Physics-Informed Multi-Scale Network with Loss-Guided Curriculum Learning for Robust Fault Diagnosis

Panfeng Bao[1,2,3], Wenjun Yi[1], Yue Zhu[2], Yufeng Shen[2], and Haotian Peng[4,*]

[1] *National Laboratory of Transient Physics, Nanjing University of Science and Technology, Nanjing 210094, China*
223121110079@njust.edu.cn
wjy@njust.edu.cn

[2] *Changsha Aeronautical Vocational and Technical College, Changsha 410124, China*
zhuyue19931277@163.com
15116220284@163.com

[3] *Hunan Provincial Innovation Center for Applied Technology in Intelligent Manufacturing of Aviation Equipment, Changsha 410124, China*

[4] *Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110069, China*
penghaotian@sia.cn

Abstract

Reliable fault diagnosis of rotating machinery is critical, yet early weak fault impulses are frequently buried in severe compound interference from mechanical harmonics and environmental noise. To address this, a novel Physics-Informed Multi-Scale Network (PI-MSN) is proposed. Variational Mode Decomposition (VMD) is first employed to decouple raw signals into distinct physical frequency bands. Subsequently, a Physics-Informed Channel Attention (PICA) module jointly evaluates the Kurtosis and Root Mean Square of each channel to autonomously highlight fault impulses and suppress harmonic interference. A Multi-Scale Feature Extractor then captures comprehensive fault characteristics. Furthermore, a closed-loop Loss-Guided Smooth Interference Scheduler (LGSIS) dynamically regulates injected interference during training based on real-time loss, significantly mitigating catastrophic forgetting. Extensive experiments on the CWRU and HUST datasets demonstrate the framework's exceptional robustness. The highly lightweight PI-MSN achieves state-of-the-art diagnostic accuracy, sustaining over 98% accuracy even under severe -4 dB compound interference, proving that physical interpretability effectively eliminates the reliance on massive parameter stacking.

https://doi.org/10.36001/IJPHM.2026.v17i2.4766

1. Introduction

Rotating machinery, such as gearboxes and induction motors, serves as the backbone of modern industrial systems. The reliable condition monitoring and early fault diagnosis of these critical components are essential to prevent unexpected breakdowns and minimize economic losses (Shao, Lin, Zhang, Galar, & Kumar, 2021; Wei, Tian, Cui, Zheng, & Liu, 2023). In recent years, data-driven methods, particularly Deep Learning (DL), have achieved remarkable success in fault diagnosis due to their powerful feature extraction capabilities (Dong, Jiang, Yao, Mu, & Yang, 2024; Y. Zhao, Zhang, Li, Bu, & Han, 2023). Concurrently, new paradigms such as transformer-based architectures (Ji & Zhao, 2024) and domain generalization techniques (C. Zhao, Zio, & Shen, 2024) are actively being explored to enhance modeling capacity and handle complex industrial scenarios. Traditional end-to-end DL models, such as one-dimensional Convolutional Neural Networks (1D-CNNs), have demonstrated near-perfect accuracy when evaluated on ideal, clean laboratory datasets (Chopra, Kumar, & Yadav, 2025; Luo, Qiu, Wu, Zhao, & Zhang, 2025).

However, a significant discrepancy exists between laboratory environments and real-world industrial applications. In actual operational settings, the early weak fault impulses of bearings are invariably buried in severe compound interference (Ruan, Wang, Yan, & Gühmann, 2023; Pancaldi, Dibiase, & Cocconcelli, 2023), leading to an extremely low Signal-to-Interference-plus-Noise Ratio (SINR). This compound interference primarily originates from two sources:

macroscopic low-frequency mechanical harmonics (e.g., rotor imbalance, gear meshing) and severe environmental Gaussian white noise (Niu, Liu, Wang, Ziehl, & Zhang, 2023). Conventional DL models operate as "black boxes" that are highly susceptible to overfitting these high-energy interference components, resulting in a dramatic degradation of diagnostic performance under complex working conditions (Su, Shi, Zhou, Bai, & Wang, 2024; Soroush, Shirazi, & Raji, 2025).

To address the feature masking caused by strong interference, researchers have attempted to integrate signal processing techniques as a front-end processor to decouple the physical frequency bands (Yan, Yan, Xu, & Yuen, 2023). A prominent example is Variational Mode Decomposition (VMD), originally proposed by Dragomiretskiy and Zosso (Dragomiretskiy & Zosso, 2014). By utilizing the Alternating Direction Method of Multipliers (ADMM) optimization scheme, VMD robustly decomposes a multi-component signal into an ensemble of band-limited Intrinsic Mode Functions (IMFs) with specific physical meanings, which has been widely adopted in recent diagnostics (R. Liu, Ding, Zhang, Zhang, & Shao, 2023; Song, Jiang, Du, Liu, & Zhu, 2023). Concurrently, attention mechanisms borrowed from Computer Vision (CV), such as Squeeze-and-Excitation (SE) networks (Hu, Shen, & Sun, 2018), have been widely adopted to assign dynamic weights to different feature channels (Z. Chen, Hu, Chen, & Zhang, 2024). Nevertheless, a critical limitation persists when transferring these mechanisms from CV to one-dimensional time-series data. Mainstream attention mechanisms heavily rely on Global Average Pooling (GAP) to extract global contextual "energy" features. However, recent studies highlight a fundamental mismatch between GAP operations and the transient nature of impulsive signals (Yan et al., 2023; D. Peng, Wang, Desmet, & Gryllias, 2023). GAP operates by averaging activations across the temporal dimension, which severely dilutes the highly localized, transient spikes characteristic of early fault impacts. Consequently, in mechanical vibration signals, channels containing continuous, pure harmonic interference often yield the highest pooled "energy", whereas channels capturing sparse early fault impacts exhibit low average energy despite their high "impulsiveness" (Li, Atoui, & Li, 2025). This energy-centric attention mechanism fails to purify the features and paradoxically amplifies the harmonic interference, revealing a fundamental incompatibility with physical fault mechanisms.

Furthermore, to enhance model robustness against noise, artificial noise injection (H. Peng, Du, Gao, Wang, & Wang, 2024) and Curriculum Learning (CL) strategies (e.g., adaptive self-paced learning (Wang, Gao, Wang, Yang, & Du, 2024)), originally pioneered by Bengio et al. (Bengio, Louradour, Collobert, & Weston, 2009) to mimic human cognitive processes by transitioning from easy to complex samples, have been introduced during the training phase (Sun, Yan, Jin, Zhao, & Chen, 2024; Wang et al., 2024). Traditional CL approaches typically employ discrete, hard-coded stages (e.g., transitioning from clean data to progressively noisier data). This open-loop paradigm introduces cumbersome manual hyperparameters and, more detrimentally, triggers catastrophic forgetting during stage transitions. When the interference difficulty increases abruptly (F. Liu et al., 2023), the network experiences severe gradient oscillation, losing the pristine physical features learned in earlier stages. More importantly, existing hybrid approaches merely stitch signal processing and training techniques together without addressing their underlying incompatibilities. The physical properties extracted by front-end signal processors are frequently distorted by subsequent energy-biased, "black-box" attention mechanisms, while rigid open-loop training fails to adapt to the dynamic learning state of these physical representations. Therefore, a fundamentally coupled architecture—featuring a physics-aware representation learning process guided by a closed-loop, continuous, and self-adaptive training strategy—is urgently required.

The primary objective of this research is to develop a highly interpretable and lightweight diagnostic framework capable of accurately isolating fault signatures under low-SINR conditions, while simultaneously ensuring stable model training. To overcome the aforementioned challenges, this paper proposes a novel framework named Physics-Informed Multi-Scale Network (PI-MSN) for robust fault diagnosis under severe compound interference. It should be clarified that rather than purely methodological or mathematical innovations in a single domain, the core novelty of PI-MSN lies in the effective architectural integration. It uniquely bridges physical prior knowledge with adaptive curriculum learning. Rather than incrementally stacking existing modules, PI-MSN substantially advances beyond conventional hybrid models by establishing a deeply coupled, closed-loop paradigm where physical interpretability and dynamic training mutually reinforce each other.

The raw signals are first decoupled in the physical frequency domain via VMD. Building upon the cutting-edge paradigm of Physics-Informed Machine Learning, which emphasizes embedding physical domain knowledge into neural architectures (Ni, Ji, Halkon, Feng, & Nandi, 2023) to overcome data dependency and improve interpretability (Z. Chen & Li, 2017; H. Peng, Wang, Gao, Wang, & Du, 2025), we eschew the traditional blind end-to-end learning approach. Instead, a novel Physics-Informed Channel Attention (PICA) mechanism is proposed. By utilizing physical priors, PICA evaluates the impulsiveness and energy of each Intrinsic Mode Function (IMF) simultaneously, applying soft masking to fundamentally correct the energy-centric bias of standard attention modules, effectively suppressing harmonic interference and highlighting fault impulses. Subsequently, a Multi-Scale Feature Extractor (MSFE) is utilized to capture both broad envelope features and transient spikes.

