SCIREA Journal of Mechanical Engineering
ISSN: 2995-7729
http://www.scirea.org/journal/Mechanical
August 17, 2026
Volume 6, Issue 1, February 2026
http://dx.doi.org/10.54647/mechanical460143

Continuous FMD-Net: A Physics-Driven Algorithm-Unrolling Network with Spatio-Temporal Dual Sparsity for Adaptive Fault Diagnosis

Meng li [1,*], Jiaqi Xie [1], Hangyuan Gao [1]

[1]School of Mechanical and Vehicle Engineering, Changchun University, Changchun, 130022 PR China

[*]Corresponding author

Abstract

Extracting and demodulating weak fault signatures from rotating machinery under heavy background noise remains a significant challenge. Conventional methods such as Feature Mode Decomposition (FMD) rely on manual parameter tuning and are highly susceptible to mode mixing. To address these limitations, this paper proposes Continuous FMD-Net, a fully differentiable framework for adaptive bearing fault diagnosis. The network unrolls the FMD algorithm into deep neural layers and incorporates three key designs: (i) a trainable time-domain spatial attention mask that adaptively determines the effective window length L of each channel's filter via backpropagation; (ii) a channel gating mechanism that prunes redundant channels to automatically optimize the number of modes K; and (iii) a cross-correlation orthogonality penalty, embedded in a physics-informed joint loss function, that enforces mutual repulsion among parallel channels to effectively suppress mode mixing. Experiments on the Case Western Reserve University (CWRU) bearing dataset demonstrate that the proposed network achieves an accuracy approaching 100% under 0 dB noise, and reaches 90.13% on the −8 dB strong-noise blind test set, significantly outperforming all compared methods.

Convergence curves and UMAP visualizations further confirm that the network learns highly decoupled and discriminative fault features. This work provides a feasible pathway for integrating physical demodulation priors with end-to-end deep learning.

Keywords: Rolling bearing fault diagnosis; Algorithm unrolling network; Feature mode decomposition; Mode mixing; Physics-informed machine learning; Adaptive filtering

1. Introduction

Rotating machinery, including bearings, gears, and shafts, constitutes a critical foundation of modern industrial systems and is widely employed in wind power generation, aero-engines, and high-speed train drivetrains [1][2]. Among various mechanical components, rolling bearings account for 40%–50% of all rotating machinery failures owing to prolonged service under extreme operating conditions [3][4]. Timely and accurate bearing fault diagnosis is a prerequisite for implementing predictive maintenance and preventing catastrophic accidents [5][6]. In industrial settings, however, weak fault transient impulses captured by sensors are often overwhelmed by strong background noise (signal-to-noise ratio [SNR] frequently below −8 dB), multipath attenuation, and electromagnetic interference [7][8]. Consequently, accurately extracting weak fault demodulation features from noise-dominated non-stationary vibration signals remains a widely acknowledged challenge in intelligent maintenance [9].

Adaptive signal decomposition methods have been extensively studied over the past two decades to separate fault components from nonlinear vibration signals [10]. Empirical Mode Decomposition (EMD) extracts intrinsic mode functions (IMFs) through local-extrema-driven adaptive sifting [11], but its lack of explicit orthogonality constraints renders it highly prone to mode mixing [12]. Although subsequent developments such as EEMD [13] and CEEMDAN [14] mitigate this issue to some extent, their noise-assisted averaging strategies incur substantial computational overhead and reconstruction residuals, and they cannot adaptively determine the optimal number of decomposition modes. Variational Mode Decomposition (VMD) transforms a signal into band-limited modes by minimizing the demodulation bandwidth [15][16], yet it depends heavily on manually prescribed mode number K and penalty coefficient; inappropriate selections lead to severe under- or over-decomposition. Feature Mode Decomposition (FMD) employs correlated kurtosis (CK) to guide filter design, enabling targeted matching of periodic bearing fault impulses [17]. Nevertheless, conventional FMD suffers from two structural limitations: (i) the filter length L relies heavily on time-consuming grid search, and (ii) the mode number K must be specified in advance [18]. As a result, the entire pipeline is highly dependent on human prior knowledge and ill-suited for online real-time diagnosis.

Furthermore, the aforementioned classical decomposition algorithms and their downstream feature recognition are embedded in an offline cascaded "decompose-then-classify" architecture, in which the decomposition process is unaware of the requirements of the classification task, leading to gradient discontinuity and non-global optima. In recent years, deep learning approaches—represented by convolutional neural networks (CNNs) [19][20], long short-term memory (LSTM) networks [21], self-attention mechanisms [22], and Transformers [23]—have transformed the research paradigm of intelligent fault diagnosis through their powerful automatic feature learning capabilities. However, these purely data-driven methods are essentially "black-box" pattern recognizers that do not explicitly incorporate bearing fault dynamics (e.g., periodic impulse decay mechanisms) [24]. Under low-SNR conditions, purely data-driven models tend to fit noise artifacts devoid of physical meaning, leading to degraded generalization and substantially reduced diagnostic reliability under unseen complex conditions [25]. To this end, Physics-Informed Machine Learning (PIML) has attracted considerable attention as an emerging paradigm that balances data flexibility with physical interpretability [26][27]. In the field of mechanical maintenance, researchers have attempted to encode physical priors such as spectral kurtosis into inductive biases or auxiliary loss terms to guide networks toward learning physically meaningful features [28],[29]. Nevertheless, existing physics-informed diagnostic methods mostly introduce only a single physical constraint and cannot simultaneously solve for the decomposition structural parameters (e.g., filter length L, mode number K) via backpropagation within a unified, fully differentiable, end-to-end framework [30]. Moreover, owing to the lack of mutual exclusion constraints among parallel channels during training, their feature spaces tend to overlap, making it difficult to fundamentally eliminate mode mixing during multi-channel extraction.

To address these engineering needs and research gaps, this paper proposes Continuous FMD-Net, an adaptive feature-mode unrolling network that integrates dynamic mechanisms. For the first time, the conventional FMD operators are restructured into a forward-fully-differentiable, end-to-end deep neural architecture. Compared with conventional externally-attached heuristic optimization or purely black-box diagnostic schemes, the proposed network breaks the offline cascaded "decompose-then-classify" limitation: the decision error of the downstream classification task (cross-entropy loss) can be directly backpropagated to the front-end physical layer, driving joint dynamic adjustment of filter parameters and thereby mitigating gradient discontinuity and non-global optima.

The main contributions of this work are organized into three synergistic aspects:

(1) Mechanism level — Adaptive window driven by time-domain sparse masking. Under the topological constraint of a maximum candidate channel number *Kmax*, a differentiable one-dimensional spatial attention mask is introduced for each parallel channel, combined with an *L1* regularization in the time domain. The network performs adaptive element-wise multiplication on the long convolution kernel in the forward pass, enabling the backpropagation gradient to implement dynamic boundary contraction at both ends of the filter. This mechanism achieves self-assembly of the receptive field mathematically, allowing each channel to adaptively crop an optimal window length L that matches the pulse width of weak fault transient impulses according to its locked fault frequency band, thereby freeing the model from dependence on fixed tap lengths.

(2) Physical level — Cross-correlation orthogonality penalty to suppress mode mixing. To address the problem that multi-channel parallel networks tend to redundantly fit the same dominant component when unconstrained—causing feature redundancy and overlap—a cross-correlation orthogonality penalty is proposed within the multi-task joint loss. This operator enforces decorrelation and orthogonalization constraints on the parallel channels' outputs in the latent feature space, compelling each activated channel to capture non-redundant physical components across the frequency domain. Combined with channel-domain *L1* soft gating that prunes low-contribution channels (automatically evolving the optimal mode number K), this mechanism suppresses mode mixing at its root and preserves the discriminative power of fault features.

