Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.085747

ARTICLE

Learnable Wavelet Convolution and Sparsity-Enhanced Feature Extraction for Unsupervised Interpretable Fault Diagnosis in Mechanical Systems

Haitao Liu[1,*], Xuyang Wang[1], Shengcheng Quan[1], Qiaosheng Guo[2], Aichun Wang[3], Lie Yang[1], Tingfang Zhang[1] and Xiaojian Wu[1,*]

[1]School of Advanced Manufacturing, Nanchang University, Nanchang, China
[2]Zhaoyang Gevotai (Xinfeng) Technology Co., Ltd., Ganzhou, China
[3]Jiangling Motors Corporation, Ltd., Nanchang, China

*Corresponding Authors: Haitao Liu. Email: liuhaitao093430@ncu.edu.cn; Xiaojian Wu. Email: saintwu520@ncu.edu.cn

Received: 17 May 2026; Accepted: 29 June 2026; Published: 28 August 2026

ABSTRACT: Rapid advances in information and automation technologies have accelerated the development of smart manufacturing, thereby heightening the importance of reliable fault diagnosis for mechanical equipment. Although neural network-based algorithms are widely adopted in industrial applications due to their strong feature extraction and classification capabilities, their deployment in safety-critical fields such as aerospace remains limited. This limitation mainly arises from poor model interpretability and a heavy reliance on large-scale labeled training data. To address these challenges, this paper proposes an interpretable neural network framework that integrates discrete wavelet transform (DWT) with neural networks. Specifically, discrete wavelet filters are embedded into convolutional kernels to construct a novel convolutional layer capable of performing time-frequency transformation, in which the filter coefficients are learnable and a learnable thresholding mechanism is introduced for adaptive denoising. For fault identification, both local and global features extracted from the wavelet decomposition layers are organized into an anomaly detection feature matrix and subsequently classified using a support vector machine. To further improve diagnostic performance under strong interference and complex operating conditions, a novel sparsity-based measurement method is incorporated during feature matrix construction, significantly enhancing the extraction of discriminative signal features from complex signals. The proposed method is validated on an open-source mechanical fault diagnosis dataset, demonstrating superior diagnostic accuracy as well as strong interpretability. Notably, the model is trained exclusively using normal operating data without any labeled fault samples, thereby enabling an unsupervised fault diagnosis framework and effectively alleviating the challenge of fault data scarcity.

KEYWORDS: Fault diagnosis; neural network; discrete wavelet filter; sparsity; interpretability; unsupervised learning

1 Introduction

With the rapid advancement of technology and the deep integration of automation, mechanical equipment is becoming increasingly precise and complex. The widespread adoption of advanced technologies has imposed more stringent requirements on the safety, reliability, and operational stability of modern machinery, particularly in critical sectors such as power generation, aerospace, maritime transportation, and the nuclear industry. In these high-risk fields, equipment failures may not only cause substantial economic losses but also result in serious safety accidents and environmental pollution. Consequently, effective and reliable fault diagnosis methods are of paramount importance. In this context, anomaly diagnosis based on acoustic and vibration signals has attracted significant attention, as it enables the early detection and identification of incipient faults, allowing potential safety hazards to be mitigated in a timely manner and thereby ensuring safe and stable industrial production.

The increasing automation and intelligence of mechanical equipment have led to more intricate and sophisticated system structures, creating both unprecedented opportunities and significant challenges for fault diagnosis research. In parallel, the rapid advancement of artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has accelerated the development of data-driven diagnostic methodologies and their widespread adoption in industrial applications. A broad spectrum of data-driven approaches has been explored for fault diagnosis, which can be broadly grouped into several methodological families. Support Vector Machine (SVM)-based methods constitute one widely adopted family owing to their strong generalization capabilities, especially in small-sample scenarios such as rolling element bearings, gears, electric motors, engines, rotor systems [1,2], and hydraulic equipment [3,4]. To further enhance the diagnostic performance of SVM-based methods, existing research has mainly focused on two directions: the development of improved SVM variants and the optimization of SVM algorithms [5]. In terms of improved SVM models, Xu et al. [6] successfully applied the Least Squares Support Vector Machine (LSSVM) to fault classification in rotating machinery. With regard to algorithm optimization, Chen et al. [7] enhanced SVM performance using a chaotic particle swarm optimization algorithm, while Zhu et al. [8] improved fault diagnosis accuracy through a quantum genetic algorithm-optimized SVM. Expert systems represent another classical family that aims to codify domain-specific diagnostic knowledge into rule-based inference engines. Berredjem and Benidir [9] proposed an improved fuzzy expert system to enhance classification accuracy, and Wu et al. [10] developed an expert system based on a probabilistic neural network (PNN) for internal combustion engine fault diagnosis. While such systems can provide transparent and logically traceable decisions, their performance is critically dependent on the completeness and accuracy of the expert knowledge base. Knowledge acquisition remains a labor-intensive and error-prone process, and the static nature of rule sets severely constrains adaptability and self-learning in evolving industrial environments.

Neural Network (NN)-based architectures offer an alternative paradigm by integrating feature extraction and classification into a unified, learnable framework. Early adoptions employed shallow networks, but the field has been progressively dominated by deep architectures capable of automatically learning hierarchical representations from raw or lightly preprocessed signals. Convolutional neural networks (CNNs) have been extensively applied to gearbox [11], bearing [12], and variable-condition gearbox diagnostics [13], while recurrent neural networks (RNNs) with long short-term memory (LSTM) units have proven effective for capturing temporal dependencies in wind-turbine monitoring [14]. To further enhance feature discrimination, advanced variants incorporating multi-scale analysis [12], dual-scale residuals [15], and attention mechanisms [16] have been proposed. Wang et al. [17] proposed an RLHF (Reinforcement Learning from Human Feedback)-optimized industrial large model (CNC-VLM) to enhance the imbalanced detection capability for small-sample faults. Collectively, these studies confirm the remarkable representational power of deep NNs in mechanical fault diagnosis. Nevertheless, a fundamental limitation of purely NN-based models is their inherent "black-box" nature, which offers little insight into the physical mechanisms underpinning a diagnosis [18]. This opacity restricts their deployment in safety-critical applications where interpretability and trustworthiness are mandatory. In response, a hybrid family of physics-informed or signal-processing-integrated neural networks has emerged, wherein well-established signal-processing transforms are embedded directly into the network architecture. Representative strategies include substituting the first convolutional layer with a continuous wavelet transform [19], employing complex-valued kernels that mimic time-frequency transforms [20], or adopting parameterized sinc-function filters [21]. Xu et al. [22] proposed the Variable Operating Condition Large Language Model (VOC-LLM), which provides a practical and interpretable solution for predictive maintenance of industrial equipment. By blending domain knowledge with deep learning, these approaches achieve a favorable trade-off between performance and interpretability. However, they still typically require substantial amounts of labeled fault data for supervised training, which is often scarce or prohibitively expensive to acquire in industrial practice.

To simultaneously address the challenges of interpretability, reliance on labeled fault samples, and model complexity, a distinct direction focuses on unsupervised learnable decomposition frameworks. A notable example is DeSpaWN [23], a fully learnable, unsupervised deep cascade that mimics the structure of the fast discrete wavelet transform (FDWT). In this architecture, both the wavelet filters and the thresholding parameters are learned end-to-end from unlabeled high-frequency time series, yielding meaningful sparse representations with only a few hundred parameters.

Building upon the unsupervised learnable decomposition paradigm [23], this paper extends the framework along two coordinated directions. On one hand, multiple orthogonal wavelet bases are embedded into a cascaded architecture to enable adaptive basis selection, while the denoising thresholds and filter coefficients are jointly optimized under a unified reconstruction-sparsity objective. On the other hand, the BCSM metric [24] is introduced during feature matrix construction to quantify the sparse structure of signals. Critically, the synergy of these two methods allows for joint optimization toward the same end-to-end goal, yielding superior performance. This coordinated optimization, together with zero reliance on labeled fault samples and the preservation of physical interpretability, constitutes a substantive contribution that goes beyond a mere combination of learnable wavelets and sparse algorithms.

The core innovation therefore lies in a unified training paradigm in which the wavelet filter coefficients, the denoising thresholds, and the wavelet-domain sparse feature construction are jointly driven by a single objective function combining the reconstruction loss and the sparsity constraint. This distinguishes the proposed framework from prior learnable wavelet decomposition networks and wavelet-kernel neural networks, in which the wavelet basis, the thresholding parameters, and the feature metrics are typically determined separately and trained in a stage-wise manner.

The main contributions of this paper are as follows:

(1) A novel interpretable unsupervised fault diagnosis framework is proposed by embedding multiple orthogonal wavelet bases into the learnable wavelet decomposition cascade network architecture, in which the filter coefficients and the denoising thresholds are jointly optimized via end-to-end training under a unified reconstruction-and-sparsity objective. This design enables adaptive matching between wavelet bases and signal characteristics, thereby significantly enhancing diagnostic accuracy. The entire framework is trained using only normal operating data, eliminating the need for labeled fault samples and effectively alleviating the data scarcity bottleneck in industrial practice.
(2) A novel sparsity-based feature construction strategy is introduced by incorporating the BCSM metric into the wavelet-domain feature extraction process. Kurtosis and negentropy theory are integrated to effectively quantify the sparsity characteristics of the data. By jointly exploiting local and global detail coefficients across multiple decomposition levels, the proposed approach significantly enhances the discriminative representation of fault-related features under strong noise and complex operating conditions. Combined with an SVM-based anomaly classification scheme, this strategy leads to substantial improvements in diagnostic accuracy, robustness, and generalization performance while maintaining clear physical interpretability.