Most importantly, the entire network is trained under a newly

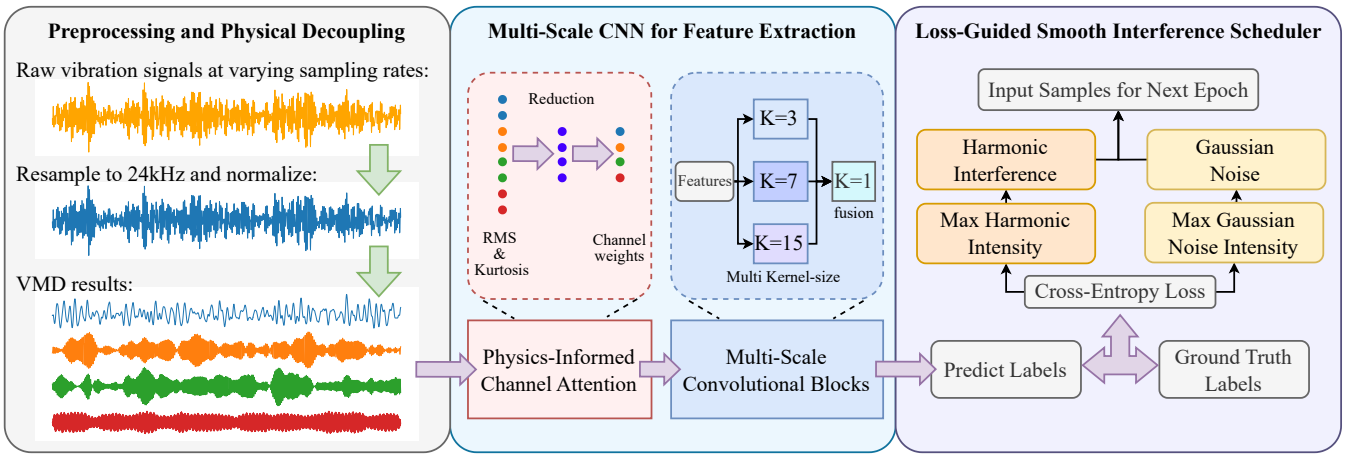

Figure 1. The overall architecture of the proposed Physics-Informed Multi-Scale Network (PI-MSN). The framework consists of three main modules: signal preprocessing and physical decoupling, Physics-Informed channel attention and multi-scale feature extracter, and a loss-guided smooth interference scheduler for closed-loop adaptive training.

designed Loss-Guided Smooth Interference Scheduler (LGSIS). Unlike rigid schedules, LGSIS leverages the high-quality, physics-purified loss signal enabled by PICA to act as a highly accurate real-time feedback indicator. LGSIS establishes a continuous, closed-loop curriculum learning paradigm driven by the dynamic training loss, ensuring a zero-bound uniform sampling mechanism that effectively prevents catastrophic forgetting. By deeply integrating physics-guided representation with self-paced adversarial training, the highly lightweight PI-MSN achieves extreme robustness in low-SINR environments, proving that physical interpretability effectively eliminates the modern reliance on massive parameter stacking.

The main contributions of this paper are summarized as follows:

1) A Physics-Informed Channel Attention (PICA) mechanism is proposed. By jointly evaluating the Kurtosis and Root Mean Square (RMS) of physical IMFs, PICA transcends the limitations of traditional energy-dependent pooling. It autonomously assigns high weights to fault impulse components while effectively suppressing low-frequency harmonic interference, significantly enhancing the physical interpretability of the deep learning model.

2) A Loss-Guided Smooth Interference Scheduler (LGSIS) is designed. As a continuous and closed-loop self-paced learning strategy, LGSIS dynamically calculates the upper bound of the injected compound interference based on the real-time training loss. Its unique uniform sampling mechanism strictly prevents catastrophic forgetting, allowing the model to smoothly adapt to extreme SINR conditions.

3) An end-to-end robust diagnostic framework, PI-MSN, is established. Extensive experiments conducted on the CWRU and HUST datasets validate that the proposed PI-MSN not only achieves superior diagnostic accuracy under severe compound noise and harmonics compared to state-of-the-art methods.

2. METHODOLOGY

2.1. Overall Framework of PI-MSN

To tackle the challenges of feature masking and catastrophic forgetting under severe compound interference (i.e., low SINR conditions), this paper proposes a novel end-to-end diagnostic framework named the Physics-Informed Multi-Scale Network (PI-MSN). The overarching architecture of PI-MSN is carefully designed to integrate signal processing priors with deep representation learning, orchestrated by a dynamic curriculum learning strategy. As illustrated in figure 1, the overall workflow of the proposed framework can be systematically divided into three major components: Physical Decoupling, Physics-Informed Representation Learning, and Closed-Loop Adaptive Training.

Signal Preprocessing and Physical Decoupling. The first stage aims to standardize the input domain and unravel the complex frequency components of the raw vibration signals. Considering that vibration data collected from different industrial scenarios often vary in sampling rates (e.g., 48 kHz or 51.2 kHz), a polyphase filtering-based resampling technique is initially applied to unify all raw signals to a standardized 24 kHz. This specific resampling target preserves the high-frequency resonance bands containing critical fault impulses while avoiding computational redundancy. Subsequently, to break the "black box" nature of pure deep learning, Variational Mode Decomposition (VMD) is employed as a physical front-end processor. The raw 1D vibration signal is adaptively decomposed into $K = 4$ distinct Intrinsic Mode Functions (IMFs), which explicitly separate macroscopic low-frequency harmonics, structural resonances, high-frequency fault impulses, and background Gaussian noise into independent phys-

ical channels.

Physics-Informed Channel Attention and Multi-Scale Extraction. Following the physical decoupling, a Physics-Informed Channel Attention (PICA) module is proposed to overcome the inherent energy-centric bias of conventional attention mechanisms. By jointly evaluating statistical priors, PICA explicitly highlights fault impulses while suppressing macro-harmonics. The re-weighted multi-channel physical tensor is then fed into a Multi-Scale Feature Extractor (MSFE), which employs parallel convolutional layers with varying kernel sizes to capture comprehensive discriminative features before classification.

Loss-Guided Smooth Interference Scheduler. To imbue the PI-MSN with exceptional robustness without suffering from the catastrophic forgetting typical of discrete curriculum learning, a closed-loop training strategy named LGSIS is integrated. LGSIS utilizes the real-time training loss to dynamically regulate the injected compound interference, ensuring a highly stable and self-paced adversarial training process.

2.2. Signal Preprocessing and Physical Decoupling

In practical industrial applications, vibration datasets collected from different testbeds or sensors frequently exhibit varying sampling rates (e.g., 48 kHz or 51.2 kHz). Directly feeding these heterogeneous signals into a unified neural network invariably leads to frequency shift and domain mismatch. To standardize the input domain while minimizing computational redundancy, a preprocessing stage utilizing polyphase filtering is initially applied to resample all raw signals to a unified target frequency of 24 kHz. This specific target is meticulously chosen because the fundamental resonance bands excited by bearing fault impulses predominantly reside within the 2 kHz to 10 kHz range; thus, a 24 kHz sampling rate fully satisfies the Nyquist theorem without losing critical high-frequency diagnostic information. To circumvent severe aliasing and boundary artifacts caused by non-integer decimation (e.g., converting 51.2 kHz to 24 kHz), a polyphase anti-aliasing filter is employed. Let the original signal be $x_{orig}[m]$ with a sampling rate of f_{orig}, and the target rate be f_{tar}. The rational resampling ratio is defined as $P/Q = f_{tar}/f_{orig}$, where P and Q are coprime integers. The discrete resampled signal $x[n]$ is mathematically formulated as:

$$x[n] = \sum_{m=-\infty}^{\infty} x_{orig}[m] \cdot h[nP - mQ] \tag{1}$$

where $h[\cdot]$ denotes the impulse response of the ideal low-pass finite impulse response (FIR) filter. Through this up-sampling by P and down-sampling by Q, the standardized 1D signal seamlessly preserves the physical integrity of the waveform.

Once standardized, the 1D vibration signal remains a complex, highly coupled mixture of macro-mechanical harmonics, environmental noise, and weak fault impulses. Feeding such a highly coupled signal directly into an end-to-end convolutional network essentially treats the physical system as a "black box," making the model highly susceptible to overfitting high-energy interference. To explicitly disentangle these components, Variational Mode Decomposition (VMD) is introduced as a robust physical front-end processor. Unlike traditional recursive methods such as Empirical Mode Decomposition (EMD), VMD is a mathematically well-founded, non-recursive technique that adaptively decomposes a multi-component signal into a predefined number of K discrete Intrinsic Mode Functions (IMFs), denoted as $u_k(t)$, each tightly clustered around a center frequency ω_k.

The core objective of VMD is to minimize the sum of the estimated bandwidths of all IMFs while ensuring that their exact reconstruction equals the original resampled signal $x(t)$. The constrained variational problem is mathematically formulated as:

$$\begin{aligned} &\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left(\left[\left(\delta(t) + \frac{\mathrm{j}}{\pi t} \right) * u_k(t) \right] \mathrm{e}^{-\mathrm{j}\omega_k t} \right) \right\|_2^2 \right\} \\ &\text{s.t.} \quad \sum_{k=1}^{K} u_k(t) = x(t) \end{aligned} \tag{2}$$

where $\delta(t)$ represents the Dirac distribution, $*$ denotes convolution, j is the imaginary unit, and ∂_t indicates the partial derivative. The term inside the square brackets represents the analytic signal of $u_k(t)$ computed via the Hilbert transform, which yields a unilateral frequency spectrum. To render this constrained problem solvable, an augmented Lagrangian $\mathcal{L}$ is constructed by introducing a quadratic penalty factor α to guarantee reconstruction fidelity and a Lagrangian multiplier $\lambda(t)$ to strictly enforce the constraints. The Alternate Direction Method of Multipliers (ADMM) is subsequently deployed to iteratively update u_k, ω_k, and λ in the frequency domain until convergence is achieved.