(3) System level — A fully differentiable physics-informed diagnostic framework with high-isolation feature embeddings. The proposed high-order joint loss forms a self-balancing system within the network: a fully differentiable Hilbert-transform-based envelope spectrum (locking 1X–3X fault harmonic energy) serves as a guiding attraction term, the cross-correlation term acts as a decorrelation repulsion constraint, and spatio-temporal dual sparsity pruning refines the structure. Under the collaborative optimization of backpropagation, the network extracts high-isolation, thoroughly denoised low-dimensional manifold embeddings prior to classification. Experiments demonstrate that under strong noise and multi-class composite fault conditions, the learned features exhibit favorable inter-class separation and intra-class compactness, achieving a balance between data flexibility and physical interpretability.

2. Continuous Feature Mode Decomposition Network

2.1 Overall Architecture and Forward Mathematical Flow

To break the offline cascaded limitation between front-end signal denoising (e.g., conventional FMD) and back-end intelligent pattern recognition in conventional mechanical fault diagnosis, this paper proposes Continuous FMD-Net, an adaptive feature-mode unrolling network integrating dynamic mechanisms. The network restructures the conventional FMD operators into a forward-fully-differentiable, end-to-end deep neural architecture, enabling the decision error of the downstream classification task (cross-entropy loss) to be directly backpropagated to the front-end physical layer and driving global joint dynamic adjustment of the filter parameters. The global forward mathematical flow topology of Continuous FMD-Net is illustrated in Fig. 2-1.

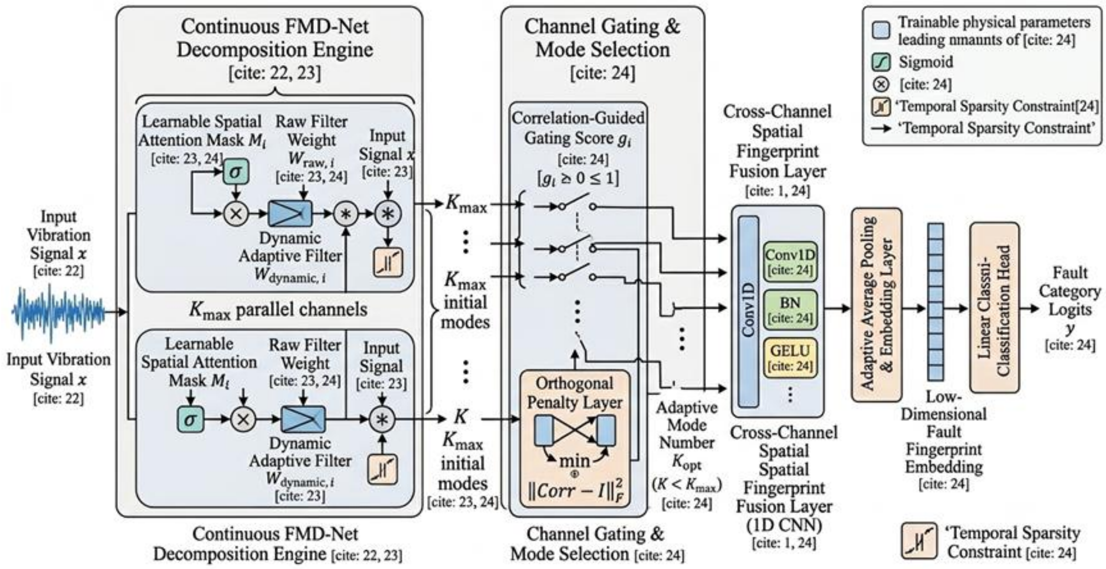

Fig. 2-1 Overall architecture schematic.

Let the input single-channel raw industrial time-series vibration signal be $\mathbf{x} \in \mathbb{R}^{1 \times N}$, where N is the number of time-domain sampling points obtained by sliding-window segmentation. The signal is sequentially passed through a differentiable time-domain spatial mask layer and a channel-domain soft gating layer to adaptively decouple multiple high-purity feature modes (FMs). The concatenated decoupled feature matrix is then fed into a cross-channel spatial fingerprint fusion layer to extract a low-dimensional manifold embedding, which is finally mapped by a linear projection classification head to output the predicted class logits $\hat{\mathbf{y}}$. The entire architecture maintains continuous derivative pathways during forward propagation, laying the mathematical foundation for deep integration of physical mechanisms and deep pattern recognition.

2.2 Differentiable Spatial Mask Layer

Conventional FMD relies heavily on one-dimensional filters with a fixed tap length, whose optimal value L typically requires time-consuming offline grid search or metaheuristic optimization. If L is too small, the fault impulse width may be incompletely captured; if L is too large, extraneous time-domain background noise is introduced.

To overcome this bottleneck, the proposed network breaks free from the dependence on fixed-tap filters. Under the topological constraint of a maximum candidate channel number $K_{\max}$, a set of learnable one-dimensional time-domain spatial attention mask parameters is introduced for each parallel demodulation channel.

Let the raw long filter kernel weight vector of the i-th parallel channel be:

$$\mathbf{W}_{\mathrm{raw},i} \in \mathbb{R}^{1\times L_{\max}}, \quad i = 1, 2, \ldots, K_{max} \tag{1}$$

where $L_{\max}$ is the maximum candidate tap length of the filter kernel. The channel-specific learnable time-domain spatial attention mask vector is defined as $\mathbf{M}_i \in \mathbb{R}^{1\times L_{\max}}$.

To map the bounded mask values to a continuous probability interval, the mask $\mathbf{M}_i$ is first activated element-wise by a Sigmoid function and then combined with the raw long kernel via a Hadamard product (element-wise multiplication), constructing a dynamic filter kernel with an adaptive receptive field boundary:

$$\mathbf{W}_{\mathrm{dynamic},i} = \mathbf{W}_{\mathrm{raw},i} \odot \sigma(\mathbf{M}_i) \quad (i = 1, 2, \ldots, K_{\max}) \tag{2}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function and $\odot$ denotes the Hadamard product operator.

The generated dynamic filter kernel is then convolved with the input raw vibration signal $\mathbf{x}$ via one-dimensional convolution to produce the primary demodulated time-domain feature sequence for each channel:

$$\mathbf{X}_{\mathrm{filtered},i} = \mathrm{Conv1D}(\mathbf{x}; \mathbf{W}_{\mathrm{dynamic},i}) \in \mathbb{R}^{1\times N} \tag{3}$$

2.3 Differentiable Channel Gating and Orthogonal Evolution

In deep networks with parallel multi-channel feature extraction, the absence of explicit mutual constraints causes the parallel channels to converge toward the same dominant noise or fault component during forward propagation, resulting in severe feature redundancy and mode mixing. Simultaneously, adaptively determining the optimal number of activated modes K is a challenge that conventional methods have yet to resolve. To this end, a soft gating pruning mechanism and a cross-correlation orthogonality constraint are jointly designed in the channel domain and the latent feature space.

First, a continuous differentiable feature activation gating vector is introduced:

$$\mathbf{G} = [g_1, g_2, \dots, g_{K_{\max}}]^T \in \mathbb{R}^{K_{\max} \times 1} \tag{4}$$

The gating score $g_i \in [0, 1]$ of the i-th channel is obtained by applying a Sigmoid mapping to the channel-specific learnable parameter θ_i: $g_i = \sigma(\theta_i)$. The final output feature mode (FM) of the i-th channel is:

$$\mathbf{FM}_i = g_i \cdot \mathbf{X}_{\text{filtered},i} \tag{5}$$

To quantitatively eliminate multi-channel feature redundancy at the mathematical level, an orthogonalization constraint is imposed on the feature space filtered by different channels. First, the time-domain vector $\mathbf{FM}_i$ is subjected to zero-mean and variance normalization to obtain the standardized mode vector $\widetilde{\mathbf{FM}}_i \in \mathbb{R}^{1 \times N}$:

$$\widetilde{\mathbf{FM}}_i = \frac{\mathbf{FM}_i - \mu_i}{\sigma_i + \epsilon} \tag{6}$$

where μ_i and σ_i are the mean and standard deviation of $\mathbf{FM}_i$, respectively, and ϵ is a small positive constant to prevent division by zero (set to 10^{-8}).