This paper is organized as follows. In Section 1, a comprehensive overview of the work undertaken by other scholars in this domain is provided. Section 2 is dedicated to a thorough exploration of the theoretical foundation for the proposed method, providing an in-depth understanding of its conceptual framework. Section 3 validates the effectiveness of the proposed method. Section 4 presents the experimental verification. Finally, Section 5 draws conclusions.

2 Theoretical Foundation

2.1 Neural Networks

Neural networks are fundamentally composed of two key processes: forward propagation and backward propagation. During forward propagation, input signals are transformed through successive network layers to generate predicted outputs. The loss function then evaluates the discrepancy between the predicted outputs and the target objectives, producing a loss value. Subsequently, during backward propagation, gradients of the loss with respect to the network parameters are computed and used to update the parameters accordingly. This forward-backward optimization procedure is iteratively performed until the model converges.

The convolutional layer serves as the core component of forward propagation. The input signal x undergoes convolution operations with randomly initialized kernels to produce feature mappings. The convolution process for the n-th channel in the m-th layer is given by Eq. (1):

$$h_n^m = w_n^m * x^m + b_n^m \tag{1}$$

where x^m denotes the input to the m-th convolutional layer, while w_n^m and b_n^m represent the weights and bias of the n-th channel in the m-th layer, respectively. The activation layer introduces nonlinear characteristics into neural networks via activation functions, enabling them to approximate complex data. The output of the m-th activation layer is Eq. (2):

$$z^m = f(h^m) \tag{2}$$

where $f(\cdot)$ is the activation function, with h^m and z^m being the input and output of the m-th activation layer, respectively. Pooling layers reduce computational complexity through downsampling while preserving essential features. The output of the m-th pooling layer is Eq. (3):

$$y^m = \mathrm{Downsample}(z^m) \tag{3}$$

During the backward propagation process, gradients of the loss function with respect to network parameters are computed using gradient descent and related algorithms, followed by iterative parameter updates. The parameter update phase relies on optimization algorithms, with gradient descent and its variants being representative.

2.2 Discrete Wavelet Transform

The convolution operation in neural networks can be expressed as the inner product of two vectors in signal processing. Similarly, the discrete wavelet transform (DWT), also grounded in inner-product operations, can be naturally formulated as a convolutional neural network architecture, enabling automatic and efficient extraction of meaningful and sparse representations from input signals.

The DWT employs wavelets constructed such that the set generated by dyadic scaling and translation forms an orthonormal basis. Together with the associated scaling function, the DWT recursively decomposes a signal into a coarse approximation c, representing a low-pass filtered version, and detail coefficients d, which correspond to high-pass (wavelet) components. This decomposition, often implemented as a cascade filter bank, is fully invertible owing to the orthonormal basis property in $L^2(\mathbb{R})$.

Its discrete expressions are shown in Eqs. (4) and (5) below.

$$c_j[n] = \sum_k h[k-2n]c_{j-1}[k] \tag{4}$$

$$d_j[n] = \sum_k g[k-2n]c_{j-1}[k] \tag{5}$$

Here, $c_j[n]$ represents the approximation coefficients at level j, with $c_0[n] = x[n]$ being the original signal; $d_j[n]$ denotes the detail coefficients at level j. The filters $h[n]$ and $g[n]$ correspond to the discretized scaling function and wavelet function, respectively, serving as low-pass and high-pass filters. Using a Quadrature Mirror Filter (QMF) bank, the filters satisfy the following relationships (Eqs. (6)–(8)):

$$g[n] = (-1)^n h[-n] \tag{6}$$

$$h_r[n] = h[-n] \tag{7}$$

$$g_r[n] = g[-n] \tag{8}$$

where $h_r[n]$ is the reconstruction low-pass filter ($h[-n]$ indicates time-reversed coefficients of h), and $g_r[n]$ is the reconstruction high-pass filter.

2.3 Box-Cox Sparse Measures (BCSM)

In signal processing, quantifying the sparsity of signals is a critical step toward extracting their essential features. The L_1 norm, as the most intuitive sparsity measure, provides a convenient basis for sparse representation. However, it merely performs linear aggregation of signal amplitudes and thus struggles to capture the concentrated distribution of signal energy in the time-frequency domain. As a result, more sophisticated metrics that reflect the underlying statistical properties of signals have been widely adopted. On one hand, kurtosis (Eq. (9)), a higher-order statistic, quantifies the deviation of a signal from a Gaussian distribution by measuring the sharpness and tail weight of its probability distribution. On the other hand, negentropy (Eq. (10)), rooted in information theory, directly measures the informational divergence between the signal distribution and a Gaussian reference. Although derived from different theoretical perspectives, both metrics provide a more intrinsic and robust characterization of non-Gaussianity and sparse structure, effectively addressing the limitations of amplitude-based measures. This enables more reliable extraction of informative components, redundancy elimination, and noise suppression in complex environments.

$$K_{l,h} = \frac{m_4\{|\bar{x}_{l,h}[n]|\}}{(m_2\{|\bar{x}_{l,h}[n]|\})^2} - 2 = \sum_{n=1}^{N}\left(\frac{SE_{l,h}[n]}{\sum_{b=1}^{N}(SE_{l,h}[b])}\right) \times \left(N\frac{SE_{l,h}[n]}{\sum_{c=1}^{N}(SE_{l,h}[c])}\right) - 2 \tag{9}$$

Here, the original signal $x[n]$ undergoes bandpass filtering (frequency band $l \le k \le h$) to obtain $x_{l,h}[n]$, which is then transformed via Hilbert transform to yield the analytic signal $\bar{x}_{l,h}[n] = x_{l,h}[n] + j \cdot \text{Hilbert}\{x_{l,h}[n]\}$, where $SE_{l,h}[n] = |\bar{x}_{l,h}[n]|^2$ represents the squared envelope of the analytic signal. The constant 2 denotes the kurtosis of complex Gaussian signals.

$$\psi_{l,h} = \left\langle \frac{|\bar{x}_{l,h}[n]|^2}{\langle|\bar{x}_{l,h}[n]|^2\rangle} \times \ln\left(\frac{|\bar{x}_{l,h}[n]|^2}{\langle|\bar{x}_{l,h}[n]|^2\rangle}\right)\right\rangle - (1-\gamma) = \sum_{n=1}^{N}\left(\frac{SE_{l,h}[n]}{\sum_{b=1}^{N} SE_{l,h}[b]}\right) \times \ln\left(N\frac{SE_{l,h}[n]}{\sum_{c=1}^{N} SE_{l,h}[c]}\right) - (1-\gamma) \tag{10}$$

Here,$\langle\cdot\rangle$ denotes the averaging operation, $1-\gamma$ represents the negentropy of complex Gaussian signals, and γ is the Euler-Mascheroni constant.

Comparing the two sparsity measurement methods reveals that the main difference in their formulas lies in the weight component $N\frac{SE_{l,h}[n]}{\sum_{c=1}^{N}(SE_{l,h}[c])}$ and $\ln\left(N\frac{SE_{l,h}[n]}{\sum_{c=1}^{N}(SE_{l,h}[c])}\right)$—specifically whether the logarithmic operation is applied. Therefore, we introduce the Box-Cox transformation to the weight component to develop a new sparsity measurement method BCSM [24], where Eqs. (11) and (12) describe the generalized Box-Cox transformation and its application to weights, respectively.

$$y^{(\lambda)} = \begin{cases} (y)^{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases} \tag{11}$$

$$\left(N\frac{SE_{l,h}[n]}{\sum_{n=1}^{N} SE_{l,h}[n]}\right)^{(\lambda)} = \begin{cases} \left(N\frac{SE_{l,h}[n]}{\sum_{n=1}^{N} SE_{l,h}[n]}\right)^{\lambda}, & \lambda \neq 0 \\ \ln\left(N\frac{SE_{l,h}[n]}{\sum_{n=1}^{N} SE_{l,h}[n]}\right), & \lambda = 0 \end{cases} \tag{12}$$

Here, λ is the transformation parameter. Eq. (13) combines the kurtosis and negentropy sparsity measures into a new method based on Eq. (12):

$$\mathrm{BCSM}_{l,h} = \begin{cases} \sum_{n=1}^{N}\left(\frac{SE_{l,h}[n]}{\sum_{b=1}^{N} SE_{l,h}[b]}\right) \times \left(N\frac{SE_{l,h}[n]}{\sum_{c=1}^{N} SE_{l,h}[c]}\right)^{\lambda} - C, & \lambda > 0 \\ \sum_{n=1}^{N}\left(\frac{SE_{l,h}[n]}{\sum_{b=1}^{N} SE_{l,h}[b]}\right) \times \ln\left(N\frac{SE_{l,h}[n]}{\sum_{c=1}^{N} SE_{l,h}[c]}\right) - C, & \lambda = 0 \end{cases} \tag{13}$$

where the weights are determined by the transformation parameter λ, and $C = \begin{cases} \frac{\mathbb{E}_{Y\sim\exp(0.5)}\{Y^{\lambda+1}\}}{(\mathbb{E}_{Y\sim\exp(0.5)}\{Y^{1}\})^{\lambda+1}}, & \lambda > 0 \\ 1-\gamma, & \lambda = 0 \end{cases}$ is the BCSM value for complex Gaussian signals. When λ is an integer, C becomes $C = \begin{cases} (\lambda+1)!, & \lambda = 1,2,3,\cdots \\ 1-\gamma, & \lambda = 0 \end{cases}$. The optimal bandpass filtered signal is then determined by $\arg\max_{l<h} \mathrm{BCSM}_{l,h}$.

