 12.83 percentage points at -9/-6/-3 dB, respectively. This significant improvement demonstrates the effectiveness of the proposed dual-branch architecture with learnable denoising modules and efficient sequence modeling.

As shown in Fig. 2 (a)–(c), the proposed model maintains strong class-wise discrimination under mixed noise. At SNR $= -9$ dB, the remaining errors mainly occur between the two fault categories (I and O), indicating that severe mixed noise can blur subtle fault-specific signatures. When SNR increases to -6 dB and -3 dB, the off-diagonal entries shrink significantly and predictions become highly concentrated on the diagonal, suggesting improved separability among H/I/O as noise contamination is reduced. These confusion matrices are consistent with the accuracy gains reported in Table II.

Table III reports the results on CWRU under mixed-noise corruption across SNRs from -9 dB to 9 dB. Our method achieves the highest accuracy across all SNR settings, demonstrating excellent robustness even under severe noise corruption. At the most challenging -9 dB SNR, the proposed

TABLE IV: Ablation Study on UofO dataset (%). Wavelet, Bi-DWConv, TF Branch, Wiener, and Time Branch are abbreviated as Wa, B-DW, TF, Wi, and T, respectively.

Methods	**Wa**	**B-DW**	**TF**	**Wi**	**T**	-9 **dB**	-6 **dB**	-3 **dB**
a		✓	✓	✓	✓	94.36	98.40	98.97
b	✓		✓	✓	✓	92.99	98.35	99.20
c	✓	✓			✓	93.16	97.89	99.03
d	✓	✓	✓		✓	93.16	98.06	98.86
e			✓	✓		89.06	97.04	98.75
DB-NAR	✓	✓	✓	✓	✓	**94.81**	**98.52**	**99.09**

model achieves 97.89% accuracy.As SNR increases to -3 dB and above, our method achieves 100.0% accuracy consistently. These results indicate that the proposed framework achieves robust fault discrimination.

D. Ablation Study

Table IV presents the ablation results. Removing wavelet denoising (a) causes a 0.45% accuracy drop at -9 dB, confirming the importance of learnable time-domain denoising. Removing Bi-DWConv (b) leads to a 1.82% drop, showing the value of efficient sequence modeling for aggregating temporally distributed fault signatures. Removing the TF branch (c) or Wiener filtering (d) each causes a 1.65% drop, showing the complementary value of time-frequency features and adaptive TF-domain denoising. The largest degradation (5.75%) occurs when removing the time branch entirely (e), confirming that time-domain provide essential discriminative information. These results validate that each component contributes to the overall performance, with the dual-branch design and channel-specific encoding being particularly important.

Overall, the experimental results validate the effectiveness of the proposed method. The dual-branch design provides complementary information: the time branch captures fine-grained temporal patterns while the TF branch reveals spectral signatures, removing either branch degrades performance. The learnable denoising modules adapt their behavior based on task supervision, outperforming fixed preprocessing. And the Bi-DWConv block enables efficient long-range temporal modeling with lower complexity. Hierarchical fusion with difference modeling and channel-specific encoders is critical for heterogeneous dual-channel inputs.

IV. Conclusion

Robust fault diagnosis under severe mixed-noise conditions remains challenging for industrial bearing monitoring systems. This paper proposes DB-NAR, a dual-location dual-branch noise-aware framework that addresses this challenge through three key innovations: learnable wavelet soft-thresholding and adaptive Wiener denoising for noise suppression, efficient bidirectional depthwise convolution for long-range temporal modeling, and hierarchical fusion with explicit difference representations for multi-source information integration. Extensive experiments on UofO and CWRU datasets validate the effectiveness of the proposed method under mixed-noise low-SNR conditions.

References

[1] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—a tutorial," *Mechanical Systems and Signal Processing*, vol. 25, no. 2, pp. 485–520, 2011. [Online]. Available:

[2] D. Liu, L. Cui, W. Cheng, D. Zhao, and W. Wen, "Rolling bearing fault severity recognition via data mining integrated with convolutional neural network," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 5768–5777, 2022.

[3] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, 2020. [Online]. Available:

[4] Y. Zhang, L. Lin, J. Wang, W. Zhang, S. Gao, and Z. Zhang, "Attention activation network for bearing fault diagnosis under various noise environments," *Scientific Reports*, vol. 15, no. 1, p. 977, 2025.

[5] 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, 2017. [Online]. Available:

[6] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5990–5998, 2018.

[7] W. Huang, J. Cheng, Y. Yang, and G. Guo, "An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis," *Neurocomputing*, vol. 359, pp. 77–92, 2019. [Online]. Available:

[8] H. Lv, J. Chen, T. Pan, T. Zhang, Y. Feng, and S. Liu, "Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application," *Measurement*, vol. 199, p. 111594, 2022. [Online]. Available:

[9] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," *ISA Transactions*, vol. 95, pp. 295–305, 2019. [Online]. Available:

[10] D. Yao, H. Liu, J. Yang, and X. Li, "A lightweight neural network with strong robustness for bearing fault diagnosis," *Measurement*, vol. 159, p. 107756, 2020. [Online]. Available:

[11] D. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[12] K. Isogawa, T. Ida, T. Shiodera, and T. Takeguchi, "Deep shrinkage convolutional neural network for adaptive noise reduction," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 224–228, 2018.

[13] H. Huang and N. Baddour, "Bearing vibration data collected under time-varying rotational speed conditions," *Data in Brief*, vol. 21, pp. 1745–1749, 2018. [Online]. Available:

[14] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve university data: A benchmark study," *Mechanical Systems and Signal Processing*, vol. 64-65, pp. 100–131, 2015. [Online]. Available:

[15] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[16] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4681–4690, 2020.

[17] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2022. [Online]. Available:

[18] M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré, "Hyena hierarchy: Towards larger convolutional language models," 2023. [Online]. Available:

[19] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available:

[21] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[22] 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 Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.