On this basis, the Pearson cross-correlation coefficient matrix $\mathbf{R} \in \mathbb{R}^{K_{\max} \times K_{\max}}$ among the parallel channels is defined as:

$$\mathrm{R} = \overline{\mathrm{FM}}\,\overline{\mathrm{FM}}^{\mathrm{T}} (\text{where } \overline{\mathrm{FM}} \in \mathbb{R}^{K_{\max} \times N}) \tag{8}$$

The off-diagonal elements of R quantitatively describe the information redundancy between different channels. Ideally, the channels should be mutually orthogonal, and R should approach the identity matrix I

2.4 Cross-Channel Spatial Fingerprint Fusion Layer

After the front-end Continuous FMD-Net engine completes adaptive decoupling and denoising, the output high-purity physical feature mode matrix is concatenated as:

$$\mathbf{FM} = [\mathbf{FM}_1^T, \mathbf{FM}_2^T, \dots, \mathbf{FM}_{K_{\max}}^T]^T \in \mathbb{R}^{K_{\max} \times N} \tag{9}$$

To mine cross-channel fingerprint interactions and perform global pattern recognition on these K high-purity features, a lightweight and robust Cross-Channel Spatial Fingerprint Fusion Layer is constructed at the end of the model. This layer fully exploits the natural local inductive bias and translation invariance of one-dimensional convolution operators. Its forward mathematical formulation is expressed as:

$$\mathbf{H} = \mathrm{GELU}\left(\mathrm{BN}\left(\mathrm{Conv1D}(\mathbf{FM}; \mathbf{W}_c, \mathbf{b}_c)\right)\right) \tag{10}$$

$$\mathbf{h}_f = \mathrm{Flatten}\left(\mathrm{AdaptiveAvgPool1D}(\mathbf{H})\right) \tag{11}$$

$$\hat{\mathbf{y}} = \mathbf{W}_{\mathrm{cls}} \mathbf{h}_f + \mathbf{b}_{\mathrm{cls}} \tag{12}$$

where $\mathrm{Conv1D}(\cdot)$ denotes one-dimensional convolution (with a temporal kernel size of 3); $\mathbf{W}_c$ and $\mathbf{b}_c$ are the learnable weight tensor and bias vector of the fusion-layer convolution kernel, respectively; $\mathrm{BN}(\cdot)$ is a one-dimensional batch normalization operator for stabilizing the internal activation distribution; $\mathrm{GELU}(\cdot)$ is the Gaussian Error Linear Unit activation function; and $\mathrm{AdaptiveAvgPool1D}(\cdot)$ is a global adaptive average pooling layer that condenses and flattens the feature manifold while preserving the spatial fault topology.

Here, $\mathbf{h}_f \in \mathbb{R}^D$ is the low-dimensional fault fingerprint embedding extracted prior to classification, which is finally mapped by the linear projection classification head weights $\mathbf{W}_{\mathrm{cls}}$ and bias $\mathbf{b}_{\mathrm{cls}}$ to output the predicted class logits $\hat{\mathbf{y}}$.

2.5 Physics-Informed Multi-Task Joint Loss Function

To guide the network to autonomously evolve optimal physical features and diagnostic decisions under a forward-fully-differentiable premise, a high-order multi-task joint loss function $\mathrm{L}_{\mathrm{total}}$ is constructed. This loss forms a self-balancing system driven by a guiding attraction term, a decorrelation constraint, and dual sparsity stabilizers:

$$\mathrm{L}_{\mathrm{total}} = \mathrm{L}_{\mathrm{cls}} + \lambda_1 \mathrm{L}_{\mathrm{envelope}} + \lambda_2 \mathrm{L}_{\mathrm{ortho}} + \lambda_3 \mathrm{L}_{\mathrm{spatial}} + \lambda_4 \mathrm{L}_{\mathrm{channel}} \tag{13}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameter weight coefficients that modulate the contribution of each physical constraint and task tendency. The mathematical definitions and physical roles of each loss term are as follows.

(1) Downstream task driver $\mathrm{L}_{\mathrm{cls}}$: The classical cross-entropy loss for classification, responsible for converting the downstream decision error into a global gradient flow:

$$\mathrm{L}_{\mathrm{cls}} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} y_{b,c}\log(\hat{y}_{b,c}) \tag{14}$$

where B is the batch size, C is the total number of fault classes, $y_{b,c}$ is the one-hot ground-truth label, and $\hat{y}_{b,c}$ is the Softmax-normalized predicted probability.

(2) Differentiable demodulation guide $\mathrm{L}_{\mathrm{envelope}}$: The physical envelope-spectrum harmonic energy ratio loss. This term employs a fully differentiable Hilbert transform to extract the analytic envelope signal of each feature mode $\mathbf{FM}_i$ and transforms it into the envelope frequency domain. By explicitly computing and maximizing the energy ratio at the bearing fault characteristic frequencies (including the rotational fundamental and its 1X–3X harmonics), each dynamic filter kernel is guided to converge toward physically meaningful frequency bands. The negative energy ratio is used as the loss contribution, so that minimizing the total loss is equivalent to maximizing the harmonic energy:

$$L_{envelope} = -\frac{1}{K_{max}}\sum_{i=1}^{Kmax}\frac{\sum_{f\in\{1X\sim3X\}}\left|\sum E_i(f)\right|^2}{\sum_{f}\left|E_i(f)\right|} \tag{15}$$

where $E_i(f)$ is the envelope spectrum amplitude of the i-th mode at frequency f.

(3) Topological decorrelation constraint $\mathrm{L}_{\mathrm{ortho}}$: The cross-correlation orthogonality penalty. To suppress feature overlap among channels, this term enforces the feature-space cross-correlation matrix $\mathbf{R}$ to approach the identity matrix $\mathbf{I}$ via the Frobenius norm:

$$\mathrm{L}_{\mathrm{ortho}} = \frac{1}{K_{\max}^2}\|\,\mathbf{R}-\mathbf{I}\|_F^2 \tag{16}$$

where $\mathbf{I}\in\mathbb{R}^{K_{\max}\times K_{\max}}$ is the identity matrix and $\|\cdot\|_F$ denotes the Frobenius norm. This term penalizes the off-diagonal correlation coefficients, compelling the channels to separate in the feature space and thereby suppressing mode mixing during multi-channel feature extraction.

(4) Spatio-temporal dual sparsity stabilizers $\mathcal{L}_{\text{spatial}}$ and $\mathcal{L}_{\text{channel}}$: To implement fine-grained pruning, *L1* norms are applied to sparsify the time-domain spatial attention mask boundaries and the channel-domain activation weights, respectively:

$$\mathcal{L}_{\text{spatial}} = \frac{1}{K_{\max}} \sum_{i=1}^{K_{\max}} \| \sigma(\mathbf{M}_i) \|_1 \tag{17}$$

$$\mathcal{L}_{\text{channel}} = \frac{1}{K_{\max}} \sum_{i=1}^{K_{\max}} |g_i| \tag{18}$$

These two sparsity penalties act as stabilizers that, in concert with the backpropagation gradient flow, drive the network to collaboratively evolve the optimal filter window length L and the optimal number of activated modes K.