3 Methodology

In this section, the structure of the proposed fault diagnosis framework, the novel sparsity-based feature construction strategy, and entire fault diagnosis procedure are introduced.

3.1 Structure of Learnable Cascade Network.

The discrete wavelet transform (DWT) offers strong interpretability in signal processing but shows limited robustness when extracting time-frequency features from complex data. Conversely, neural networks possess powerful feature-learning capabilities, yet their decision processes are often opaque and difficult to interpret. To combine the strengths of both approaches, this paper adopts a strategy by embedding wavelet filters within convolutional layers to construct a learnable network architecture that emulates the DWT's multiscale decomposition and reconstruction procedures. This design endows the model with data-adaptive feature-extraction capabilities while simultaneously enhancing its interpretability.

Wavelet decomposition can be mathematically expressed as the inner product between a signal and wavelet basis functions, a process that is highly analogous to the convolution operation in neural networks: both extract features via local weighted summations. Based on this similarity, we employ a convolutional neural network to emulate the cascaded structure of multi-scale wavelet decomposition. The network comprises L decomposition modules and L reconstruction modules, where the number of layers L is constrained to not exceed the base-2 logarithm of the input size, i.e., $L \leq \log_2(\text{input size})$. The main architecture of this method is shown in Fig. 1.

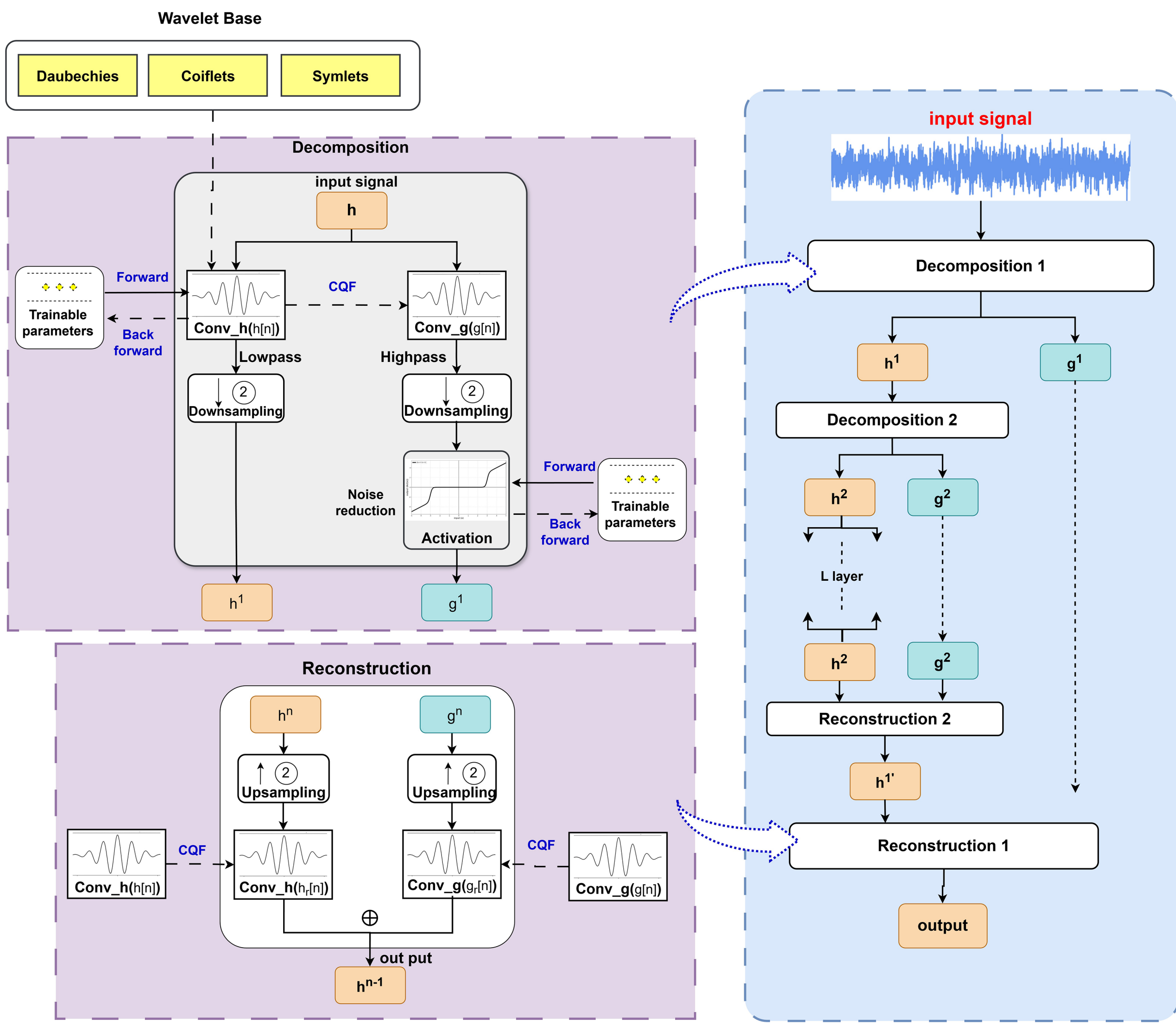

Figure 1: Structure of decomposition and reconstruction layers.

During decomposition, each encoder block accepts a single input and produces two output channels. The low-pass filter coefficients $h[n]$ corresponding to the scaling function and the high-pass filter coefficients $g[n]$ corresponding to the wavelet are embedded into the initial convolution kernels of these two channels, respectively. Convolving the input signal with these two sets of kernels yields the approximation coefficients and the detail coefficients, which is mathematically equivalent to performing a wavelet decomposition. The input to the current encoder block is the approximation coefficients output by the previous block, enabling a layerwise, scale-by-scale decomposition. To satisfy the Nyquist sampling theorem, each layer down-samples both convolution outputs by a factor of two (stride = 2), halving the data volume at each stage. This operation corresponds to the progressive reduction of the signal bandwidth after wavelet decomposition, thereby preserving multi-scale information while achieving dimensionality reduction and a structured feature representation.

The detail coefficients are then fed through a thresholding activation layer with learnable thresholds for noise suppression. This activation function is formed by a pair of oppositely directed sigmoid functions, and its expression is given in Eq. (14). After this layer, the encoder outputs the denoised detail coefficients $g[n]$ and the approximation coefficients $h[n]$. The approximation coefficients are used as the input to the next encoder block, and this process iterates to realize multi-level decomposition.

$$Act(x) = x\left[\frac{1}{1+\exp(10\cdot(x+b_-))} + \frac{1}{1+\exp(-10\cdot(x-b_+))}\right] \tag{14}$$

Here, b_+, b_- denote the positive and negative thresholds, respectively, both initialized to 3. As shown in Fig. 2, these denoising thresholds can be adaptively learned and updated for different datasets, enabling the model to denoise varying data and thereby enhance its feature-extraction capability.

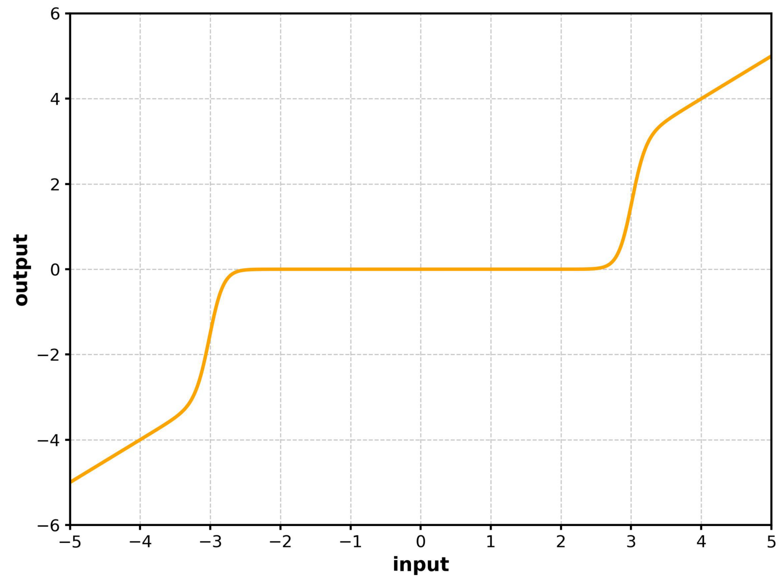

Figure 2: The dual-sigmoid activation function for denoising.

In reconstruction, the decoder functions as the inverse process of the encoder, responsible for progressively reconstructing the original signal from the multiscale coefficients. Each decoder block receives two inputs: the thresholded detail coefficients from the encoder at the same level and the reconstruction output from the previous decoder block. The decoder block performs reconstruction operations through two sets of transposed convolutions, embedding the reconstruction filter coefficients $h_r[n]$ and $g_r[n]$ into the initial kernels of these two channels, respectively. The results of the convolutions in both channels are then upsampled and combined, yielding the reconstructed signal at the current scale, which serves as one of the inputs to the next decoder block. Each signal undergoes upsampling with a stride of 2 to restore the original sampling rate, gradually recovering the signal's temporal resolution and bandwidth. This process strictly corresponds to the coefficient upsampling and filter-convolution steps in wavelet reconstruction, ultimately achieving precise, layer-by-layer reconstruction from the multiscale sparse representation to the original signal.