A critical hyperparameter determining the success of VMD is the number of decomposition modes, K. Over-decomposition leads to severe mode splitting, which dilutes the fault features across multiple channels, whereas under-decomposition results in feature mixing. Grounded in the fundamental physical mechanisms of rotating machinery, K is optimally configured to 4 in this study. This configuration is not arbitrary but physically motivated: IMF1 captures the low-frequency macroscopic mechanical harmonics (e.g., gear meshing and rotor imbalance); IMF2 represents intermediate-frequency structural resonances; IMF3 isolates the high-frequency resonance bands strictly excited by early fault impulses; and IMF4 absorbs ultra-high-frequency background Gaussian noise.

Consequently, the original 1D time-series signal of length L is physically mapped into a multi-channel tensor $\mathbf{X} \in \mathbb{R}^{K \times L}$. This rigorous physical decoupling transforms the initial un-

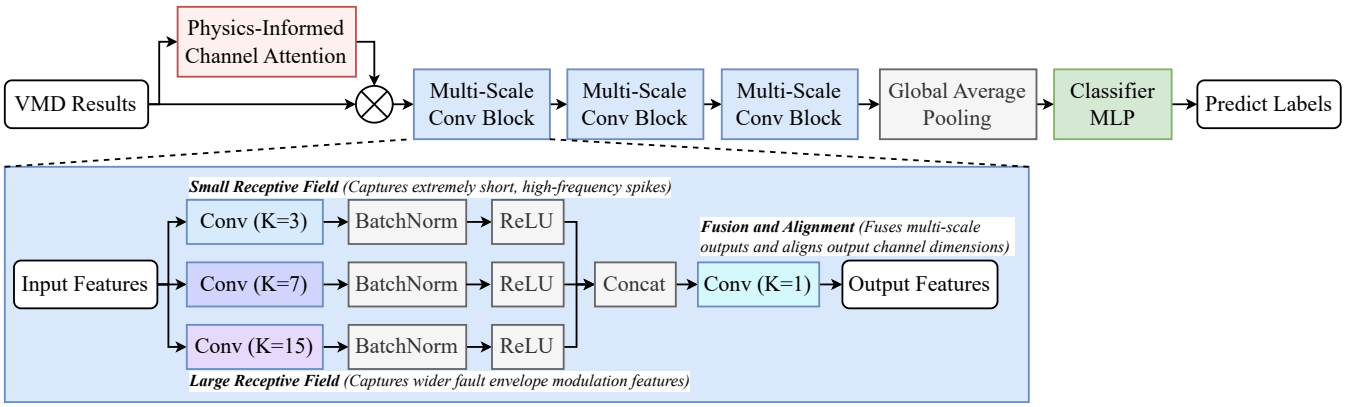

Figure 2. The proposed network architecture featuring Physics-Informed Channel Attention and Multi-Scale Feature Extraction. The expanded view demonstrates the Multi-Scale Conventional Block's use of $K = 3$, 7, and 15 parallel 1D convolutions to capture diverse frequency components before fusion via a 1×1 convolution.

interpretable signal into highly interpretable, band-limited physical channels, laying a solid foundation for the subsequent Physics-Informed representation learning and attention distribution.

2.3. Physics-Informed Channel Attention and Multi-Scale Feature Extraction

Following the physical decoupling phase, the raw 1D vibration signal is transformed into a highly interpretable multi-channel tensor $\mathbf{X} \in \mathbb{R}^{K \times L}$, where K denotes the number of IMFs and L represents the sequence length. As discussed in Section 1, conventional channel attention mechanisms (e.g., SE networks) rely heavily on Global Average Pooling (GAP), an "energy-centric" evaluator that erroneously assigns high weights to harmonic interference channels, exacerbating feature masking.

To overcome this critical physical mismatch, a novel Physics-Informed Channel Attention (PICA) module is proposed to replace the blind end-to-end pooling strategy. Instead of evaluating mere energy, PICA leverages domain-specific statistical priors to jointly assess both the energy and the impulsiveness of each physical channel. Specifically, for the k-th channel of the input tensor $\mathbf{x}^{(k)} = [x_1^{(k)}, x_2^{(k)}, \ldots, x_L^{(k)}]$, the Root Mean Square (RMS) is calculated to quantify its effective energy:

$$RMS^{(k)} = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\left(x_i^{(k)}\right)^2 + \epsilon} \tag{3}$$

Simultaneously, the Kurtosis ($K_u^{(k)}$), a dimensionless statistical metric highly sensitive to transient impacts and widely recognized in bearing fault diagnostics, is computed to quantify the impulsiveness of the k-th IMF:

$$K_u^{(k)} = \frac{\frac{1}{L}\sum_{i=1}^{L}\left(u_i^{(k)} - \mu_k\right)^4}{\left(\frac{1}{L}\sum_{i=1}^{L}\left(u_i^{(k)} - \mu_k\right)^2 + \epsilon\right)^2} \tag{4}$$

where μ_k is the mean of the k-th channel (IMF), and ϵ is a remarkably small constant added to prevent zero division. The physical feature descriptors for all K channels are then concatenated to form a comprehensive physical prior vector $\mathbf{v}_{phy} \in \mathbb{R}^{2K}$. This vector is subsequently fed into a bottleneck Multi-Layer Perceptron (MLP) with a reduction ratio r, followed by a Sigmoid activation function $\sigma(\cdot)$, to generate the final Physics-Informed attention weights $\mathbf{w} \in \mathbb{R}^K$:

$$\mathbf{w} = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1\mathbf{v}_{phy}\right)\right) \tag{5}$$

where $\delta(\cdot)$ denotes the ReLU activation function, $\mathbf{W}_1 \in \mathbb{R}^{\frac{2K}{r} \times 2K}$ and $\mathbf{W}_2 \in \mathbb{R}^{K \times \frac{2K}{r}}$ are the learnable weight matrices of the MLP. Finally, a soft masking operation is performed by element-wise multiplication ($\otimes$) between the derived attention weights and the original IMF tensor, yielding the purified feature tensor $\mathbf{X}' = \mathbf{w} \otimes \mathbf{X}$. This PICA mechanism autonomously forces the network to focus on high-kurtosis fault impulse channels while penalizing high-energy but low-kurtosis harmonic interference channels.

Subsequently, the purified tensor $\mathbf{X}'$ is fed into a deep representation learning backbone named the Multi-Scale Feature Extractor (MSFE). Traditional CNNs utilizing single-sized kernels struggle to capture both the broad, low-frequency modulation envelopes and the sharp, high-frequency transient spikes simultaneously. To achieve comprehensive receptive fields, the MSFE employs a parallel trident architecture within each convolutional block. Each branch utilizes a distinct 1D convolution kernel size $s \in \{3, 7, 15\}$. The parallel 1D convolution

operation for the m-th branch can be expressed as:

$$y_j^{(m)} = f\left(\sum_{c=1}^{C_{in}} x'_c * k_{j,c}^{(m)} + b_j^{(m)}\right) \tag{6}$$

where $y_j^{(m)}$ is the j-th output feature map of the m-th branch, x'_c is the input purified channel, $k_{j,c}^{(m)}$ denotes the learnable convolution kernel with a specific temporal scale, $*$ represents the 1D convolution operator, $b_j^{(m)}$ is the bias term, and $f(\cdot)$ is the non-linear ReLU activation function.

The multi-scale feature maps extracted by the three branches are concatenated along the channel dimension to form a dense, rich representation. This multi-scale extraction process is repeated iteratively to deepen the feature abstraction. Finally, to resolve the constraint of variable input sequence lengths inherent in dynamic industrial environments, an Adaptive Global Average Pooling (Adaptive GAP) layer is deployed at the terminal end of the convolutional backbone. This layer autonomously compresses the arbitrary temporal dimension L' into a fixed unit length, generating a flattened, length-agnostic feature vector that is seamlessly mapped to the final diagnostic probability distribution via fully connected layers.

2.4. Loss-Guided Smooth Interference Scheduler

While the PI-MSN architecture provides a robust foundation for feature extraction, the training strategy is equally critical. Conventional Curriculum Learning (CL) approaches typically rely on discrete, hard-coded stages, which introduces cumbersome manual hyperparameters and, as previously mentioned, frequently triggers catastrophic forgetting due to abrupt difficulty transitions. To eradicate these fundamental flaws, a Loss-Guided Smooth Interference Scheduler (LGSIS) is proposed to establish a continuous, closed-loop, and self-adaptive adversarial training paradigm.

To accurately simulate the compound interference prevalent in real-world rotating machinery, LGSIS dynamically synthesizes and injects both macroscopic mechanical harmonics and environmental noise into the batched signals during the forward propagation. Let $\mathbf{x} \in \mathbb{R}^L$ denote the pure 1D vibration signal sequence. The dynamically contaminated signal $\tilde{\mathbf{x}}$ is mathematically modeled as:

$$\tilde{\mathbf{x}} = \mathbf{x} + \alpha_h \sin(2\pi f_h \mathbf{t} + \phi_h) + \alpha_n \mathbf{z} \tag{7}$$

where $f_h \sim \mathcal{U}(10, 50)$ Hz represents the random fundamental frequency of the low-frequency mechanical harmonic (mimicking rotor imbalance or gear meshing interference), $\phi_h \sim \mathcal{U}(0, 2\pi)$ is the random phase, and $\mathbf{t}$ is the time vector. The vector $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ represents standard Gaussian white noise. The coefficients α_h and α_n control the instantaneous intensity of the harmonic interference and the Gaussian noise, respectively.