3. Experimental Results and Analysis

3.1 Experimental Setup

To comprehensively evaluate the denoising capability and classification performance of the proposed Continuous FMD-Net, experiments are conducted on the publicly available Case Western Reserve University (CWRU) bearing benchmark dataset. The raw time-domain vibration signals are taken from the drive-end bearing (6205-2RS JEM SKF) at a rotational speed of 1797 rpm (0 hp load) with a sampling frequency of 12 kHz. The samples cover four typical health conditions: Normal, Inner Race Fault (IRF), Outer Race Fault (ORF), and Ball Fault (BF). Each defective condition includes three damage diameters (0.007 in., 0.014 in., and 0.021 in.), yielding a total of 10 classification categories.

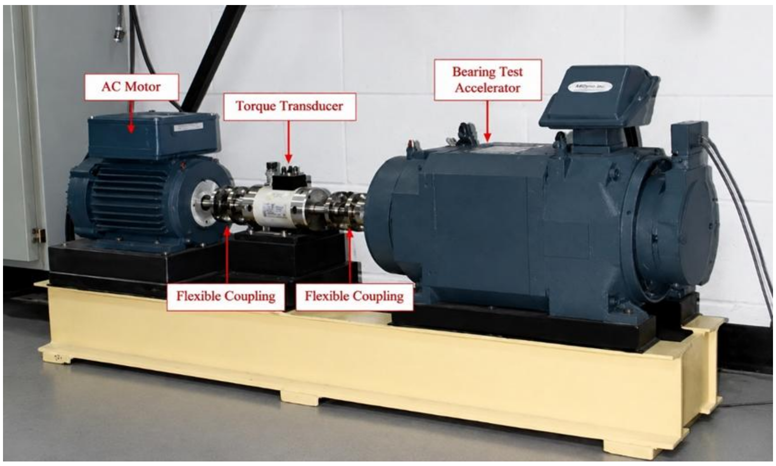

Fig. 3-1 CWRU test rig.

The raw signal is uniformly truncated to 119,808 sampling points, and a sliding-window segmentation strategy is employed to generate multiple samples. The window length is set to 1024 sampling points with a 50% overlap, yielding 233 samples per raw signal. The dataset is then partitioned into training, validation, and test sets in a 7:2:1 ratio.

To simulate the strong background noise encountered in modern industrial settings, Gaussian white noise is artificially added to the raw noise-free vibration signals. Two noise levels are employed in this study:

• 0 dB: Used for the self-evaluation of the proposed model in Section 3 (§3.2–3.5), examining the network's feature learning and diagnostic capability under moderate noise.

• −8 dB: Used for the fair comparison of the four models in Section 4, examining the robustness of each method under extreme strong noise.

The SNR is defined as:

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) \tag{19}$$

where P_{signal} and P_{noise} denote the average power of the raw noise-free vibration signal and the injected noise, respectively.

The proposed model is an end-to-end network consisting of four components: the continuous feature mode decomposition module, a CNN-based feature extractor, an embedding layer, and a fault classifier. The decomposition module contains 8 adaptive channels, and the maximum filter length is set to 257. The model is trained using the AdamW optimizer with a cosine annealing learning rate schedule to improve convergence stability. The main hyperparameters are listed in Table 3-1.

Table 3-1 Main hyperparameters.

Parameter	Value
Sampling frequency	12000 Hz
Fault characteristic frequency	141 Hz
Maximum number of decomposition channels K_max	8
Maximum filter length L_max	257
Batch size	64

Number of epochs	80
Initial learning rate	5×10^{-4}
Weight decay coefficient	1×10^{-4}
Envelope energy loss weight λ_1	1.0
Orthogonality penalty loss weight λ_2	1.0
Time-domain sparsity loss weight λ_3	0.01
Channel sparsity loss weight λ_4	0.05

3.2 Convergence Trajectory Analysis

The end-to-end training trajectory and multi-metric convergence curves of Continuous FMD-Net under 0 dB noise are shown in Fig. 3-2. The total number of training epochs is set to 80.

From the evolution trajectory in Fig. 3-2, it can be observed that, benefiting from the fully differentiable continuous gradient pathway constructed by algorithm unrolling, the composite optimization system exhibits high efficiency and stability. The total loss decreases rapidly from an initial value of 2.25 and enters a stable asymptotic convergence region at approximately the 50th epoch, eventually stabilizing at around 0.40. Both the physics penalty term (FMD Physics Loss) and the harmonic envelope energy term (Harmonic Loss) exhibit coordinated monotonic convergence: the Harmonic Loss continues to decrease with training (from −0.03 to below −0.085), indicating that the dynamic filters in the front-end differentiable physical engine are progressively locking onto the fault characteristic frequency and its harmonics, consistent with the design intent of the physical inductive bias.

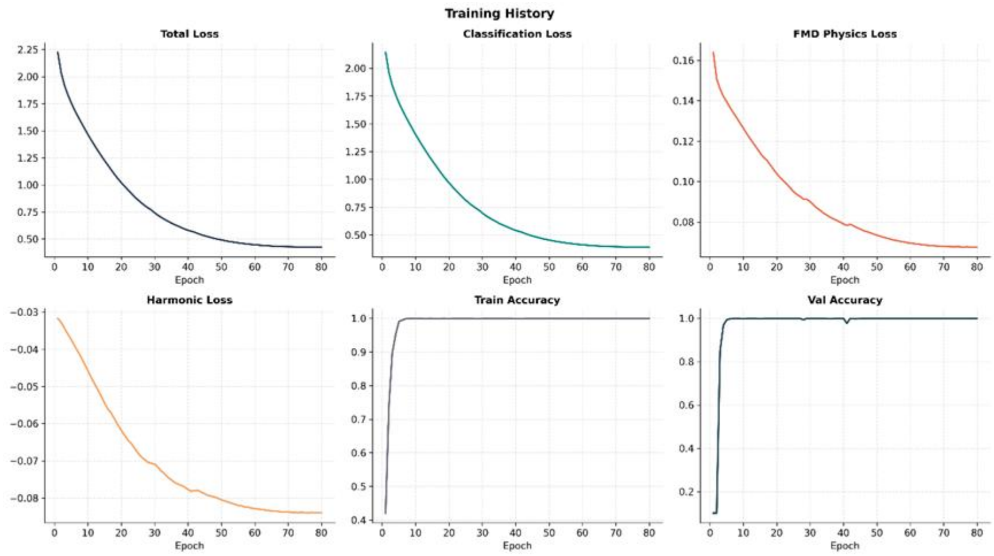

Fig. 3-2 Convergence curves.

Ultimately, the training accuracy (Train Accuracy) and validation accuracy (Val Accuracy) demonstrate good consistency, rising above 99% at approximately the 10th epoch and maintaining a level approaching 100% in subsequent iterations, indicating that the proposed method does not suffer from gradient collapse or convergence to poor local minima.

Although the model achieves excellent classification performance, classification accuracy alone is insufficient for a comprehensive evaluation. Therefore, this paper further conducts multiple interpretability analyses to investigate the network's internal mechanisms.

3.3 Interpretability Analysis of Differentiable Spatial Mask Evolution

To reveal the internal mechanism by which Continuous FMD-Net frees itself from dependence on a fixed tap length L and autonomously crops the optimal adaptive filter window, a visualization analysis is performed on the weight distribution of the learned time-domain spatial attention masks $\mathbf{M}_i$ after convergence and on the corresponding effective filter lengths of each channel, as shown in Fig. 3-3.

During network initialization, the maximum candidate tap length for each channel is set to $L_{\max} = 257$ (indicated by the red dashed line). To investigate the model's adaptive decomposition capability, the spatial attention masks learned by the network are visualized. Fig. 3-3 shows that each decomposition channel autonomously generates a differentiated spatial weight distribution rather than uniformly utilizing all filter coefficients, indicating that the model can selectively emphasize the time-domain regions containing fault impulse features while suppressing irrelevant redundant information. The effective filter lengths of all channels are shown below Fig. 3-3: despite the preset maximum filter length of 257, the effective lengths learned by the network are substantially shorter. The effective filter lengths of the eight channels are:

$$[143,115,106,130,119,126,118,123]$$

The average effective filter length is $L_{\text{avg}} = 122.5$, only 47.7% of the preset maximum length.