To achieve perfect reconstruction of the signal during the encoding and decoding processes, the design of the filters embedded in the convolutional layers strictly adheres to the properties of Conjugate Quadrature Filters (CQF), as specified in Eqs. (6)–(8). This property ensures that the filter sets used for decomposition and reconstruction satisfy an exact mathematical duality, thereby enabling lossless representation and reconstruction of the signal in the transform domain. Only the low-pass filter $h[n]$ needs to be designed, as the other three filters are derived directly from the CQF constraints. In designing the low-pass filter $h[n]$, this study selects filters corresponding to three wavelet families: Daubechies, Symlets, and Coiflets. All three families satisfy the conditions for orthogonal wavelets but exhibit distinct characteristics due to differing optimization objectives, such as vanishing moments, regularity, and symmetry. Daubechies wavelets are optimized to maximize vanishing moments, making them suitable for applications that demand strong sparsity. Symlets wavelets enhance symmetry while preserving the same vanishing moments, rendering them appropriate for applications sensitive to phase distortion. Coiflets wavelets jointly optimize the vanishing moments and symmetry of both the scaling and wavelet functions, offering balanced performance in signal approximation and image processing. Selecting these distinct wavelet filter families not only facilitates the investigation of how different wavelet embeddings affect model performance but also enables the choice of appropriate filters based on data characteristics, thereby enhancing model diagnostic accuracy. Furthermore, the coefficients of $h[n]$ in each layer are iteratively updated during neural network training, allowing the model to adaptively learn optimal filter sets from the data. This approach preserves the theoretical completeness of wavelet transforms while incorporating learnable filter design, enabling the transformation process to flexibly adapt to the time-frequency structure of diverse signals.

3.2 Objective Function and N-Adam

To learn optimal filters, we establish a robust loss function consisting of reconstruction loss and sparsity terms, as formulated in Eq. (15):

$$Loss = \frac{1}{n}\sum_{i=1}^{n}\|x_i - y_i\|_1 + \frac{1}{\mathrm{Card}(\{g^L\}_{L\in[1\cdots l]}, h^l)}\left(\sum_{i=1}^{l}|g^i| + |h^l|\right) \tag{15}$$

Here, x represents the original signal, and y represents the reconstructed signal. The first term of the loss function uses the L1 norm of the difference between the original signal and the reconstructed signal to represent the reconstruction term, which is used to constrain the reconstructed signal to approximate the original signal. $\mathrm{Card}(A)$ represents the number of elements in A. The second term of the loss function is the sparsity term, defined as the average of the absolute values of the detail coefficients g^l of each decomposition layer and the approximate coefficients h^l of the final layer. This sparsity constraint promotes discriminative feature learning by eliminating irrelevant components, thereby enhancing the model's fault identification capability.

After clarifying the composition of the loss function and the constraint objectives, it is necessary to iteratively update the network parameters through an optimization algorithm to minimize the loss value [25]. This paper introduces the Nadam optimizer, which retains the advantages of adaptive learning rates and independent updates for each parameter from Adam, while incorporating the Nesterov momentum's predictive gradient estimation mechanism. This allows for stronger directional foresight and curvature adaptability in parameter updates, thereby improving optimization efficiency. The computation process of the Nadam optimizer is as follows:

When updating parameters θ, the Nadam optimizer first computes the gradient g_t of the loss function with respect to the parameters and then corrects the gradient as shown in Eq. (16):

$$g' = \frac{g_t}{1 - \prod_{i=1}^{t} \mu_i} = \frac{\nabla_{\theta_{t-1}} f(\theta_{t-1})}{1 - \prod_{i=1}^{t} \mu_i} \tag{16}$$

where $\mu_t = \beta_1 \left(1 - 0.5 \times 0.96^{0.004t}\right)$ is the momentum decay coefficient at the current time step, controlled by β_1 as the base decay rate. The coefficient 0.004 regulates the decay speed, with $0.96^{0.004t}$ serving as the exponential decay term, which decays rapidly initially and gradually stabilizes later. By dynamically adjusting μ_t, the momentum remains small in early stages to prevent oscillations and gradually increases later to accelerate convergence. The factor 0.5 in the formula is used to smooth the decay process.

Next, the first moment estimate of the gradient m_t is calculated (17), and then it is corrected (18):

$$m_t = \mu_t m_{t-1} + (1 - \mu_t) g_t \tag{17}$$

$$m_t{}' = \frac{m_t}{1 - \prod_{i=1}^{t+1} \mu_i} \tag{18}$$

On this basis, the Nesterov momentum term $\hat{m}_t$ is introduced, as shown in Eq. (19):

$$\hat{m}_t = (1 - \mu_t) g' + \mu_{t+1} m_t{}' \tag{19}$$

In this context, the first term $(1 - \mu_t) g'$ directly uses the current corrected gradient, weighted by the momentum coefficient. The second term $\mu_{t+1} m_t{}'$ represents the momentum direction trend for the next step.

Next, the second moment estimate of the gradient is calculated (20) and then corrected (21):

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{20}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{21}$$

Finally, the parameters θ_t are updated based on the corrected second moment estimate of the gradient and the Nesterov momentum (22):

$$\theta_t = \theta_{t-1} - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{22}$$

where ε is a small value to prevent division by zero, and α is the learning rate. The Nadam optimizer performs predictive updates by using the "current corrected gradient" combined with the "future momentum foresight" from the Nesterov momentum. This approach retains the acceleration benefits of momentum while avoiding directional lag, leading to faster and more stable convergence and reducing oscillations.

3.3 Sparsity-Constrained Latent Space Construction

The model proposed in this paper abandons the traditional approach of adding labels to different data for training in fault diagnosis, and instead achieves diagnostic classification by constructing a latent space that characterizes signal features, thereby enhancing the interpretability of the diagnostic process.

During the latent space construction process, it is necessary to consider not only the global trend changes of the signal but also the local transient anomalies. For reconstructed signals, this paper selects the reconstruction residual features of the signals and calculates two types of features between the original signals and the reconstructed signal: Global Mean Residual (GMR) and Local Extreme Residual (LER) between original and reconstructed signals. The mathematical expressions of the two features, GMR and LER, are shown in Eqs. (23) and (24).

$$\mathrm{GMR}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\left|X(i) - \hat{X}(i)\right| \tag{23}$$

$$\mathrm{LER}(\mathbf{X}) = \max_{i}\left|X(i) - \hat{X}(i)\right| \tag{24}$$

Here, N is the total number of sample points in the signal, $X(i)$ is the original signal, and $\hat{X}(i)$ is the reconstructed signal. GMR characterizes the total energy loss during signal reconstruction, reflecting the overall signal deviation. In the early stages of faults in bearings and similar components, there is often a slow and gradual evolution of anomalies. During this process, errors are distributed throughout the time series and accumulate continuously, leading to a significant increase in the GMR value. Therefore, this feature is particularly sensitive to slow-evolving trend anomalies in the early stages of faults. LER can capture transient singular points in the signal, identifying the most significant local anomalies, and thus has high sensitivity for detecting impact-type faults.

For each decomposition level, we select wavelet-domain sparsity features, calculating Average Sparsity (AS) and Peak Sparsity (PS) of detail coefficients g^n. AS computes the mean L1-norm of coefficients at each level, quantifying energy distribution characteristics across different frequency bands and reflecting system vibration modal characteristics in the frequency domain. PS computes the maximum absolute values of coefficients at each level, detecting local singularities in the frequency domain with particular sensitivity to transient impacts. This yields 2L features in total. AS and PS are shown in Eqs. (25) and (26).

$$\mathrm{AS}_l(x) = \frac{\sum_x |D_l(x)|}{N_l} \tag{25}$$

$$\mathrm{PS}_l(x) = \max|D_l(x)| \tag{26}$$

where $D_l(x)$ is the sequence of detail coefficients at the l-th level, and N_l is the number of coefficients at that level.

The aforementioned four types of features are capable of effectively extracting both local and global feature information, as well as sparsity, at each decomposition level, thereby enabling the construction of a sparsity-constrained latent space for subsequent identification and classification. However, these features are overly simplistic and exhibit limited representational capability, making it difficult for the constructed latent space to accurately characterize data characteristics under complex operating conditions. To address this, this paper introduces the higher-precision BCSM sparsity feature based on the aforementioned features to accurately describe the sparsity characteristics of the data. The maximum BCSM value of the data, along with its corresponding frequency band ranges L and H-three features in total-are incorporated into the latent space to construct a sparsity-constrained latent space with a dimensionality of 2L+5. This latent space is then input into a support vector machine for classification and diagnosis, and the diagnostic accuracy is subsequently calculated.