Instead of manually defining when and how to increase the interference intensities α_h and α_n, LGSIS treats the deep neural network as a dynamic control system and utilizes the real-time training loss as a negative feedback variable. At the initial epoch ($t = 0$), the model is trained entirely on uncontaminated data to establish a pristine baseline training loss, denoted as $\mathcal{L}_0$. For any subsequent epoch t, the scheduler monitors the average training loss of the previous epoch, $\mathcal{L}_{t-1}$. A dynamic difficulty factor $\gamma^{(t)}$ is continuously calculated by evaluating the model's current learning capacity relative to the initial baseline:

$$\gamma^{(t)} = \max\left(0, 1 - \frac{\mathcal{L}_{t-1}}{\mathcal{L}_0}\right) \tag{8}$$

Consequently, the dynamic upper bounds for the harmonic and noise intensities at epoch t, denoted as $D_h^{(t)}$ and $D_n^{(t)}$, are strictly governed by this adaptive factor:

$$D_h^{(t)} = D_{h,max} \cdot \gamma^{(t)}, \quad D_n^{(t)} = D_{n,max} \cdot \gamma^{(t)} \tag{9}$$

where $D_{h,max}$ and $D_{n,max}$ are the predefined maximum tolerance limits for the respective interferences. This closed-loop mechanism ensures that when the training loss decreases (indicating the model is learning proficiently), the scheduler automatically elevates the interference upper bounds, challenging the network with harsher SINR conditions. Conversely, if the loss spikes, indicating the network is struggling, the bounds autonomously plateau or decrease, allowing the model sufficient epochs to digest the current difficulty without undergoing mathematical collapse.

The most crucial innovation of the LGSIS lies in its zero-bound continuous sampling mechanism. During the batch loading process within epoch t, the actual injected interference intensities α_h and α_n for each specific sample are continuously sampled from uniform distributions whose lower bounds are strictly anchored at zero:

$$\alpha_h \sim \mathcal{U}(0, D_h^{(t)}), \quad \alpha_n \sim \mathcal{U}(0, D_n^{(t)}) \tag{10}$$

By perpetually anchoring the lower bound at zero, a proportion of the training batch always remains near-pristine. This ensures the network continuously rehearses pure physical fault mechanisms while simultaneously defending against the dynamically escalating upper bounds $D_h^{(t)}$ and $D_n^{(t)}$. Mathematically, this continuous sampling smooths the gradient landscape, fundamentally preventing catastrophic forgetting and imparting the PI-MSN with high diagnostic robustness.

3. EXPERIMENTAL STUDY

3.1. Experimental Setup

To comprehensively evaluate the diagnostic performance of the proposed PI-MSN framework , extensive experiments are conducted on two distinct rotating machinery datasets: the Case

Table 1. Quantitative comparison results among different diagnostic models on the CWRU dataset.

Method	FLOPs (G)	Params (M)	Accuracy$^{\uparrow}$	F1-Score$^{\uparrow}$	Precision$^{\uparrow}$	Recall$^{\uparrow}$	FPR$^{\downarrow}$	FNR$^{\downarrow}$
WDCNN (2017)	0.85	0.60	0.9929	0.9927	0.9930	0.9925	0.0020	0.0075
MCNN (2021)	3.50	1.80	0.9580	0.9577	0.9605	0.9550	0.0120	0.0450
QCNN (2023)	**0.15**	0.64	0.9817	0.9815	0.9825	0.9805	0.0055	0.0195
MSAWS (2024)	12.50	4.80	**0.9979**	**0.9977**	**0.9980**	**0.9975**	**0.0008**	**0.0025**
DDDGN (2025)	2.80	1.50	0.9945	0.9942	0.9950	0.9935	0.0015	0.0065
PI-MSN (Ours)	0.20	**0.45**	0.9969	0.9970	0.9971	0.9969	0.0010	0.0031

Note: FLOPs and Params refer to floating-point operations and the number of parameters, respectively. FPR and FNR denote false positive rate and false negative rate. The arrows $^{\uparrow}$ and $^{\downarrow}$ indicate the preferred direction of metrics. Best results are highlighted in bold; second-best results are underlined. For the proposed PI-MSN, the reported 0.20 G FLOPs represent the entire diagnostic pipeline, consisting of the CNN backbone (0.17 G) and the statistically averaged overhead of the iterative VMD preprocessing (approx. 0.03 G).

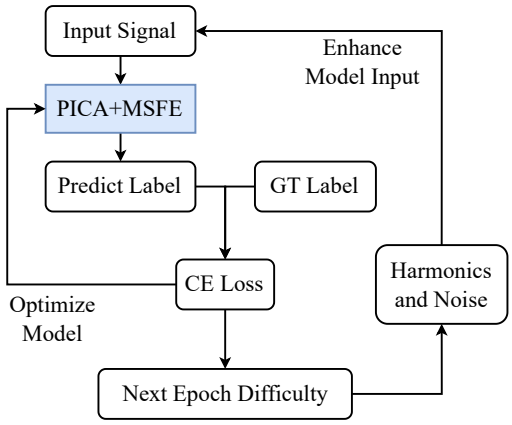

Figure 3. Schematic diagram of the Loss-Guided Smooth Interference Scheduler (LGSIS). The scheduler utilizes the cross-entropy (CE) loss as negative feedback to dynamically adjust the intensity of harmonics and noise for the next epoch, establishing a closed-loop, self-adaptive training paradigm for the PI-MSN.

Western Reserve University (CWRU) bearing dataset (Case Western Reserve University, 2019) and the Hanoi University of Science and Technology (HUST) bearing dataset (Thuan & Hong, 2023). The CWRU dataset, collected at a sampling rate of 48 kHz, serves as the primary benchmark for evaluating fundamental diagnostic accuracy and robustness. To further validate the consistent robustness of the model under varying sampling rates and rotational speeds, the HUST dataset (sampled at 51.2kHz) is utilized as a supplementary testbed. It is important to clarify that the evaluations in this study strictly focus on noise robustness under varying Signal-to-Interference-plus-Noise Ratios (SINR), while the evaluation of cross-domain generalization across distinct operating conditions (e.g., fluctuating speeds and loads) is reserved for future exploration. For both datasets, the mechanical health states are uniformly categorized into four distinct classes: normal condition, inner race fault, rolling element (ball) fault, and outer race fault. To facilitate model optimization and rigorous evaluation, all vibration sequences are strictly partitioned into training, validation, and testing sets according to a 7:2:1 ratio.

The experimental framework is implemented on a workstation running Debian 12, equipped with an Intel Core i7-12700 processor and an NVIDIA GeForce RTX 4090 GPU, utilizing Python 3.12 and the PyTorch 2.9 deep learning library. The network is trained end-to-end for 200 epochs with a batch size of 64. The optimization process utilizes an initial learning rate of 0.001, which is scheduled to decay by a factor of 0.9 every 10 epochs to ensure stable convergence. Regarding the specific hyperparameters of the proposed PI-MSN, the VMD quadratic penalty factor is configured to 2000 to maintain optimal physical decoupling. As suggested by the physical properties of VMD, α strictly controls the bandwidth of the decomposed IMFs. The empirically selected value of 2000 effectively balances the physical trade-off between mode mixing and mode splitting, successfully isolating the structural resonance bands without losing critical diagnostic information. Within the Loss-Guided Smooth Interference Scheduler (LGSIS), the maximum bounds for harmonic interference and Gaussian white noise are empirically set to 3.0 and 1.5, respectively.

The evaluation protocol is carefully designed to assess baseline accuracy and anti-interference robustness. During the baseline training phase, the optimal model weights are selected based on the highest performance achieved on the validation set before being evaluated on the unseen test set. To demonstrate the superiority of the proposed approach, PI-MSN is compared against a spectrum of established baseline models. These include classical deep learning architectures such as the Wide First-layer Deep Convolutional Neural Network (WDCNN) (W. Zhang, Peng, Li, Chen, & Zhang, 2017), Quadratic Convolutional Neural Network (QCNN) (Liao et al., 2023) and the Multi-Scale Convolutional Neural Network (MCNN) (X. Chen, Zhang, & Gao, 2021), alongside recent state-of-the-art diagnostic frameworks including the Dual Decoupling Domain Generalization Network (DDDGN) (G. Zhang et al., 2025) and the Multi-Sensor Adaptive Weighting Strategy (MSAWS) (Jiang et al., 2024).

To guarantee absolute fairness and scientific rigor during the comparative analysis, all baseline models are trained and cross-tested under identical hardware environments and hyperpa-

rameter configurations. For peer-reviewed methods lacking open-source code, the algorithms were meticulously reproduced strictly adhering to their published methodologies. Furthermore, if a comparative model explicitly mandates a specific data preprocessing or augmentation strategy—such as frequency-domain transformation, normalization, or artificial noise injection—it is applied accordingly; otherwise, raw time-domain vibration signals are directly utilized as inputs. For subsequent noise robustness, the models optimized during the baseline phase are directly evaluated on the target datasets without further parameter fine-tuning. Finally, to provide a holistic assessment, the diagnostic performance is quantified using a comprehensive suite of metrics: overall Accuracy (Acc), Precision, Recall, F1-Score, False Positive Rate (FPR), and False Negative Rate (FNR).