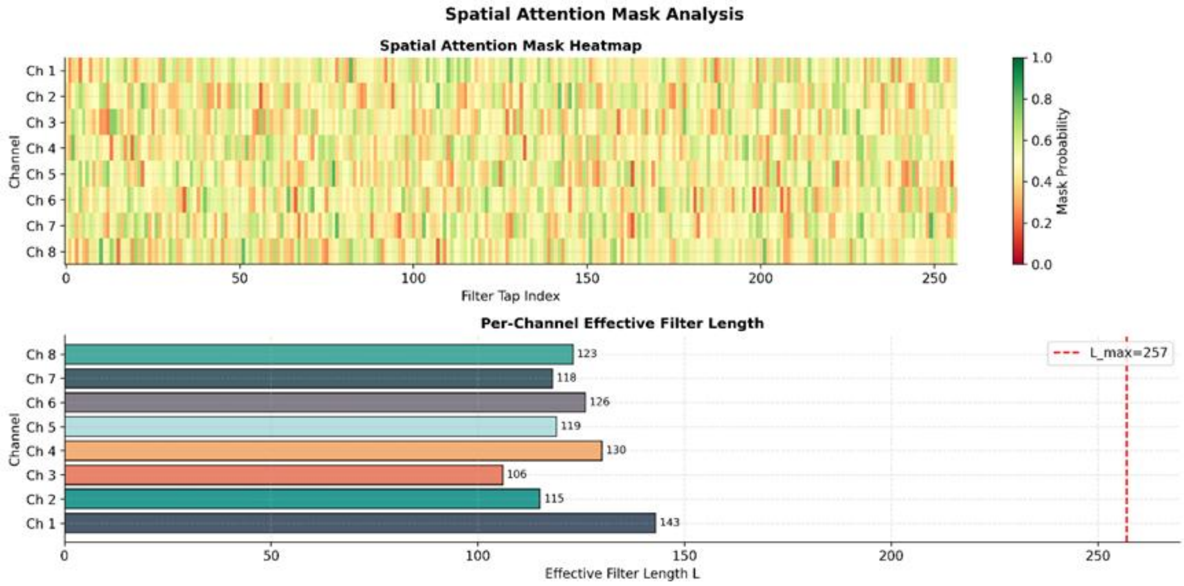

Fig. 3-3 Visualization of the differentiable spatial masks.

These results demonstrate that the proposed adaptive mask mechanism can effectively compress redundant filter coefficients and improve parameter efficiency. Moreover, the differentiated receptive field sizes of different channels indicate that the model can perform multi-scale decomposition autonomously: some channels capture short-duration impulse features, while others extract long-period fault components. Conventional FMD requires manual specification of the filter length, whereas the proposed framework automatically matches the optimal decomposition scale during training. In summary, the spatial attention mechanism not only enhances the flexibility of signal decomposition but also improves the interpretability of the model.

3.4 Channel Orthogonality Analysis

In multi-channel parallel signal processing frameworks, whether mode mixing—caused by all channels blindly fitting the same highest-energy noise or a single dominant component—can be effectively suppressed is a key criterion for evaluating the algorithm's effectiveness. To this end, the Pearson cross-correlation matrix $\mathbf{R}$ is computed on the parallel multi-channel feature modes (FMs) output in the latent feature space, and the quantitative orthogonality analysis results are shown in Fig. 3-4.

The heatmap on the left of Fig. 3-4 exhibits a diagonal-dominant pattern. Apart from the main diagonal elements (the correlation of each channel with itself, which is identically 1.00), all off-diagonal elements—which represent the information redundancy between heterogeneous channels—are close to zero (e.g., 0.01 between Ch1 and Ch3, and 0.00 between Ch2 and Ch8).

From the off-diagonal correlation distribution histogram on the right of Fig. 3-4, it can be seen that the global statistical mean of the absolute Pearson correlation coefficients between heterogeneous channels is only:

$$\text{Mean} = 0.042$$

This near-zero mean indicates that the cross-correlation orthogonality penalty (L_{ortho}), incorporated into the physics-informed joint loss, successfully imposes a decorrelation constraint on the heterogeneous channel features in the latent space, suppressing inter-channel information overlap. It compels each gated physical channel to capture non-redundant demodulated components, thereby effectively mitigating the mode mixing problem that plagues conventional adaptive signal decomposition frameworks such as EMD and VMD.

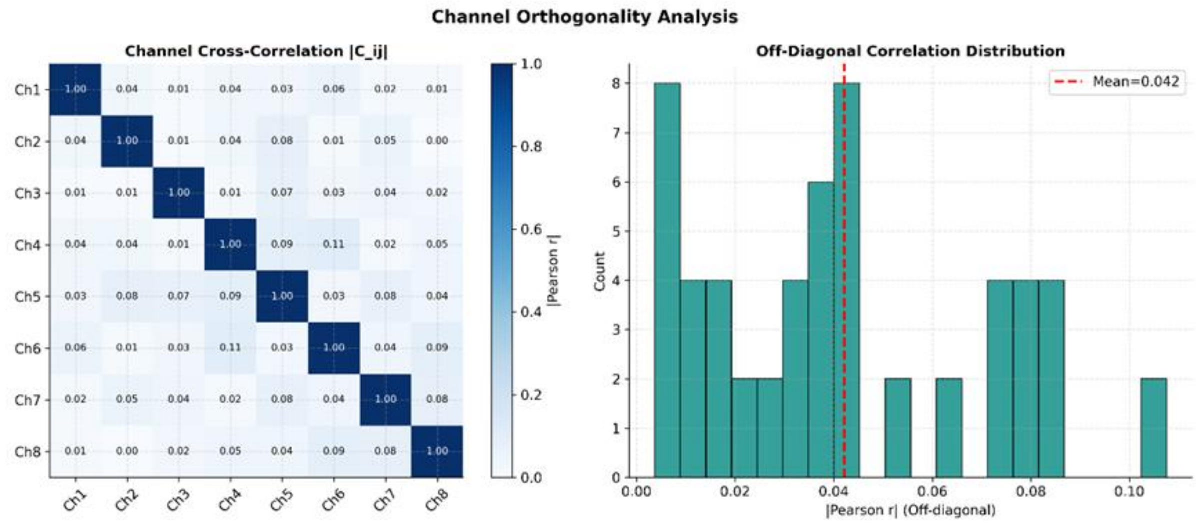

Fig. 3-4 Visualization of channel orthogonality.

3.5 Feature Distribution Visualization

To further examine the pattern recognition accuracy of Continuous FMD-Net, the classification confusion matrix on the 0 dB noise blind test set and the UMAP dimensionality-reduced visualization of the feature embedding space are exported, as shown in Fig. 3-5.

Under the experimental conditions of this section, the model produces no misclassifications across all test samples, achieving a classification accuracy of 100.00%. The UMAP visualization reveals that samples of the same class cluster tightly around their respective cluster centers, while the boundaries between different fault classes are clear with no significant overlap. The intra-class compactness and inter-class separation indicate that the proposed embedding layer can map high-dimensional raw features into latent representations with strong discriminative power. The large inter-cluster distances suggest that the extracted features contain sufficient fault information to accurately distinguish among different fault types and damage severities. The visualization results further validate the effectiveness of the physics-guided decomposition mechanism. Compared with conventional step-wise feature extraction algorithms, the proposed framework can simultaneously accomplish signal decomposition, feature learning, and fault classification within a unified network structure.