In the calculation of the BCSM sparsity measure, the transformation parameter λ takes values 0 and 1, corresponding to negative entropy and kurtosis, respectively. When $0 \le \lambda \le 1$, the BCSM can stably reflect changes in signal sparsity; when $\lambda > 1$, it responds noticeably only when the signal distribution is extremely sparse, and is insensitive to general sparsity changes. In this paper, to balance the smooth sensitivity of negative entropy with the responsiveness of kurtosis to extreme sparsity, λ is set to 0.5. Consequently, the mathematical expression of the BCSM sparsity measure is obtained as shown in Eq. (27). In terms of frequency band selection, a filter bank is employed to divide the entire frequency range into several candidate bands, and the optimal band is automatically selected by maximizing the BCSM. The sensitivity analysis justifying this choice of λ and the influence of the frequency band division strategy on diagnostic accuracy and computational cost are reported in Section 4.3. Fig. 3 illustrates the framework of the fault diagnosis process.

$$\mathrm{BCSM}_{l,h} = \sum_{n=1}^{N} \frac{\mathrm{SE}_{l,h}[n]}{\sum_{b=1}^{N} \mathrm{SE}_{l,h}[b]} \times \left(N \frac{\mathrm{SE}_{l,h}[n]}{\sum_{c=1}^{N} \mathrm{SE}_{l,h}[c]} \right)^{0.5} - \frac{\mathbb{E}_{Y\sim\exp(0.5)}\{Y^{1.5}\}}{\left(\mathbb{E}_{Y\sim\exp(0.5)}\{Y^{1}\}\right)^{1.5}} \tag{27}$$

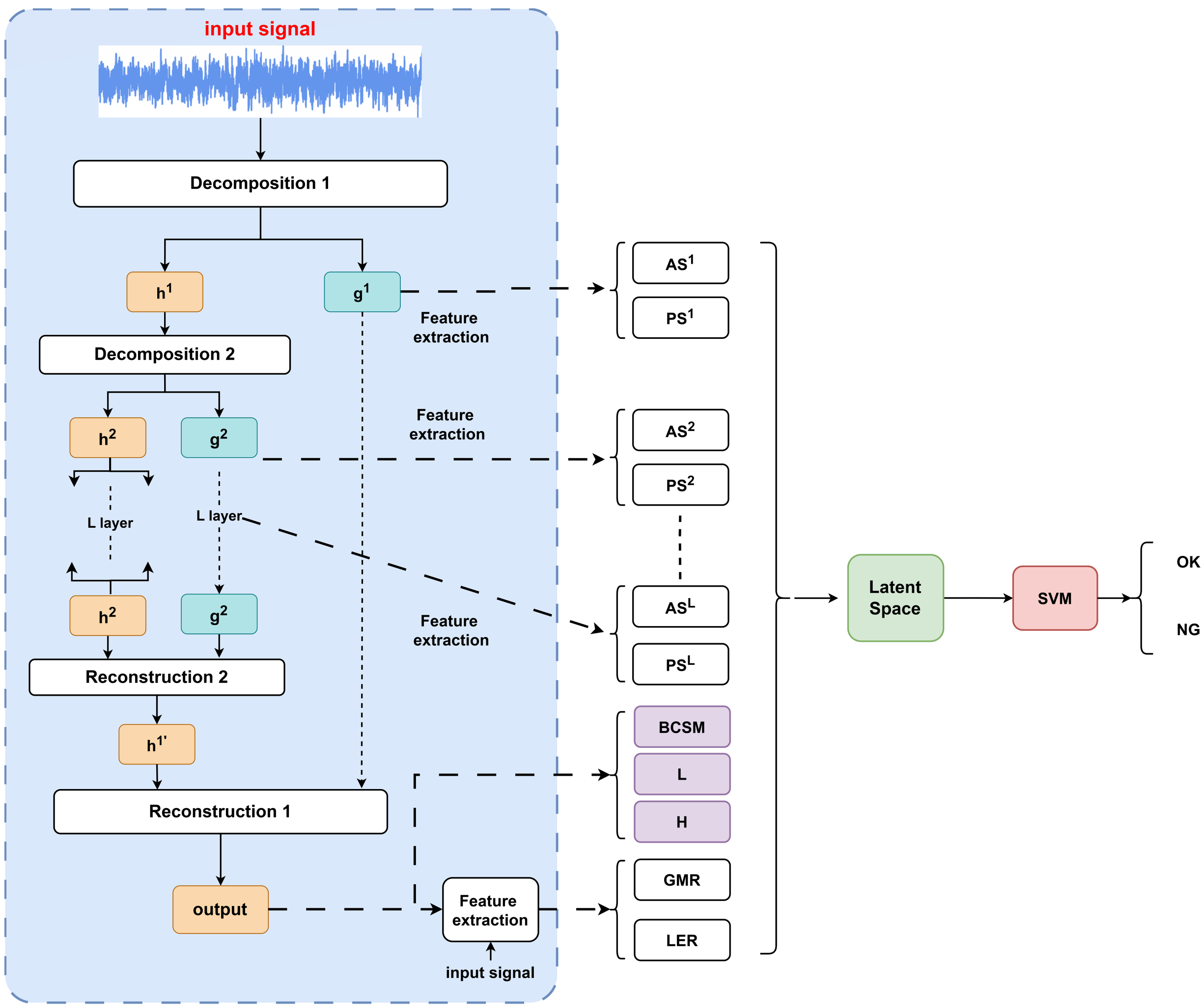

Figure 3: The framework of the fault diagnosis.

3.4 Training Procedure and SVM Classification

The proposed framework combines unsupervised feature extraction with supervised classification in a unified training pipeline that is driven by a single end-to-end objective on the wavelet decomposition side and a separate supervised objective on the classifier side. Training begins with data preparation, in which the original acoustic signals are segmented into fixed-length samples; only normal samples are used to optimize the wavelet decomposition network, while both normal and abnormal samples together with their labels are reserved for the downstream support vector machine classifier and for evaluation. The wavelet decomposition network is initialized with filter coefficients from a selected wavelet family (for example, the Coiflet 1 wavelet with filter length 6, as listed in Table 1) and is trained to minimize the joint reconstruction-and-sparsity loss in Eq. (15) via the Nadam optimizer, with the filter coefficients, the denoising thresholds, and all other learnable parameters updated simultaneously through end-to-end backpropagation without any label information. Once this unsupervised training converges, the trained network is frozen and reused as a fixed feature extractor: each sample, whether normal or abnormal, is passed through the network to obtain the latent-space features described in Section 2.3, namely the Global Mean Residual (GMR), the Local Extreme Residual (LER), the Average Sparsity (AS) and Peak Sparsity (PS) of the detail coefficients at each decomposition level, the maximum BCSM value, and its corresponding band indices (L, H), together yielding a $(2L + 5)$-dimensional feature vector. These feature vectors, paired with their normal/abnormal labels, are then fed into a support vector machine with a radial basis function (RBF) kernel, which constructs the maximum-margin decision boundary in the latent feature space; the class_weight parameter is set to balanced to mitigate the class imbalance inherent in the MIMII dataset, where abnormal samples are far less frequent than normal ones. During inference, the same trained network extracts the feature vector for each test sample, which is then classified by the trained SVM. Ground-truth labels are therefore used only in the SVM training and evaluation stages; the wavelet decomposition network is trained in a completely unsupervised manner, and the overall framework follows a hybrid paradigm of unsupervised feature extraction plus supervised classification rather than a fully unsupervised method. The SVM is employed not because the framework relies on supervision, but as a standardized evaluator to obtain quantitative accuracy measures that objectively assess the discriminative quality of the unsupervisedly learned features; this design allows us to rigorously validate the model's performance while retaining the core advantage of unsupervised representation learning, ensuring that the effectiveness of the proposed feature extractor is demonstrated through clear, comparable metrics without compromising its unsupervised nature.

Table 1: Filter lengths corresponding to different wavelet bases.

Family	**Daubechies**				**Symlets**			**Coiflets**		
Wavelet	db4	db5	db6	db8	sym4	sym5	sym8	coif1	coif2	coif3
Filter length	8	10	12	16	8	10	16	6	12	18

The convergence behavior of the wavelet decomposition network is illustrated in Fig. 4, which plots the training loss against the number of epochs on the Slider dataset. As can be observed, the loss drops rapidly within the first few epochs, indicating that the network quickly learns meaningful wavelet filters and denoising thresholds that capture the dominant time-frequency structure of the normal signals. Around the 50th epoch, the loss curve enters a relatively stable plateau with only minor fluctuations, suggesting that the filter coefficients and the threshold parameters have approached a near-optimal regime and that further updates only fine-tune the solution. By the 100th epoch, the loss has fully converged and remains essentially constant, indicating that the network has reached a stable reconstruction-and-sparsity trade-off and that additional training would not yield meaningful improvement. This rapid and stable convergence behavior validates the choice of the Nadam optimizer together with the joint reconstruction-and-sparsity objective, and it also justifies the use of 100 epochs as the default training horizon in the subsequent experiments.

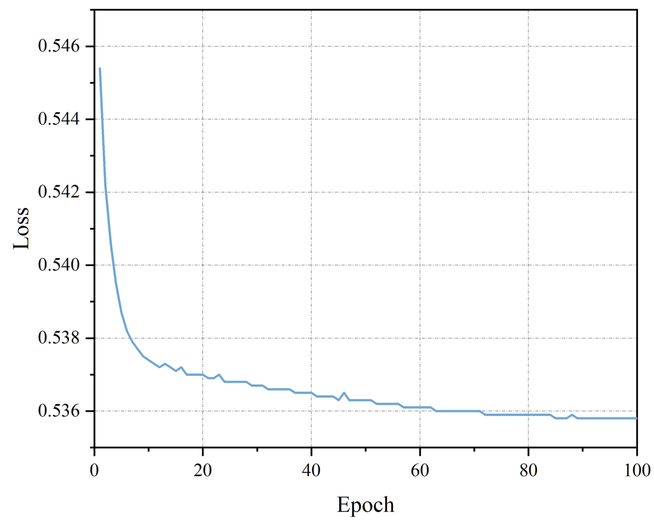

Figure 4: Training loss curve of the wavelet decomposition network on the Slider dataset.

4 Experiments and Verification

This section employs an open-source fault diagnosis dataset to validate the diagnostic performance of the proposed model. During the experiments, considering the core role of the filters in the convolutional layers, different wavelet basis functions are first embedded as initial filters in the convolutional layers to conduct ablation experiments, thereby verifying the feasibility of the proposed model architecture. The ten wavelet filters listed in Table 1, spanning the Daubechies, Symlets, and Coiflets families, are used as the initial filter coefficients for the comparative experiments. Additionally, to enhance the representation capability for weak fault features, a new sparsity measure method (BCSM) is introduced during the feature extraction phase, and its impact on model performance is systematically assessed through ablation experiments.