3.2. Baseline Results

Table 1 presents a comprehensive quantitative comparison of the proposed PI-MSN against various established baseline models on the CWRU dataset. The experimental results clearly demonstrate the exceptional balance between diagnostic precision and computational efficiency achieved by the proposed framework. While the recent MSAWS model achieves a marginally higher accuracy of 0.9979 , it heavily relies on a massive computational overhead, demanding 12.50 G FLOPs and 4.80 M parameters, which severely limits its deployment on edge devices in real-world industrial settings. In stark contrast, the proposed PI-MSN attains a highly competitive accuracy of 0.9969 and an F1-Score of 0.9970 while requiring only 0.20 G FLOPs and a mere 0.45 M parameters. It is worth noting that this 0.20 G FLOPs accounts for the entire pipeline, comprising 0.17 G for the CNN backbone and approximately 0.03 G (statistically averaged) for the VMD front-end. Although the iterative nature of VMD introduces slight variations in inference time per sample, the overall computational footprint remains highly manageable. This indicates that PI-MSN drastically reduces the floating-point operations by approximately 98.4% compared to MSAWS, without suffering a significant degradation in diagnostic reliability. Regarding actual deployment, the average inference latency of PI-MSN per sample on the test workstation is approximately 2.3 ms, which fully meets the strict real-time requirements for online industrial condition monitoring. Furthermore, when compared to other lightweight or recent state-of-the-art models such as QCNN and DDDGN, which achieve accuracies of 0.9817 and 0.9945 respectively , PI-MSN exhibits superior performance across all evaluation metrics, including lower false positive (0.0010) and false negative rates (0.0031). This overwhelming efficiency and robustness validate the effectiveness of the Physics-Informed attention mechanism and the closed-loop adaptive training strategy, proving that PI-MSN can accurately capture critical fault features without relying on brute-force parameter stacking.

Table 2. Quantitative evaluation results on the HUST dataset

Method	Acc.	F1	Prec.	Recall	FPR	FNR
WDCNN	0.9915	0.9915	0.9920	0.9910	0.0025	0.0090
MCNN	0.9650	0.9652	0.9665	0.9640	0.0105	0.0360
QCNN	0.9785	0.9785	0.9790	0.9780	0.0065	0.0220
MSAWS	0.9985	0.9985	0.9986	0.9984	0.0005	0.0016
DDDGN	0.9972	0.9972	0.9975	0.9970	0.0010	0.0030
PI-MSN	**0.9992**	**0.9993**	**0.9993**	**0.9993**	**0.0003**	**0.0007**

To further validate the wide applicability and robust learning capability of the proposed framework across different testbed configurations and sampling conditions, an independent evaluation was conducted on the HUST dataset. Table 2 details the diagnostic performance of all comparative models when trained and tested entirely within the HUST data distribution. Unlike the CWRU dataset results where the MSAWS model exhibited a marginal accuracy advantage, the proposed PI-MSN framework achieves the absolute best performance across all evaluation metrics on the HUST dataset, reaching a remarkable accuracy of 0.9992 and an F1-score of 0.9993. This superiority is also distinctly evident in the significantly minimized false positive rate (0.0003) and false negative rate (0.0007), strictly outperforming complex state-of-the-art models such as MSAWS and DDDGN. The consistent excellence demonstrated on both the 48 kHz CWRU dataset and the 51.2 kHz HUST dataset proves that the PI-MSN is not exclusively tailored to a single specific data distribution or testbed setup. Instead, the integration of the Physics-Informed channel attention and the multi-scale extraction module inherently empowers the network to effectively decouple and capture robust physical fault features regardless of the underlying sensor configurations. By maintaining state-of-the-art precision across distinct, independent datasets, PI-MSN confirms its high reliability and broad structural suitability for handling varying hardware specifications and severe compound interference in mechanical condition monitoring.

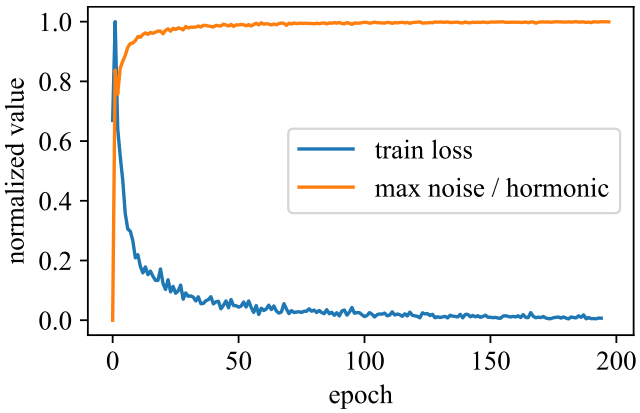

Figure 4. The dynamic evolution of the normalized training loss and the adaptive upper bound of injected compound interference during the PI-MSN training process.

To deeply understand the training stability and the functional efficacy of the proposed Loss-Guided Smooth Interference Scheduler (LGSIS), the dynamic evolution of the normalized training loss and the corresponding interference intensity boundaries over 200 epochs are visualized in Figure 4. As clearly illustrated, the training loss exhibits a rapid initial descent followed by a remarkably smooth and stable convergence trajectory, completely devoid of the severe gradient oscillations or sudden loss spikes typically associated with conventional hard-coded curriculum learning. Correspondingly, the maximum upper bound of the injected compound interference dynamically mirrors the learning progress. In the initial phase, the interference remains minimal, allowing the network to establish a pristine baseline for physical feature extraction. As the model proficiently minimizes the loss, the LGSIS autonomously escalates the interference boundary, continuously challenging the network with progressively harsher conditions. Crucially, the curves demonstrate the closed-loop adaptive nature of the LGSIS; any microscopic transient increase or fluctuation in the training loss immediately triggers a proportional plateau or a slight reduction in the injected noise and harmonic intensity. This continuous, negative-feedback adjustment strictly prevents the network from being overwhelmed by abrupt difficulty transitions, ensuring that the model has sufficient epochs to digest the complex features without undergoing mathematical collapse. Consequently, this self-paced adversarial training paradigm significantly mitigates catastrophic forgetting and guarantees that the PI-MSN achieves highly robust and stable convergence under extreme compound interference scenarios.

To rigorously demonstrate the physical interpretability and statistical superiority of the proposed Physics-Informed Channel Attention (PICA) mechanism over traditional attention modules, Figure 5 presents the statistical distributions of energy (RMS), kurtosis, and assigned attention weights evaluated across 100 testing samples. As visualized in the physical distribution plots (Figure 5a and b), IMF 1 and IMF 2 possess the highest average signal energy (mean RMS of 4.45 and 6.01, respectively), driven predominantly by macroscopic mechanical harmonics and structural resonances. However, the true transient fault impacts are deeply embedded within IMF 3, which exhibits the highest mean kurtosis of 8.53, indicating profound impulsiveness. Meanwhile, IMF 4 is heavily corrupted by high-frequency background Gaussian noise, displaying a near-normal kurtosis (2.90).

Traditional energy-centric attention mechanisms, such as Squeeze-and-Excitation (SE) networks reliant on Global Average Pooling (GAP), are inherently deceived by this physical disparity. As shown in Figure 5c, the SE-Net paradoxically assigns the highest average weights to the high-energy interference channels (0.44 for IMF 2 and 0.35 for IMF 1), while erroneously suppressing the fault-rich IMF 3 with a marginal weight of 0.15, thereby exacerbating feature masking.

By jointly evaluating the physical priors of Kurtosis and RMS, the proposed PICA module fundamentally overcomes this energy-centric limitation. PICA statistically and dynamically allocates the highest attention weight (mean 0.43) to the highly impulsive IMF 3, ensuring that the critical fault signatures are aggressively amplified. Simultaneously, it leverages the joint non-linear mapping to effectively penalize the high-energy, low-kurtosis harmonic interference in IMF 1 (reduced to 0.19) and firmly suppresses the noise-dominated IMF 4 (0.04). This adaptive, physics-informed weighting strategy unequivocally proves that the PI-MSN transcends the traditional "black box" paradigm of deep learning. Furthermore, this statistically significant attention distribution well aligns with the rigorous spectral verification provided in Section 3.4 (Figure 7), confirming that the network autonomously targets the most diagnostic structural resonance bands rather than being overwhelmed by compound interference.

3.3. Robustness analysis under interference environment

To rigorously evaluate the robustness of the proposed framework against complex industrial environments, compound interference comprising varying levels of Gaussian white noise and mechanical harmonics was injected into the original CWRU dataset. The diagnostic accuracy of all comparative models was tested under different Signal-to-Interference-plus-Noise Ratio (SINR) conditions, ranging from 8dB down to a severe -4dB. As detailed in Table 3, traditional and state-of-the-art models exhibit severe performance degradation as the interference intensity increases. For instance, while the MSAWS model achieves an impressive baseline accuracy under clean conditions, its performance catastrophically plummets to 0.8145 at 0dB and further down to 0.6486 at -4dB. Similarly, the accuracies of WDCNN, MCNN, and QCNN all drop below 0.63 under the most extreme -4dB condition, demonstrating their vulnerability to high-energy interference and their inability to extract effective fault features in low-SINR environments.

In stark contrast, the proposed PI-MSN framework demonstrates significant advantage and exceptional stability across all interference levels. Even when the interference energy matches the signal energy at 0dB, PI-MSN maintains a near-perfect diagnostic accuracy of 0.9949. Furthermore, under the extremely harsh -4dB condition, where the weak fault impulses are almost completely submerged in noise, PI-MSN still achieves an accuracy of 0.9816, significantly outperforming the second-best baseline model, DDDGN*, which only reaches 0.9014.