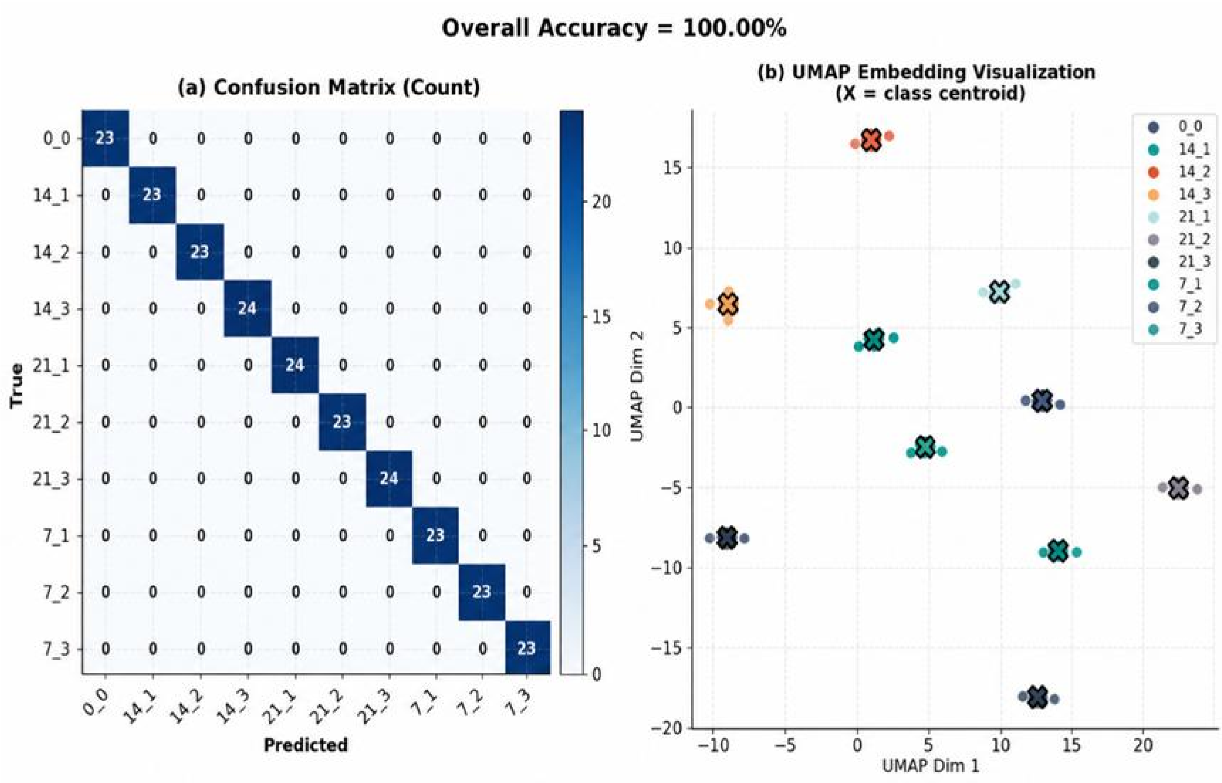

Fig. 3-5 Result visualization (confusion matrix and UMAP).

4. Comparative Experiments

To comprehensively highlight the technical advancement of Continuous FMD-Net, three representative industrial diagnostic benchmark models are introduced for quantitative and visual comparison. The compared methods include:

• **Pure 1D-CNN:** A purely data-driven black-box network;

•**FMD+1D-CNN:** A conventional fixed-parameter feature mode decomposition cascaded network (with manually fixed L = 256, K = 10);

• **GWO-FMD+1D-CNN:** An externally-attached cascaded network based on the Grey Wolf Optimizer (GWO), which searches for L and K by maximizing correlated kurtosis (CK) and then feeds the filtered signal into the CNN after the parameters are locked.

All models are tested fairly under the same −8 dB SNR extreme strong noise and the 7:2:1 dataset partition.

4.1 Characteristics of the Raw Strongly-Noise-Corrupted Signal

Before conducting the multi-model comparison, it is instructive to first analyze the intrinsic characteristics of the raw input data under strong noise. Fig. 4-1 presents a raw −8 dB time-domain input signal sample and its spatial distribution after direct UMAP manifold projection. From the time-domain waveform on the left of Fig. 4-1, it can be seen that, owing to the dominance of Gaussian white noise, the weak periodic bearing fault impulses are completely buried in the background noise.

The right panel of Fig. 4-1 further demonstrates that, without physical denoising and demodulation, the raw feature space exhibits a highly mixed and inseparable state: the projection points of the 10 fault classes are thoroughly intertwined, and the class centers (marked with ✖) are severely collapsed. This indicates that, in the absence of physical denoising, the strong background noise undermines the discriminability of conventional statistical features, confirming the engineering necessity of front-end physical engine decoupling.

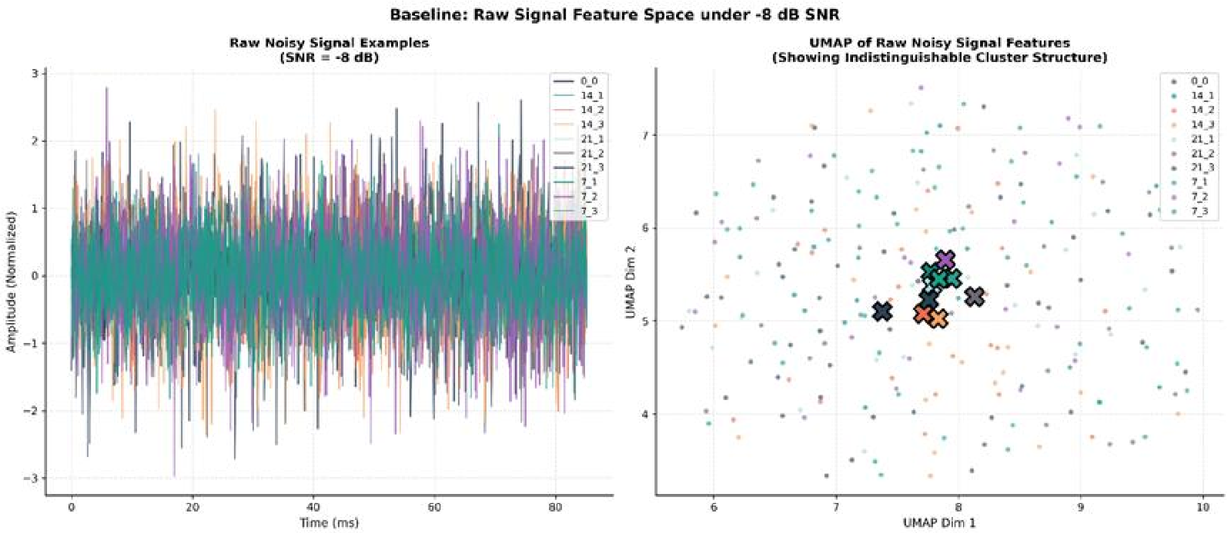

Figure 4-1 Characteristic Distribution of Signals under Strong Noise

4.2 Quantitative Comparison of Classification Performance

Combined with the visual bar chart in Fig. 4-2, the following quantitative conclusions can be drawn regarding the performance of each model:

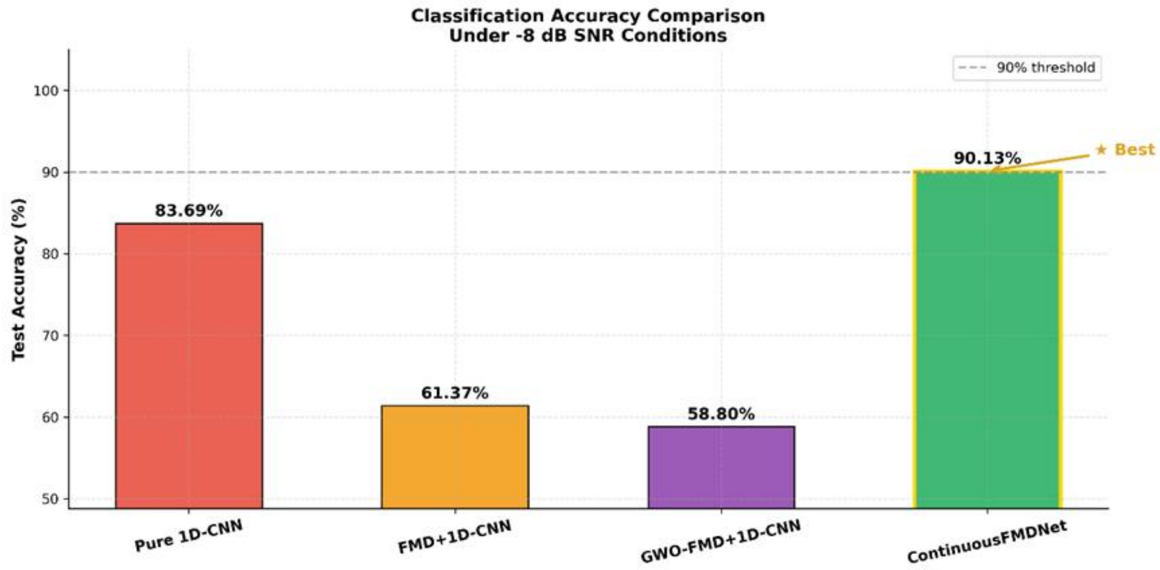

Fig. 4-2 Accuracy curves.

Generalization bottleneck of the purely black-box network: Pure 1D-CNN achieves an accuracy of 83.69%. Although the convolutional layers learn some noise-robust features in a data-driven manner, the model fails to elevate its classification accuracy to the 90% industrial practicality threshold in the absence of physical prior constraints, exposing its tendency to overfit strong-noise artifacts.

Performance degradation of conventional fixed and externally-optimized cascades: The accuracies of FMD+1D-CNN and GWO-FMD+1D-CNN drop to 61.37% and 58.80%, respectively. The underlying reason is that conventional FMD and its externally-attached intelligent optimizer (GWO), under −8 dB extreme noise, easily misidentify high-energy random noise artifacts as fault impulses owing to their single-objective maximization of correlated kurtosis (CK). Owing to the gradient fracture between the front-end signal decomposition and the back-end CNN classifier, the erroneous front-end filtering directly contaminates the input, leading to numerous misclassifications downstream.

Performance advantage of the proposed network: In contrast, the proposed Continuous FMD-Net surpasses the 90% technical threshold with a significant margin, achieving a classification accuracy of 90.13%, and its four core evaluation metrics (accuracy, precision, recall, and F1-score) all outperform those of all compared methods. This demonstrates that the proposed design—unfolding physical FMD into a differentiable layer and enabling the decomposition and classification to be collaboratively optimized within a unified global gradient flow—achieves high-purity feature demodulation within a single network.

Table 4-1 Comprehensive diagnostic performance comparison on the −8 dB strong-noise blind test set.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Pure 1D-CNN	83.69	83.38	83.64	83.38
FMD+1D-CNN	61.37	62.53	61.36	60.97
GWO-FMD+1D-CNN	58.80	61.72	58.55	57.89
Continuous FMD-Net (Ours)	**90.13**	**91.06**	**90.05**	**89.96**

4.3 Cross-Model Feature Manifold Visualization

To analyze the root causes of the compared models' failure from the feature-space perspective, the UMAP two-dimensional projections of the test-set feature embeddings for the four methods are further exported, as shown in Fig. 5-1. A horizontal comparison of the four sub-figures reveals the following: in the manifold sub-figures of FMD+1D-CNN and GWO-FMD+1D-CNN, the sample points of each fault class exhibit large-scale interwoven mixing, with severely deviated intra-class centers, indicating that the externally-attached optimization falls into local optima under strong noise. In the Pure 1D-CNN sub-figure, although some clustering tendencies are evident, the manifold boundaries of multiple fault states remain adhered in the central and right regions, leading to decision ambiguity.

By contrast, the Continuous FMD-Net sub-figure shows that, even under −8 dB strong background noise, the sample clusters belonging to the 10 different damage conditions are separated relatively cleanly, exhibiting clearer topological boundaries and favorable intra-class compactness. These visualization results further demonstrate the effectiveness of the proposed network in separating the underlying fault feature manifolds.

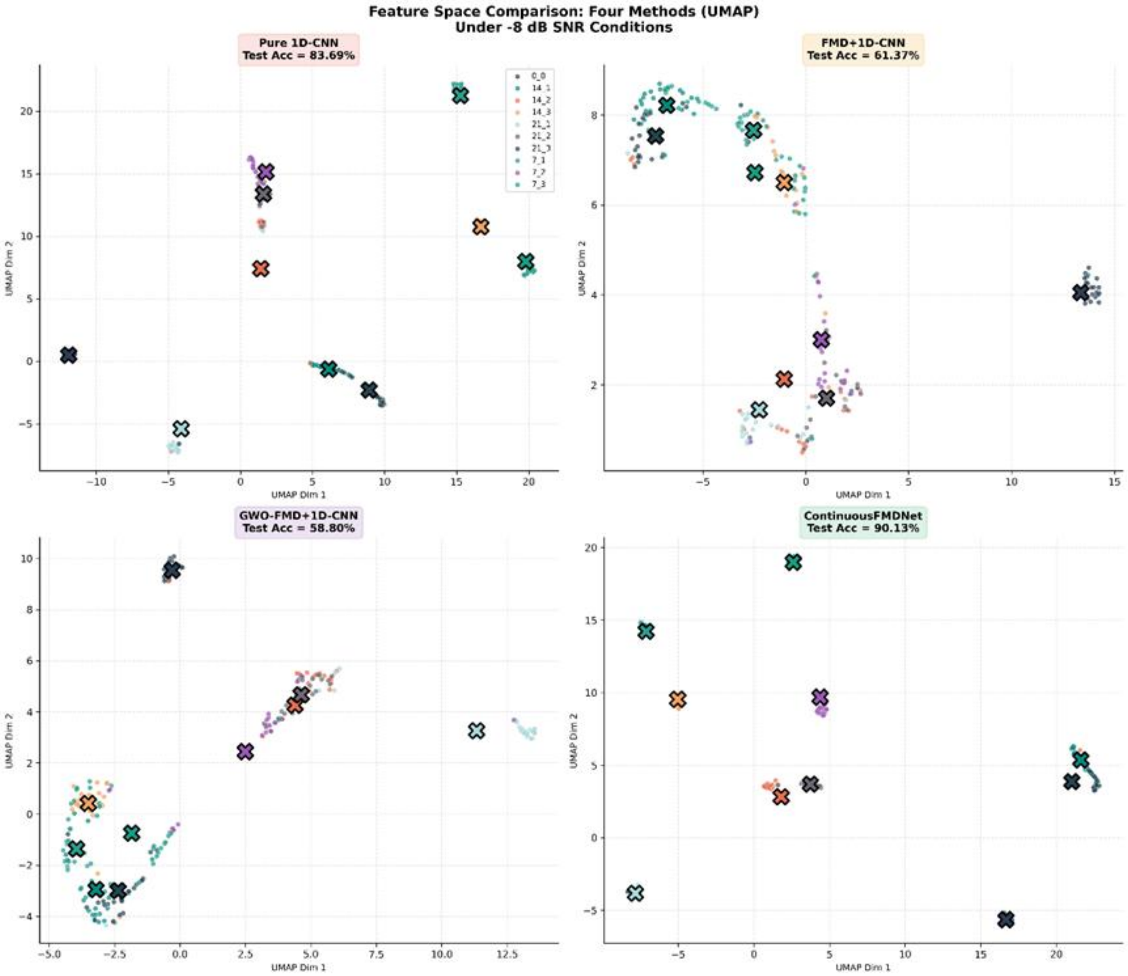

Fig. 5-1 UMAP two-dimensional projection comparison.