Table 1 reports the filter lengths corresponding to different wavelet orders. The filter length directly affects the time-localization capability and frequency resolution: a longer filter yields higher frequency resolution but weaker time localization, and vice versa. Selecting an appropriate filter length according to signal characteristics is therefore critical for diagnostic accuracy. The ten wavelet filters listed are embedded into the convolutional layers, respectively for comparative experiments.

The experimental environment and parameter configuration are summarized as follows. The self-learning wavelet network is trained in an unsupervised feature extraction manner, with the number of training epochs set to 100, the batch size set to 4, and the learning rate set to 0.001. The initial filter length is configured according to Table 1 (filter initial length). All models are trained on an NVIDIA RTX 2080Ti GPU. The downstream SVM classifier employs an RBF kernel with the penalty parameter set to $C = 1.0$. The dataset is partitioned into training and test sets at a 6:4 ratio, and the random seed is fixed to 42 to ensure reproducibility.

4.1 MIMII Dataset

The MIMII dataset [26] is an acoustic dataset designed for investigating and inspecting faulty industrial machinery. It records various types of industrial equipment including valves, pumps, fans, and Slide rail, capturing different abnormal sound patterns. Table 2 describes the detailed contents of the MIMII dataset, where ID represents different machine models, along with the sample counts for normal and abnormal data, and the corresponding fault types. Each sample was acquired at a 16 kHz sampling frequency using an 8-channel microphone array, with a duration of 10 s per sample(Wavelet decomposition levels = 17). Additionally, the acoustic data for each machine type includes three different signal-to-noise ratios (with factory noise): 6, 0, and −6 dB. For fault diagnosis, we exclusively used the audio data from the first microphone channel. In the fault classification process, a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel is employed; the detailed training procedure is described in Section 3.4.

Table 2: MIMII dataset content details.

Machine Type	ID	Normal	Abnormal	Anomalous Conditions
Valve	00	991	119	Open/close repeat with different timing.
	02	708	120	
	04	1000	120	More than two kinds of contamination.
	06	992	120	
Pump	00	1006	143	Suction from/discharge to a water pool.
	02	1005	111	
	04	702	100	Leakage, contamination, clogging, etc.
	06	1036	102	
Fan	00	1011	407	Normal operation. Unbalanced, voltage change, clogging, etc.
	02	1016	359	
	04	1033	348	
	06	1015	361	
Slide	00	1068	356	Slide repeat at different speeds. Rail damage, loose belt, no grease, etc.
	02	1068	267	
	04	534	178	
	06	534	89	

To test the diagnostic performance of convolutional layers embedded with discrete wavelet filters, we selected 10 wavelet filters of different families and orders from Table 1 and embedded them into convolutional layers for comparative experiments. Table 3 details the accuracy rates of the slider under different signal-to-noise ratios (SNRs) across various convolutional layers.

Based on the diagnostic results from the MIMII dataset, we draw the following conclusions:

1. The proposed model demonstrates high diagnostic accuracy across different machine models and various signal-to-noise ratio (SNR) conditions. Particularly at 6 dB SNR, the diagnostic accuracy exceeds 95% for all four machine models, reaching up to 100% in certain operational conditions. Compared with the baseline values from the MIMII dataset, models embedded with different filters show significant advantages in fault diagnosis accuracy, with improvements exceeding 30% in some datasets. These results fully validate the effectiveness of the proposed model for fault diagnosis tasks.
2. Models embedded with different filters exhibit variations in diagnostic accuracy, which become particularly pronounced in low SNR conditions and challenging datasets. For instance, on the −6 dB SNR "06"

slider data, the maximum accuracy difference among models reaches 3.2%. Different wavelet families possess distinct time-frequency characteristics (e.g., support length, symmetry, number of vanishing moments) that directly affect diagnostic performance. Nevertheless, diagnosis models learned from different wavelet filters generally maintain high accuracy across various data. Therefore, for specific data types, wavelet filters with better adaptability can be selected as the basis for model training to optimize performance.

3. Analysis of optimal diagnostic models across various datasets reveals that the best-performing models are predominantly trained using Coiflet wavelet filters. This indicates that models embedded with Coif-series filters are more suitable for fault diagnosis of the slider. The superior performance of Coiflet wavelet filters may stem from their excellent symmetry and compact support properties, which enable effective capture of transient impact features commonly found in mechanical signals, thereby enhancing diagnostic accuracy.

Table 3: The accuracy of slider.

Wavelet	Accuracy											
	−6 dB_slider				6 dB_slider				0 dB_slider			
	00	02	04	06	00	02	04	06	00	02	04	06
db4	98.42	94.21	88.81	85.2	99.82	99.63	98.95	95.6	99.3	98.88	94.76	91.2
db5	97.55	95.51	89.86	86	100	99.63	98.6	97.2	99.3	98.69	95.8	92.8
db6	98.07	95.89	89.16	86	100	99.63	98.95	93.2	99.65	99.07	96.5	92.8
db8	98.42	95.7	91.96	83.6	100	99.63	98.25	94.98	99.65	98.88	95.45	93.2
sym4	98.25	95.7	92.66	85.2	100	100	99.3	96.0	99.82	98.69	96.85	89.2
sym5	98.25	94.77	90.21	83.2	100	99.81	98.25	97.2	99.65	98.69	94.76	90.8
sym8	97.9	95.7	90.21	84.4	100	100	98.25	93.6	99.65	98.88	96.5	92.4
coif1	98.77	97.2	87.76	84.8	100	100	99.65	98.0	100	99.44	96.15	92.8
coif2	98.42	95.51	90.56	86.4	100	99.44	98.6	95.6	99.82	98.88	97.55	91.6
coif3	98.07	95.51	89.86	85.2	100	99.63	98.6	94.4	99.12	98.69	95.1	94.0
base	93	74	61	52	99	93	88	71	99	79	78	56

4.2 Improve

To address the issue of insufficient diagnostic accuracy of the model when handling certain complex data, this section focuses on enhancing its feature extraction capability. To this end, a novel sparsity measure method is introduced to strengthen the expression of key features. Additionally, ablation experiments are designed and conducted to systematically validate the actual contribution of the new method to the improvement of the model's performance.

To validate the effectiveness of the BCSM sparsity measure method, Fig. 5 presents the analysis results of the BCSM sparsity characteristics of the Slide data under different signal-to-noise ratio conditions. The main findings are as follows:

1. Abnormal data exhibit significantly higher overall sparsity (BCSM) than normal data. This primarily stems from transient features commonly present in abnormal data, such as impulsive bursts, spikes, abrupt transitions, or irregular oscillations. These features create highly non-uniform energy distributions in the sample signals, resulting in higher sparsity measures.

2. Normal signals typically demonstrate periodicity, stationarity, and gradual variations, exhibiting relatively uniform energy distributions. Consequently, their optimal BCSM values are generally distributed evenly across frequency bands. In contrast, the sparsity of abnormal data is predominantly governed by transient random signals like abnormal impacts or sudden disturbances. The inherent high-frequency characteristics of such signals concentrate their maximum sparsity primarily in higher frequency regions.
3. Under varying signal-to-noise ratio (SNR) conditions, the maximum BCSM values of samples and their corresponding frequency band ranges exhibit overall stability in distribution patterns, while clearly distinguishing characteristic differences between normal and abnormal signals. This phenomenon fully validates the strong robustness of the BCSM sparse metric method against noise interference. The results demonstrate that BCSM not only possesses exceptional feature extraction capability to effectively capture key discriminative features reflecting the underlying faults but also maintains high reliability in complex noise environments, thereby providing a robust feature metric basis for fault diagnosis.