To explicitly disentangle the contributions of the proposed model architecture and the training strategy, controlled experiments were conducted by equipping the best-performing baseline models with the proposed LGSIS training strategy (denoted as WDCNN*, MSAWS*, and DDDGN* in Table

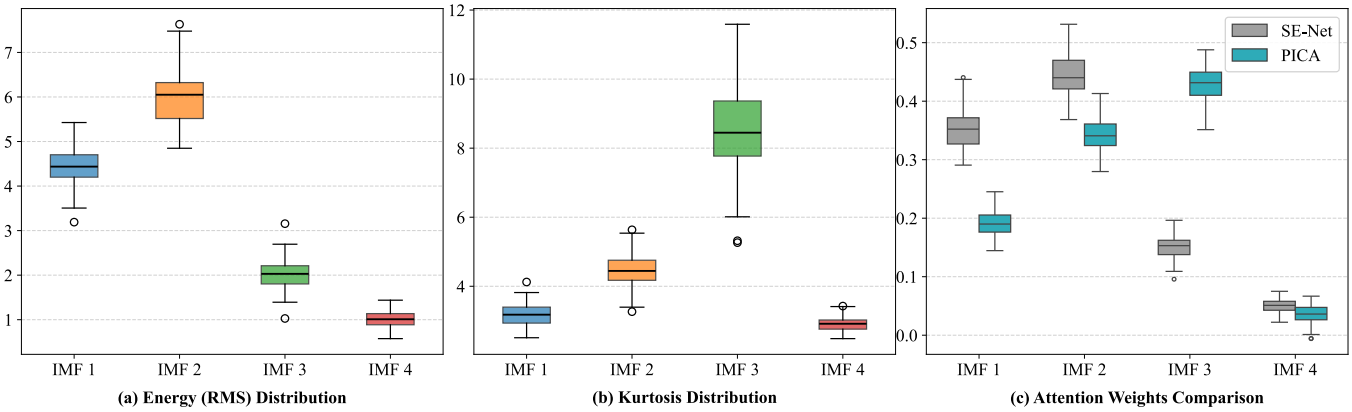

Figure 5. Statistical evaluation of physical priors and attention weight distributions evaluated across 100 testing samples. (a) Energy (RMS) distribution. (b) Kurtosis distribution. (c) Quantitative comparison of attention weights assigned by the conventional SE-Net and the proposed PICA mechanism.

Table 3. Diagnostic Accuracy of Different Models Under Varying compound Interference Levels on the CWRU Dataset

Method	None	8 dB	4 dB	0 dB	-4 dB
WDCNN	0.9929	0.9412	0.8856	0.7634	0.5842
WDCNN*	0.9934	0.9587	0.9213	0.8351	0.7218
MCNN	0.9580	0.9524	0.8965	0.7812	0.6215
QCNN	0.9817	0.9256	0.8524	0.7156	0.5123
MSAWS	0.9979	0.9615	0.9128	0.8145	0.6486
MSAWS*	**0.9982**	0.9784	0.9457	0.8876	0.7935
DDDGN	0.9945	0.9765	0.9528	0.9145	0.8724
DDDGN*	0.9953	0.9812	0.9681	0.9328	0.9014
PI-MSN	0.9969	**0.9969**	**0.9959**	**0.9949**	**0.9816**

Note: The dB values represent the signal-to-interference-plus-noise ratio (SINR) under compound interference. For instance, 0dB indicates that the original signal is mixed with both white noise and harmonic interference, each having energy equal to the signal itself. Methods marked with an asterisk (*) indicate that the baseline model was trained using the proposed LGSIS strategy instead of the standard clean data training.

3). The results demonstrate that the integration of LGSIS significantly improves the diagnostic capabilities of the baseline models under severe noise environments, boosting the accuracy of DDDGN at -4 dB from 0.8724 to 0.9014. This explicitly validates that the LGSIS method can be universally applied to other architectures to enhance noise robustness. Nevertheless, even with the aid of LGSIS, these baseline models still fall short of the performance achieved by PI-MSN. This substantial performance gap successfully isolates and proves the standalone contribution of the Physics-Informed Channel Attention (PICA) and Multi-Scale Feature Extractor (MSFE) structures, confirming that the physical decoupling and attention mechanisms are indispensable for reaching state-of-the-art diagnostic performance under extreme compound interference. This remarkable overall robustness can be primarily attributed to the synergistic effect of PICA and LGSIS. The closed-loop adversarial training paradigm of LGSIS effectively immunizes the network against catastrophic forgetting, allowing the PICA module to consistently filter out high-energy harmonics and focus on the pristine physical fault signatures regardless of the background noise intensity.

To further dissect the individual and coupled impacts of different interference sources, an ablation-like evaluation was conducted by cross-combining varying intensities of mechanical harmonics and Gaussian white noise. As depicted in the performance matrix (Figure 6), PI-MSN exhibits remarkable immunity to mechanical harmonic interference. When the Gaussian noise is held constant, varying the harmonic interference from a clean state down to a severe -4 dB condition causes negligible fluctuations in diagnostic accuracy, consistently remaining above 0.99. This demonstrates that the physical decoupling via VMD, combined with the penalization strategy of the PICA module, effectively mitigates the masking effect of low-frequency macro-mechanical harmonics. Conversely, the model maintains highly stable performance across the noise spectrum from the clean state to 0 dB, but experiences a slight degradation when the Gaussian noise reaches the extreme -4 dB level (yielding accuracies ranging from 0.9816 to 0.9857).

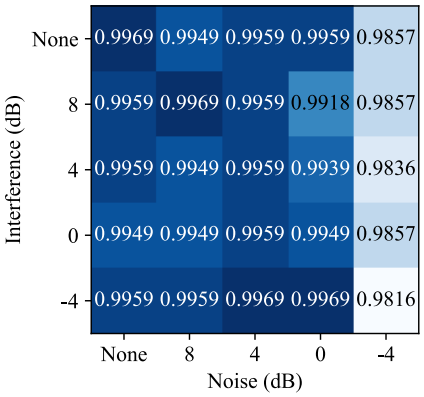

Figure 6. Diagnostic accuracy matrix of the proposed PI-MSN under various cross-combinations of Gaussian white noise and mechanical harmonic interference.

Table 4. Ablation study of the proposed PI-MSN framework under the 0dB compound interference condition (CWRU dataset).

Model Variant	**Key Components**				**Metrics (0dB)**	
	VMD	Attention	MSFE	Training Strategy	Accuracy$^{\uparrow}$	F1-Score$^{\uparrow}$
M1 (Baseline 1D-CNN)	×	×	×	Standard	0.7634	0.7582
M2 (w/o VMD)	×	PICA	✓	LGSIS	0.8542	0.8521
M3 (w/ SE-Net)	✓	SE-Net	✓	LGSIS	0.8915	0.8894
M4 (w/o Attention)	✓	×	✓	LGSIS	0.9123	0.9105
M5 (Single-Scale)	✓	PICA	× (K=3)	LGSIS	0.9712	0.9708
M6 (Static Aug.)	✓	PICA	✓	Static Aug.	0.8856	0.8812
M7 (Discrete CL)	✓	PICA	✓	Discrete CL	0.9345	0.9328
M8 (Linear Sched.)	✓	PICA	✓	Linear Sched.	0.9427	0.9413
M9 (Cosine Sched.)	✓	PICA	✓	Cosine Sched.	0.9582	0.9568
PI-MSN (Ours)	✓	**PICA**	✓	**LGSIS**	**0.9949**	**0.9945**

Note: '✓' indicates the proposed module is utilized. '×' denotes the module is removed or replaced by a standard operation. 'Standard' refers to training on clean data only. 'SE-Net' replaces PICA with conventional Squeeze-and-Excitation pooling. 'K=3' denotes using a single-scale convolutional kernel instead of the parallel trident architecture.

This specific behavioral pattern is not arbitrary but is mathematically bounded by the hyperparameter configuration of the Loss-Guided Smooth Interference Scheduler (LGSIS) during the training phase. Assuming the normalized raw signal power is approximately 1, the empirical upper bounds for harmonic interference ($D_{h,max}$) and Gaussian noise ($D_{n,max}$) were set to 3.0 and 1.5, respectively. Consequently, the maximum injected noise power is $P_n = 1.5^2 = 2.25$, which theoretically corresponds to a minimum training Signal-to-Noise Ratio (SNR) of $10\log_{10}(1/2.25) \approx -3.52$ dB. For the harmonic interference, the maximum injected power of the sinusoidal wave is $P_h = 3.0^2/2 = 4.5$, establishing a minimum training Signal-to-Interference Ratio (SIR) of $10\log_{10}(1/4.5) \approx -6.53$ dB. This theoretical calculation well aligns with the empirical matrix: testing the model under -4 dB harmonic interference falls comfortably within its training cognitive boundary (-6.53 dB < -4 dB), resulting in flawless robustness. In contrast, evaluating the model at -4 dB Gaussian noise slightly exceeds its dynamic training curriculum boundary (-3.52 dB > -4 dB), forcing the network into a zero-shot extrapolation state, which accounts for the minor accuracy dip. Nevertheless, an accuracy of 0.9816 under such unencountered, ultra-low SINR conditions still profoundly eclipses existing state-of-the-art baselines. While the diagnostic efficacy at -4 dB noise could be further optimized by marginally increasing the $D_{n,max}$ hyperparameter, the current configuration is deemed optimal, as complex industrial environments rarely deteriorate beyond a 0 dB noise threshold.

3.4. Hyperparameter Sensitivity and Physical Verification Analysis

To deeply understand the internal mechanism of the PI-MSN framework, a comprehensive physical verification of the decoupling stage and a sensitivity analysis of the core hyperparameters are conducted in this section.

3.4.1. Physical Verification of VMD Decomposition

The number of decomposition modes, K, acts as a critical prior in the physical decoupling stage. To quantify its impact, the diagnostic accuracy under 0 dB compound interference is evaluated for $K \in \{3, 4, 5, 6\}$, as detailed in Table 5.

Table 5. Impact of the number of VMD modes (K) on diagnostic accuracy under 0 dB SINR.

K	3	**4**	5	6
Accuracy	0.9125	**0.9949**	0.9634	0.9412
F1-Score	0.9102	**0.9945**	0.9610	0.9385

The empirical results demonstrate that $K = 4$ achieves the optimal performance. When $K \leq 3$, the network suffers from feature mixing, where high-frequency fault impulses blend with background noise, leading to a notable drop in accuracy. Conversely, configurations with $K \geq 5$ trigger mode splitting, diluting the fault energy across multiple channels and weakening the targeted enhancement of the PICA module.