5. Conclusion

Addressing the challenges that conventional signal decomposition methods for rotating machinery rely on manual parameter tuning and are prone to mode mixing under extreme industrial strong-noise conditions, while purely data-driven methods lack dynamic physical priors, this paper proposes Continuous FMD-Net, a fully differentiable, end-to-end physics-driven algorithm-unrolling network. Experiments on the CWRU bearing benchmark dataset yield the following conclusions:

(1) Diagnostic performance: On the 10-class task covering different defect locations and damage diameters, the proposed network achieves an accuracy approaching 100% under 0 dB noise, and reaches 90.13% accuracy and 91.06% precision on the −8 dB strong-noise blind test set—representing a significant improvement over the fixed-parameter FMD+1D-CNN (61.37%) and the GWO-optimized GWO-FMD+1D-CNN (58.80%).

(2) Adaptive evolution of the filter window: Each channel can autonomously contract a differentiated effective window length according to its locked fault frequency band (e.g., 143 for Channel 1 and 106 for Channel 3), with the average effective length being only 47.7% of the preset maximum, providing interpretability evidence for the network's internal mechanism.

(3) Effective suppression of mode mixing: The global mean of the off-diagonal Pearson correlation coefficients in the latent feature space is only 0.042, indicating that the cross-correlation orthogonality penalty effectively constrains inter-channel redundancy and alleviates the mode mixing problem prevalent in conventional adaptive decomposition methods.

Acknowledgements

This work was supported by the Natural Science Foundation of Jilin Province (Grant No. 20230101208JC) and the Science and Technology Development Program of Jilin Province (Grant No. 20240304123SF). The authors would like to express their sincere gratitude for the financial support.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Author contributions

Meng Li: Supervision, Software, Method guidance, Review.

Jiaqi Xie: Conceptualization, Methodology, Writing, Investigation.

Hangyuan Gao: Writing of partial content, Literature collection.

References

[1] Lei, Y., Lin, J., He, Z., & Zuo, M. J. (2013). A review on empirical mode decomposition in fault diagnosis of rotating machinery. Mechanical Systems and Signal Processing, 35(1–2), 108–126. https://doi.org/10.1016/j.ymssp.2012.09.015

[2] Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics — A tutorial. Mechanical Systems and Signal Processing, 25(2), 485–520. https://doi.org/10.1016/j.ymssp.2010.07.017

[3] Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In IEEE International Conference on Prognostics and Health Management (pp. 1–8). Denver, CO, USA.

[4] Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7), 1483–1510. https://doi.org/10.1016/j.ymssp.2005.09.012

[5] Peng, Z. K., & Chu, F. L. (2004). Application of the wavelet transform in machine condition monitoring and fault diagnostics: A review with bibliography. Mechanical Systems and Signal Processing, 18(2), 199–221. https://doi.org/10.1016/S0888-3270(03)00075-X

[6] Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050

[7] Antoni, J. (2006). The spectral kurtosis: A useful tool for characterising non-stationary signals. Mechanical Systems and Signal Processing, 20(2), 282–307. https://doi.org/10.1016/j.ymssp.2004.09.001

[8] Wang, D., Tsui, K. L., & Miao, Q. (2018). Prognostics and health management: A review of vibration based bearing and gear health indicators. IEEE Access, 6, 665–676. https://doi.org/10.1109/ACCESS.2017.2774261.

[9] He, M., & He, D. (2017). Deep learning based approach for bearing fault diagnosis. IEEE Transactions on Industry Applications, 53(3), 3057–3065. https://doi.org/10.1109/TIA.2017.2661250

[10] Yan, R., Gao, R. X., & Chen, X. (2014). Wavelets for fault diagnosis of rotary machines: A review with applications. Signal Processing, 96, 1–15.

https://doi.org/10.1016/j.sigpro.2013.04.015

[11] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London Series A, 454(1971), 903–995. https://doi.org/10.1098/rspa.1998.0193

[12] Rilling, G., & Flandrin, P. (2008). One or two frequencies? The empirical mode decomposition answers. IEEE Transactions on Signal Processing, 56(1), 85–95. https://doi.org/10.1109/TSP.2007.906771

[13] Wu, Z., & Huang, N. E. (2009). Ensemble empirical mode decomposition: A noise-assisted data analysis method. Advances in Adaptive Data Analysis, 1(1), 1–41. https://doi.org/10.1142/S1793536909000047

[14] Torres, M. E., Colominas, M. A., Schlotthauer, G., & Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4144–4147). Prague, Czech Republic. https://doi.org/10.1109/ICASSP.2011.5947265

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

[16] Li, Z., Chen, J., Zi, Y., & Pan, J. (2017). Independence-oriented VMD to identify fault feature for wheel set bearing fault diagnosis of high speed locomotive. Mechanical Systems and Signal Processing, 85, 512–529. https://doi.org/10.1016/j.ymssp.2016.08.042

[17] McDonald, G. L., & Zhao, Q. (2017). Multipoint optimal minimum entropy deconvolution and convolution fix: Application to vibration fault detection. Mechanical Systems and Signal Processing, 82, 461–477. https://doi.org/10.1016/j.ymssp.2016.05.036

[18] Miao, Y., Zhao, M., & Lin, J. (2020). Identification of mechanical compound-fault based on the improved parameter-adaptive variational mode decomposition. ISA Transactions, 84, 82–95. https://doi.org/10.1016/j.isatra.2018.10.008

[19] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[20] Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2017). A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors, 17(2), 425. https://doi.org/10.3390/s17020425

[21] Zhao, R., Yan, R., Wang, J., & Mao, K. (2017). Learning to monitor machine health with

convolutional bi-directional LSTM networks. Sensors, 17(2), 273. https://doi.org/10.3390/s17020273

[22] Hou, L., & Bergmann, N. W. (2013). Novel industrial wireless sensor networks for machine condition monitoring and fault diagnosis. IEEE Transactions on Instrumentation and Measurement, 61(10), 2787–2798. https://doi.org/10.1109/TIM.2012.2200817

[23] Chen, Z., Gryllias, K., & Li, W. (2020). Mechanical fault diagnosis using convolutional neural networks and extreme learning machine. Mechanical Systems and Signal Processing, 133, 106272. https://doi.org/10.1016/j.ymssp.2019.106272

[24] Antoni, J., & Randall, R. B. (2006). The spectral kurtosis: Application to the vibratory surveillance and diagnostics of rotating machines. Mechanical Systems and Signal Processing, 20(2), 308–331. https://doi.org/10.1016/j.ymssp.2004.09.002

[25] Shao, H., Jiang, H., Zhao, H., & Wang, F. (2017). A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. Mechanical Systems and Signal Processing, 95, 187–204. https://doi.org/10.1016/j.ymssp.2017.03.034

[26] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

[27] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

[28] Zhang, S., Zhang, S., Wang, B., & Habetler, T. G. (2020). Deep learning algorithms for bearing fault diagnostics — A comprehensive study. IEEE Access, 8, 29857–29881. https://doi.org/10.1109/ACCESS.2020.2972859

[29] Wang, H., Liu, Z., Peng, D., & Qin, Y. (2022). Understanding and learning discriminant features based on multiattention 1DCNN for wheelset bearing fault diagnosis. IEEE Transactions on Industrial Informatics, 18(9), 5735–5745. https://doi.org/10.1109/TII.2021.3136069

[30] Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. Mechanical Systems and Signal Processing, 64–65, 100–131. https://doi.org/10.1016/j.ymssp.2015.04.021