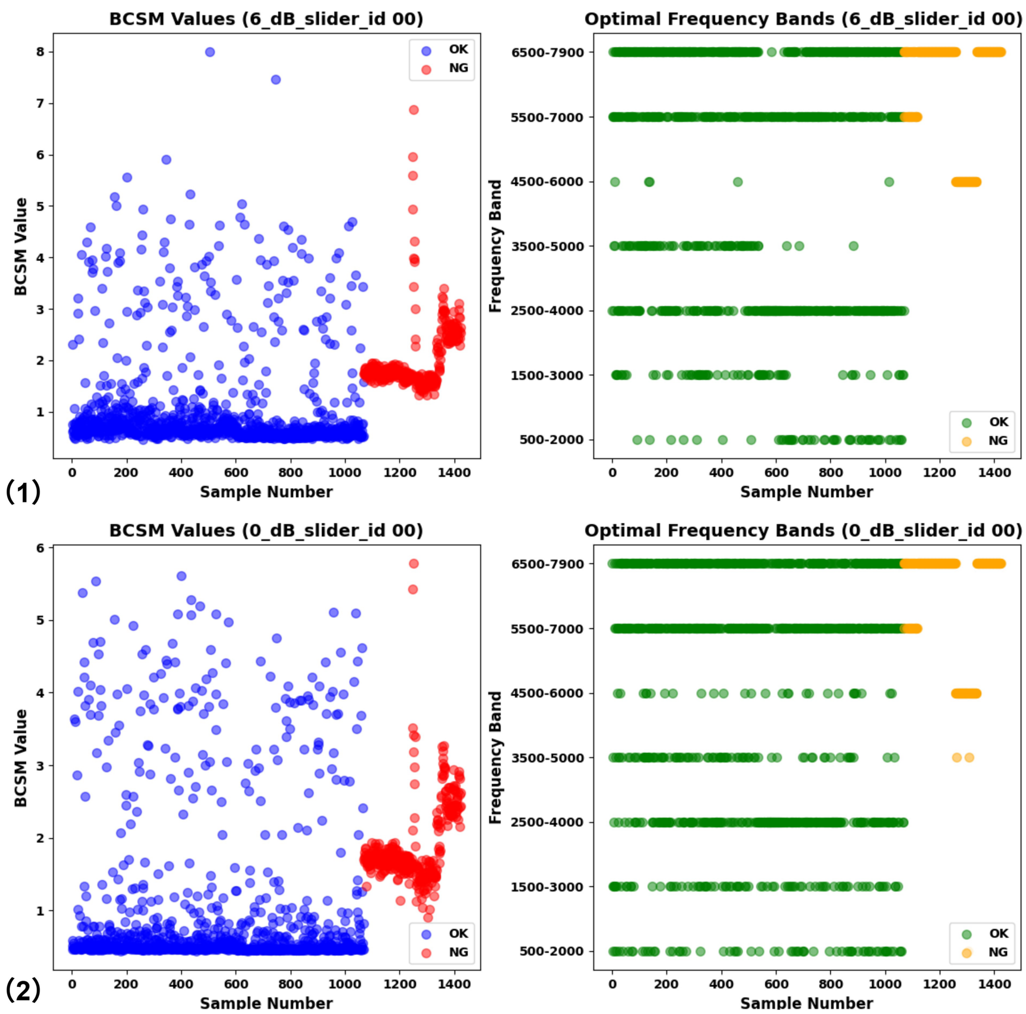

Figure 5: (Continued)

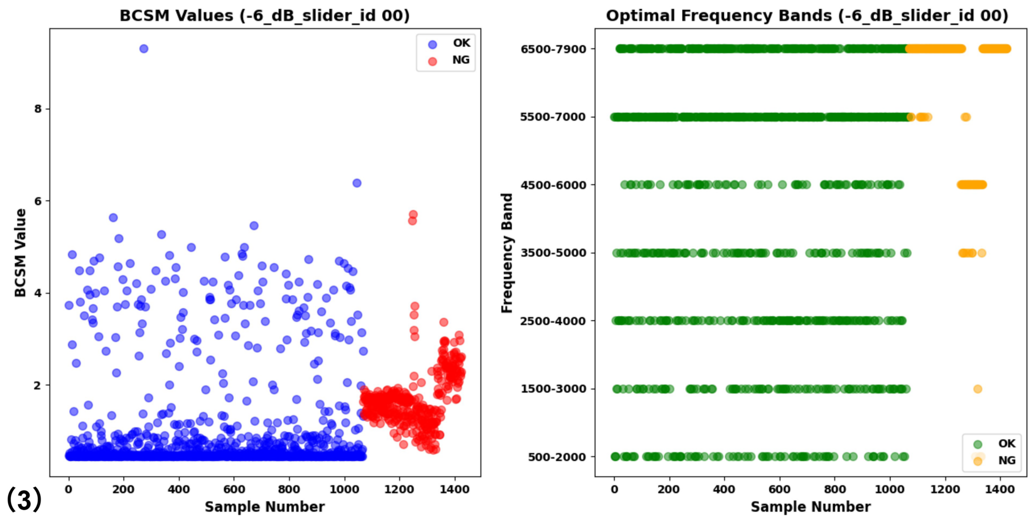

Figure 5: Sparsity scatter plots of the "00" model Slide under different signal-to-noise ratio (SNR) conditions. The horizontal axes represent sample indices. The left plot's vertical axis shows the maximum BCSM values of samples, while the right plot's vertical axis indicates the frequency bands corresponding to these maximum BCSM values. "OK" samples (blue) denote normal data, and "NG" samples (red) represent abnormal data. The transformation parameter was set to $\lambda = 0.5$ following reference [24]. (1) Sparsity scatter plot of slider sample id 00 and recognition results across different frequency bands at 6 dB; (2) Sparsity scatter plot of slider sample id 00 and recognition results across different frequency bands at 0 dB; (3) Sparsity scatter plot of slider sample id 00 and recognition results across different frequency bands at −6 dB.

Given the significant characteristics of the BCSM sparsity measure in abnormal data, such as the overall higher sparsity of fault data, the concentration of maximum sparsity in the high-frequency range, and its reliable performance even under noise interference-we fully leverage its representation capability to extract two key discriminative features from the reconstructed signal samples: the maximum BCSM value and its corresponding frequency range (l, h). Subsequently, the BCSM value and the frequency range (l, h) are fused as enhanced features into the original feature matrix, thereby constructing a new feature matrix. This feature enhancement strategy directly utilizes the dual advantages of BCSM in terms of fault sensitivity and strong noise resistance, providing a comprehensive characterization of the differences in the sparse structure of the signals. For the slider data at 6 dB SNR, the original model already achieves high diagnostic accuracy (average correct rate reaches 98%), making improvement unnecessary. Therefore, comparisons were only made for the other two SNR conditions.

To validate the enhancement provided by the BCSM sparsity measure, three models are constructed by incorporating the frequency range (l, h), the maximum BCSM value, and the combination of the BCSM value with frequency range (l, h) into the feature matrix, allowing for comparative experiments to analyze the performance of BCSM. Table 4 describes the diagnostic accuracy of different models. "db4 + LH" embeds the db4 wavelet filter and appends l, h values to the original feature matrix. "db4 + BCSM" embeds the db4 wavelet filter and appends the maximum BCSM value to the original feature matrix. "db4 + BCSM + LH" embeds the db4 wavelet filter and appends both the maximum BCSM value and l, h values to the original feature matrix. Other model structures are similar to the above description. The experimental results indicate:

1. In the models constructed using the db4, sym4, and coif1 wavelet bases, after introducing the BCSM sparsity features, the LH frequency range features, or their combinations, the diagnostic accuracy is higher than that of the baseline model in most cases. This confirms from a data perspective the general

effectiveness of the proposed feature enhancement strategy in improving the model's discriminative capability.

2. The model that combines the maximum BCSM value and the frequency range features achieves optimal or near-optimal performance in most cases. For example, in the four operating conditions at a 0 dB signal-to-noise ratio, the coif1 + BCSM + LH model achieves two accuracies of 100% and two accuracies near 99%. This suggests that the "global sparsity strength" information provided by BCSM, along with the "key frequency location" information provided by LH, can effectively complement each other, allowing for a more comprehensive characterization of fault features.
3. Overall, the family of models based on the coif1 wavelet demonstrates the highest performance limits under both signal-to-noise ratio conditions. Particularly at 0 dB, the coif1 series models generally outperform the db4 and sym4 series. This may relate to the inherent shape characteristics of the coif1 wavelet being more compatible with the fault features of the analyzed signals, and it indirectly reflects the potential value of treating the wavelet basis as a learnable parameter.

In summary, the fault diagnosis model proposed in this study, which embeds wavelet filter coefficients into the convolutional layers, possesses the capability to adaptively learn the optimal time-frequency representation from the data. This model architecture consistently achieves high-precision diagnostics, and its performance is influenced by the time-frequency characteristics of the wavelet basis. This directly validates the rationale and potential of treating the wavelet basis as a learnable parameter. Furthermore, by introducing the BCSM sparsity measure and key frequency range features to enhance the model, the diagnostic accuracy under complex conditions such as low signal-to-noise ratios is significantly improved. This further confirms that the architecture has good scalability and feature fusion capabilities, effectively integrating more robust feature extraction modules.

Table 4: The accuracy rates of models.

Model	**−6 dB Slider**				**0 dB Slider**			
	00	**02**	**04**	**06**	**00**	**02**	**04**	**06**
db4	98.42	94.21	88.81	85.20	99.30	98.88	94.76	91.20
db4 + LH	98.95	95.14	91.26	85.60	99.30	98.88	94.41	90.80
db4 + BCSM	98.60	94.21	93.01	87.20	99.30	98.88	97.55	92.40
db4 + BCSM + LH	99.12	95.14	92.66	86.80	99.47	98.88	96.50	92.40
sym4	98.25	95.70	92.66	85.20	99.82	98.69	96.85	89.20
sym4 + LH	98.60	95.89	93.01	84.80	99.82	99.07	98.25	90.40
sym4 + BCSM	98.42	95.70	95.10	85.60	99.82	98.88	97.20	89.20
sym4 + BCSM + LH	98.60	95.70	94.41	86.00	99.65	98.88	98.60	89.60
coif1	98.77	97.20	87.76	84.80	100.00	99.44	96.15	92.80
coif1 + LH	99.30	96.45	90.91	85.20	99.82	99.63	97.90	92.80
coif1 + BCSM	98.95	97.01	91.96	85.60	100.00	99.44	98.25	93.20
coif1 + BCSM + LH	99.12	96.07	93.01	85.20	100.00	100.00	98.95	93.20

To quantify the contribution of each component under different SNR conditions, the performance gains of the BCSM and LH modules relative to the corresponding baseline model are computed from Table 4. At 0 dB, the BCSM module yields gains of 1.00%, 0.14%, and 0.63% for db4, sym4, and coif1, respectively, while the LH module yields gains of −0.19%, 0.75%, and 0.44%; the combined BCSM + LH yields gains of 0.78%, 0.54%, and 0.94%, with coif1 + BCSM + LH achieving the highest accuracy of 98.04%. At −6 dB, the BCSM module yields gains of 1.60%, 0.75%, and 1.25%, while the LH module yields gains of 1.08%, 0.12%, and 0.83%; the combined BCSM + LH yields gains of 1.77%, 0.73%, and 1.22%, with sym4 + BCSM achieving the highest accuracy of 93.71%.