To explicitly elucidate the physical characteristics of the decoupled channels and provide mechanistic support for the optimal configuration, Figure 7 visualizes the envelope spectra of all four decomposed IMFs for a representative outer race fault sample.

As visually verified in the spectra, the $K = 4$ setting successfully maps distinct physical phenomena into independent channels. Specifically, IMF1 is dominated by macroscopic low-frequency harmonics (e.g., shaft rotation frequency), while IMF4 exhibits a flat spectrum indicative of wideband Gaussian noise. Most crucially, the envelope spectrum of IMF3 displays prominent and distinct spectral peaks exactly at the theoretical Ball Pass Frequency Outer race (BPFO) and its higher-order harmonics. This comprehensive spectral verification unequivocally proves that the proposed VMD configuration effectively

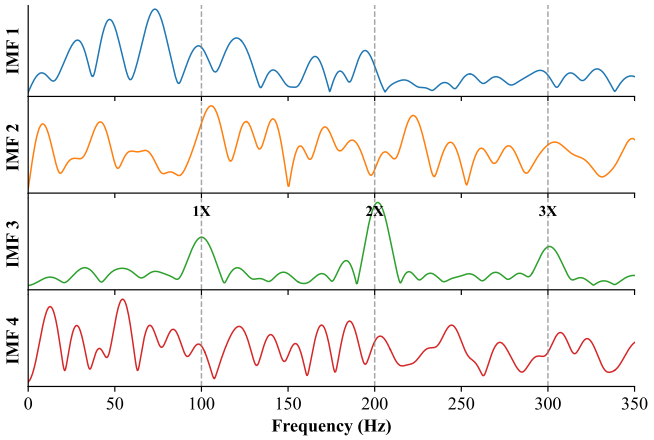

Figure 7. Envelope spectra of the four decomposed Intrinsic Mode Functions (IMFs). IMF1 isolates low-frequency macroscopic harmonics; IMF2 captures intermediate structural resonances; IMF3 explicitly isolates the high-frequency resonance band dominated by the theoretical Ball Pass Frequency Outer race (BPFO) and its harmonics; IMF4 predominantly contains wideband Gaussian noise.

and cleanly isolates the fault-induced resonance band without frequency aliasing.

3.4.2. Sensitivity Analysis of PICA and LGSIS Parameters

To ensure the framework does not rely on overly specific empirical values, the stability of the PI-MSN is evaluated against the reduction ratio r in the PICA module and the dynamic interference boundaries in the LGSIS.

Table 6. Sensitivity analysis of the reduction ratio r in the PICA module.

Reduction Ratio r	2	4	**8**	16
Accuracy (0 dB)	0.9951	0.9948	**0.9949**	0.9932
Params (M)	0.462	0.455	**0.450**	0.447

As illustrated in Table 6, the diagnostic accuracy remains highly stable across a wide range of $r \in \{2, 4, 8\}$, with precision fluctuations strictly bounded within 0.2%. The parameter $r = 8$ is selected to achieve the optimal trade-off between diagnostic precision and computational compactness.

Furthermore, a grid search was conducted to evaluate the robustness of the closed-loop LGSIS strategy under various maximum interference boundary configurations, $(D_{h,max}, D_{n,max})$.

Table 7 reveals that transitioning from a conservative (C1) to an aggressive (C3) boundary setting significantly enhances the model's robustness in extreme SINR environments (-4 dB). However, overly aggressive noise injection (C3) introduces a slight degradation in clean-data accuracy due to the excessive adversarial difficulty during training. The proposed configuration ($D_{h,max} = 3.0, D_{n,max} = 1.5$) exhibits low sensitivity to minor boundary perturbations and secures a superior balance, ensuring near-perfect accuracy under standard conditions while maintaining robust defense mechanisms against severe compound interference.

Table 7. Diagnostic performance under various LGSIS maximum boundary configurations.

Config	$D_{h,max}$	$D_{n,max}$	Acc (Clean)	Acc (-4 dB)
C1 (Conservative)	2.0	1.0	0.9972	0.9245
C2 (Proposed)	**3.0**	**1.5**	**0.9969**	**0.9816**
C3 (Aggressive)	4.0	2.0	0.9854	0.9842

3.5. Ablation studies

To rigorously validate the necessity and individual contribution of each proposed module within the PI-MSN framework, a comprehensive ablation study was conducted. The models were evaluated under the challenging 0dB compound interference condition, where the severity of noise and harmonics equals the true fault signal energy. As presented in Table 4, a baseline 1D-CNN (M1) stripped of all specialized components suffers a catastrophic performance collapse, yielding a mere accuracy of 0.7634. By systematically reintegrating or substituting specific modules, several critical observations can be derived.

Most notably, replacing the proposed PICA with a conventional attention mechanism (M3, SE-Net) paradoxically degrades the accuracy to 0.8915, which is even lower than the variant without any attention module (M4, 0.9123). This counter-intuitive phenomenon exposes a fundamental domain mismatch when directly applying vision-based attention to physical vibration signals. While the VMD front-end successfully isolates high-energy macroscopic harmonics and weak, high-kurtosis fault impulses into distinct channels, traditional attention modules like SE-Net rely heavily on Global Average Pooling (GAP), which inherently acts as an energy-centric evaluator. Consequently, SE-Net is easily deceived by high-energy harmonic interference, erroneously assigning it higher weights while actively suppressing the critical but sparse fault impacts.

In contrast, the proposed PICA is specifically designed to deeply couple with the physical properties of the VMD outputs. By jointly evaluating the physical priors of Kurtosis and RMS, PICA fundamentally rectifies this energy bias. It successfully aligns the attention mechanism with actual physical fault characteristics, precisely highlighting the high-kurtosis fault impulses while robustly suppressing the harmonic interference. This synergistic coupling explicitly confirms that physical decoupling and physics-informed attention are mutually indispensable for achieving robust diagnosis under severe interference.

Furthermore, to comprehensively evaluate the training strategy, LGSIS was substituted with various baselines. Replacing it with standard Static Augmentation (M6) or Discrete Curriculum Learning (M7) results in significant performance drops. To further investigate continuous curriculum paradigms, Linear Scheduling (M8) and Cosine Scheduling (M9) were introduced. To ensure a strictly fair comparison, both continuous baselines dynamically escalate the interference intensities from zero to the exact same maximum bounds ($D_{h,max} = 3.0$, $D_{n,max} = 1.5$) over the total epochs using linear and cosine functions, respectively. However, these open-loop strategies only achieve suboptimal accuracies of 0.9427 and 0.9582. Because they blindly increase the interference difficulty based solely on the predefined time step, they remain entirely oblivious to the network's actual learning state. This rigid open-loop progression frequently forces the model into high-difficulty regimes before it has adequately digested simpler features, triggering gradient oscillation and partial catastrophic forgetting. In contrast, the closed-loop, loss-guided nature of LGSIS dynamically calibrates the difficulty, achieving the optimal accuracy of 0.9949 by ensuring a truly self-paced adversarial training process.

Ultimately, the synergistic integration of all modules in PI-MSN proves that physical decoupling, Physics-Informed representation, and closed-loop continuous training are indispensable for robust industrial fault diagnosis.

4. CONCLUSION AND FUTURE WORK

In this paper, a novel and highly interpretable framework, named the Physics-Informed Multi-Scale Network (PI-MSN), is proposed to address the critical challenge of rotating machinery fault diagnosis under severe compound interference. Breaking away from the traditional "black-box" paradigm of end-to-end deep learning, the PI-MSN synergistically integrates physical signal processing priors with advanced representation learning. Specifically, Variational Mode Decomposition (VMD) is first employed to explicitly decouple the raw vibration signals into distinct physical frequency bands. To overcome the inherent flaw of energy-centric pooling, a Physics-Informed Channel Attention (PICA) mechanism is designed to jointly evaluate the Kurtosis and Root Mean Square (RMS) of each channel, autonomously highlighting weak fault impulses while aggressively suppressing high-energy macroscopic mechanical harmonics. Furthermore, a closed-loop curriculum learning strategy, the Loss-Guided Smooth Interference Scheduler (LGSIS), is proposed to dynamically regulate the injected compound interference based on real-time training loss. By enforcing a zero-bound uniform sampling mechanism, LGSIS significantly mitigates the catastrophic forgetting dilemma, ensuring highly stable adversarial training.

Extensive empirical evaluations conducted on the CWRU and HUST datasets conclusively demonstrate the superiority and exceptional robustness of the proposed framework. Compared to existing state-of-the-art diagnostic models, PI-MSN not only achieves near-perfect baseline accuracy but also maintains an overwhelming advantage in extremely low Signal-to-Interference-plus-Noise Ratio (SINR) environments. Even under a severe -4 dB compound interference condition, where fault signatures are almost entirely submerged, PI-MSN sustains a remarkable accuracy of over 98%, significantly eclipsing complex counterparts. Crucially, this robust diagnostic performance is achieved with a highly lightweight architecture, requiring merely 0.20 G FLOPs and 0.45 M parameters, thereby validating that the integration of physical interpretability eliminates the reliance on brute-force parameter stacking.

While PI-MSN exhibits profound resilience to dynamic noise and harmonic fluctuations, we acknowledge that the current evaluation is primarily bounded to consistent operating regimes. Future research will focus on extending its capabilities toward broader operational scenarios. One highly promising direction is Domain Generalization (DG) under severe working condition discrepancies (e.g., transferring diagnostic knowledge from a laboratory motor operating at a constant speed to a complex wind turbine under variable speeds and loads without target-domain fine-tuning). Investigating how to further extract invariant physical representations across massive domain shifts will be a primary objective. Additionally, considering the lightweight nature of PI-MSN, optimizing the framework for real-time inference and deploying it on resource-constrained industrial edge computing devices (such as FPGA or ARM architectures) represents a vital step toward practical, large-scale industrial prognostic health management. While the current evaluation demonstrates high robustness on benchmark datasets, validating the framework on real-world industrial datasets with cross-domain transfer learning remains an important direction for our future work.

ACKNOWLEDGMENT

This research was funded by Hunan Provincial Department of Education Excellent Youth Project under grant number 24B1167.

REFERENCES

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 41–48). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1553374.1553380

Case Western Reserve University. (2019). *Case western reserve university bearing data center.* https://engineering.case.edu/bearingdatacenter/download-data-file.

Chen, X., Zhang, B., & Gao, D. (2021, April). Bearing fault diagnosis base on multi-scale CNN and LSTM model.

Journal of Intelligent Manufacturing, *32*(4), 971–987. doi: 10.1007/s10845-020-01600-2

Chen, Z., Hu, B., Chen, Z., & Zhang, J. (2024, October). Progress and thinking on self-supervised learning methods in computer vision: A review. *IEEE Sensors Journal*, *24*(19), 29524–29544. doi: 10.1109/JSEN.2024.3443885

Chen, Z., & Li, W. (2017, July). Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network. *IEEE Transactions on Instrumentation and Measurement*, *66*(7), 1693–1702. doi: 10.1109/TIM.2017.2669947

Chopra, P., Kumar, H., & Yadav, S. (2025, March). *PNN: A Novel Progressive Neural Network for Fault Classification in Rotating Machinery under Small Dataset Constraint* (No. arXiv:2503.18263). arXiv. doi: 10.48550/arXiv.2503.18263

Dong, Y., Jiang, H., Yao, R., Mu, M., & Yang, Q. (2024, March). Rolling bearing intelligent fault diagnosis towards variable speed and imbalanced samples using multiscale dynamic supervised contrast learning. *Reliability Engineering & System Safety*, *243*, 109805. doi: 10.1016/j.ress.2023.109805

Dragomiretskiy, K., & Zosso, D. (2014, February). Variational mode decomposition. *IEEE Transactions on Signal Processing*, *62*(3), 531–544. doi: 10.1109/TSP.2013.2288675

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141).

Ji, M., & Zhao, G. (2024). DEViT: Deformable convolution-based vision transformer for bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1–13. doi: 10.1109/TIM.2024.3440383

Jiang, X., Li, X., Wang, Q., Song, Q., Liu, J., & Zhu, Z. (2024, January). Multi-sensor data fusion-enabled semi-supervised optimal temperature-guided pcl framework for machinery fault diagnosis. *Information Fusion*, *101*, 102005. doi: 10.1016/j.inffus.2023.102005

Li, G., Atoui, M. A., & Li, X. (2025, April). *Attention-Based Multi-Scale Temporal Fusion Network for Uncertain-Mode Fault Diagnosis in Multimode Processes* (No. arXiv:2504.05172). arXiv. doi: 10.48550/arXiv.2504.05172

Liao, J.-X., Dong, H.-C., Sun, Z.-Q., Sun, J., Zhang, S., & Fan, F.-L. (2023). Attention-embedded quadratic network (qttention) for effective and interpretable bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–13. doi: 10.1109/TIM.2023.3259031

Liu, F., Zhang, T., Zhang, C., Liu, L., Wang, L., & Liu, B. (2023, April). A Review of the Evaluation System for Curriculum Learning. *Electronics*, *12*(7).

Liu, R., Ding, X., Zhang, Y., Zhang, M., & Shao, Y. (2023, February). Variable-scale evolutionary adaptive mode denoising in the application of gearbox early fault diagnosis. *Mechanical Systems and Signal Processing*, *185*, 109773. doi: 10.1016/j.ymssp.2022.109773

Luo, T., Qiu, M., Wu, Z., Zhao, Z., & Zhang, D. (2025, March). *Bearing fault diagnosis based on multi-scale spectral images and convolutional neural network* (No. arXiv:2503.21566). arXiv. doi: 10.48550/arXiv.2503.21566

Ni, Q., Ji, J., Halkon, B., Feng, K., & Nandi, A. K. (2023, October). Physics-informed residual network (PIResNet) for rolling element bearing fault diagnostics. *Mechanical Systems and Signal Processing*, *200*, 110544. doi: 10.1016/j.ymssp.2023.110544

Niu, G., Liu, E., Wang, X., Ziehl, P., & Zhang, B. (2023, January). Enhanced Discriminate Feature Learning Deep Residual CNN for Multitask Bearing Fault Diagnosis With Information Fusion. *IEEE Transactions on Industrial Informatics*, *19*(1), 762–770. doi: 10.1109/TII.2022.3179011

Pancaldi, F., Dibiase, L., & Cocconcelli, M. (2023, April). Impact of noise model on the performance of algorithms for fault diagnosis in rolling bearings. *Mechanical Systems and Signal Processing*, *188*, 109975. doi: 10.1016/j.ymssp.2022.109975

Peng, D., Wang, H., Desmet, W., & Gryllias, K. (2023, April). RMA-CNN: A residual mixed-domain attention CNN for bearings fault diagnosis and its time-frequency domain interpretability. *Journal of Dynamics, Monitoring and Diagnostics*, 1–18. doi: 10.37965/jdmd.2023.156

Peng, H., Du, J., Gao, J., Wang, Y., & Wang, W. (2024, May). Adversarial training of multi-scale channel attention network for enhanced robustness in bearing fault diagnosis. *Measurement Science and Technology*, *35*(5), 056204. doi: 10.1088/1361-6501/ad2828

Peng, H., Wang, W., Gao, J., Wang, Y., & Du, J. (2025, September). A lightweight triple-stream network with multisensor fusion for enhanced few-shot learning fault diagnosis. *IEEE Transactions on Reliability*, *74*(3), 4062–4075. doi: 10.1109/TR.2025.3540500

Ruan, D., Wang, J., Yan, J., & Gühmann, C. (2023, January). CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis. *Advanced Engineering Informatics*, *55*, 101877. doi: 10.1016/j.aei.2023.101877

Shao, H., Lin, J., Zhang, L., Galar, D., & Kumar, U. (2021, October). A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance. *Information Fusion*, *74*, 65–76. doi: 10.1016/j.inffus.2021.03.008

Song, Q., Jiang, X., Du, G., Liu, J., & Zhu, Z. (2023, April). Smart multichannel mode extraction for enhanced bearing fault diagnosis. *Mechanical Systems and Signal Processing*, *189*, 110107. doi: 10.1016/

j.ymssp.2023.110107

Soroush, K., Shirazi, N., & Raji, M. (2025, July). *Efficient Triple Modular Redundancy for Reliability Enhancement of DNNs Using Explainable AI* (No. arXiv:2507.08829). arXiv. doi: 10.48550/arXiv.2507.08829

Su, Y., Shi, L., Zhou, K., Bai, G., & Wang, Z. (2024, April). Knowledge-informed deep networks for robust fault diagnosis of rolling bearings. *Reliability Engineering & System Safety*, *244*, 109863. doi: 10.1016/j.ress.2023.109863

Sun, W., Yan, R., Jin, R., Zhao, R., & Chen, Z. (2024, December). Curriculum-Based Federated Learning for Machine Fault Diagnosis With Noisy Labels. *IEEE Transactions on Industrial Informatics*, *20*(12), 13820–13830. doi: 10.1109/TII.2024.3435449

Thuan, N. D., & Hong, H. S. (2023, July). HUST bearing: A practical dataset for ball bearing fault diagnosis. *BMC Research Notes*, *16*(1), 138. doi: 10.1186/s13104-023-06400-4

Wang, Y., Gao, J., Wang, W., Yang, X., & Du, J. (2024, April). Curriculum learning-based domain generalization for cross-domain fault diagnosis with category shift. *Mechanical Systems and Signal Processing*, *212*, 111295. doi: 10.1016/j.ymssp.2024.111295

Wei, Q., Tian, X., Cui, L., Zheng, F., & Liu, L. (2023, September). WSAFormer-DFFN: A model for rotating machinery fault diagnosis using 1D window-based multi-head self-attention and deep feature fusion network. *Engineering Applications of Artificial Intelligence*, *124*, 106633. doi: 10.1016/j.engappai.2023.106633

Yan, X., Yan, W.-J., Xu, Y., & Yuen, K.-V. (2023, November). Machinery multi-sensor fault diagnosis based on adaptive multivariate feature mode decomposition and multi-attention fusion residual convolutional neural network. *Mechanical Systems and Signal Processing*, *202*, 110664. doi: 10.1016/j.ymssp.2023.110664

Zhang, G., Kong, X., Ma, H., Wang, Q., Du, J., & Wang, J. (2025, April). Dual disentanglement domain generalization method for rotating Machinery fault diagnosis. *Mechanical Systems and Signal Processing*, *228*, 112460. doi: 10.1016/j.ymssp.2025.112460

Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2017, February). A New Deep Learning Model for Fault Diagnosis with Good Anti-Noise and Domain Adaptation Ability on Raw Vibration Signals. *Sensors*, *17*(2), 425. doi: 10.3390/s17020425

Zhao, C., Zio, E., & Shen, W. (2024, May). Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study. *Reliability Engineering & System Safety*, *245*, 109964. doi: 10.1016/j.ress.2024.109964

Zhao, Y., Zhang, Y., Li, Z., Bu, L., & Han, S. (2023, April). AI-enabled and multimodal data driven smart health monitoring of wind power systems: A case study. *Advanced Engineering Informatics*, *56*, 102018. doi: 10.1016/j.aei.2023.102018