The distinct optimal configurations under the two SNR conditions reflect the noise-dependent behavior of each component. The BCSM metric integrates kurtosis and negentropy to quantify the sparsity structure of the signal and, as analyzed in Section 2.3, maintains high reliability in complex noise environments; consequently, its contribution is consistently positive under both SNR conditions and becomes the dominant discriminative feature at −6 dB. The LH feature provides the frequency band indices (L, H) corresponding to the maximum BCSM value, and its effectiveness depends on the accuracy of the optimal band selection. Under the 0 dB condition, the band selection remains relatively reliable, so the LH feature complements the BCSM value and coif1 + BCSM + LH achieves the best performance. Under the −6 dB condition, however, strong noise contaminates the band selection process and reduces the reliability of the LH feature, so that appending LH to BCSM no longer improves and may even slightly degrade the accuracy; in this case, sym4 + BCSM, which relies solely on the noise-robust BCSM sparsity value, achieves the best performance. The superiority of sym4 over coif1 at −6 dB is attributable to the enhanced symmetry of the Symlets wavelet, which provides better stability under strong noise, whereas the balanced vanishing moments and symmetry of the Coiflets wavelet are more advantageous under the moderate noise condition at 0 dB.

4.3 Hyperparameter Analysis

This subsection systematically examines the impact of the key hyperparameters of the proposed framework on diagnostic performance, namely the BCSM transformation parameter λ and the frequency band division strategy used for the optimal band search. All experiments in this subsection are conducted on the Slider dataset under the same experimental configuration as described in Section 4.

Sensitivity to $\boldsymbol{\lambda}$. According to previous studies [24], when $0 \leq \lambda \leq 1$, BCSM can stably reflect changes in signal sparsity; when $\lambda > 1$, it only responds significantly when the signal distribution is extremely sparse, and is insensitive to general changes in sparsity. Therefore, in this paper, we only consider the case of $0 \leq \lambda \leq 1$. To verify the rationality of setting $\lambda = 0.5$ in Eq. (27), a sensitivity analysis is conducted on the Slider dataset for $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$, corresponding to negentropy ($\lambda = 0$), kurtosis ($\lambda = 1$), and three intermediate states. As shown in Fig. 6, the model achieves the best performance at $\lambda = 0.5$; as λ deviates from 0.5, accuracy decreases on both sides. This trend is consistent with the theoretical analysis in Section 2.3: negentropy responds smoothly to general sparsity changes but lacks sensitivity to extreme sparsity, whereas kurtosis is sensitive to extreme sparsity but insensitive to general changes. The intermediate value $\lambda = 0.5$ balances the two and therefore yields the most discriminative sparsity representation for the fault-related transient components in the acoustic signals. The accuracy drop is more pronounced on the $\lambda = 0$ side than on the $\lambda = 1$ side, indicating that for the Slider data, missing the response to extreme sparsity (kurtosis) is more harmful than losing the smooth response to general sparsity changes; this further justifies the use of a λ value closer to the middle of the interval rather than one biased toward either end.

Frequency band division strategy. The sparsity computation of BCSM in Section 2.3 depends on the frequency band division scheme. In this work, the entire frequency range is divided into multiple sub-bands, and the granularity of this division directly determines both the frequency resolution of the BCSM-based feature and the computational cost. To evaluate this impact, Table 5 compares three schemes, all with a sampling frequency of 16 kHz. According to the Nyquist sampling theorem, the highest recoverable signal frequency without aliasing is 8 kHz; hence the upper bound of the band is set to 7.9 kHz. The three schemes are as follows: Scheme 1 divides the 0.5–7.9 kHz range into 5 non-overlapping bands; Scheme 2 uses 7 bands with a 0.5 kHz overlap between adjacent bands; Scheme 3 uses 9 bands with a 0.8 kHz overlap. As the number of bands and the overlap increase, the accuracy first increases and then decreases (84.4%, 85.2%, and 83.2%, respectively), while the inference time grows steadily from 0.4497 to 0.7590 s. Moderate subdivision improves frequency resolution and enables precise localization of fault-related frequency bands; however, excessive subdivision reduces the number of samples per band, undermining the statistical stability of the BCSM estimate, and the added overlap introduces redundant computation without commensurate discriminative gain. Therefore, Scheme 2 is adopted as the default configuration, offering a balanced trade-off between accuracy and efficiency, and is used throughout the remaining experiments.

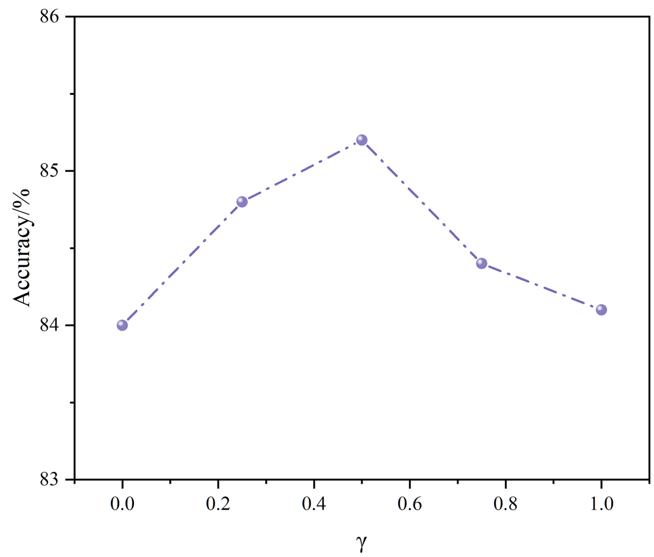

Figure 6: Sensitivity of diagnostic accuracy to the BCSM transformation parameter λ on the Slider dataset.

Table 5: Influence of frequency band division strategy on diagnostic accuracy and inference time.

ID	Frequency Bands (kHz)	Acc./Time (s)
1	0.5–2, 2–3.5, 3.5–5, 5–6.5, 6.5–7.9	84.4/0.4497
2	0.5–2, 1.5–3, 2.5–4, 3.5–5, 4.5–6, 5.5–7, 6.5–7.9	85.2/0.6063
3	0.5–2, 1.2–2.7, 1.9–3.4, 2.6–4.1, 3.3–4.8, 4–5.5, 4.7–6.2, 5.4–6.9, 6.1–7.9	83.2/0.7590

5 Discussion and Comparative Analysis

To address the insufficient diagnostic accuracy of the model when processing certain complex data, this section focuses on enhancing its feature extraction capability. For this purpose, we introduce a novel sparsity measurement method to strengthen the representation of critical features. Concurrently, we design and conduct ablation experiments to systematically validate the actual contribution of the new method to model performance improvement, quantify the effects of individual components, and perform comparative analysis with existing models. Furthermore, to comprehensively evaluate the model's practicality, we conduct rigorous verification from both unsupervised learning and interpretability perspectives, ensuring the model not only possesses efficient diagnostic capabilities but also provides clear decision-making basis.

5.1 Comparative Analysis

Based on the systematic evaluation of the differences in feature extraction and diagnostic performance among model variants constructed with different wavelet bases through ablation experiments, this section establishes a comprehensive multi-level and multi-dimensional evaluation system to thoroughly validate the overall superiority of the proposed model. Firstly, representative similar methods in the current field are selected for comparative experiments. Starting from the two core indicators of diagnostic accuracy and robustness, the system evaluates the performance improvements of the proposed model relative to existing methods, objectively presenting its practical value in complex fault diagnosis tasks. Subsequently, to further examine the model's potential from both engineering applications and academic innovation perspectives, this paper conducts in-depth analyses from two key dimensions: interpretability and unsupervised learning capability, to elucidate its applicability in real-world scenarios and its theoretical contributions.

Table 6 shows the average accuracy rates for different models across various datasets. WKN_Laplace and WKN_Morlet [19] embed the Laplace kernel and Morlet kernel, respectively, into the first layer of a conventional convolutional neural network. SincNet [21] embeds parametrized sinc functions into the first layer of a neural network. Two representative unsupervised deep learning methods are further introduced as comparative baselines: One-ClassDeepSVDD, which maps normal samples into a hypersphere space and uses the hypersphere radius as the anomaly threshold, and OutlierNet, which performs anomaly detection based on autoencoder reconstruction error. All methods are trained and tested on the same MIMII dataset under identical settings.

Table 6: The average accuracy rates for different models across various datasets.

Model	Average Accuracy (%)	
	−6 dB Slider	0 dB Slider
WKN_Laplace [19]	88.19	89.30
WKN_Morlet [19]	84.9	85.12
SincNet [21]	79.57	91.43
One-ClassDeepSVDD [27]	67.0	79.1
OutlierNet [28]	81.0	89
base [26]	70	78
db4	91.66	96.04
db4 + LH	92.74	95.85
db4 + BCSM	93.26	97.03
db4 + BCSM + LH	93.43	96.81
sym4	92.95	96.14
sym4 + LH	93.08	96.88
sym4 + BCSM	93.70	96.28
sym4 + BCSM + LH	93.68	96.68
coif1	92.13	97.09
coif1 + LH	92.97	97.54
coif1 + BCSM	93.38	97.72
coif1 + BCSM + LH	93.35	98.04

The comparative results in Table 6 indicate that the proposed coif1 + BCSM + LH model outperforms all comparative methods, including both supervised and unsupervised baselines, under both SNR conditions. At −6 dB, it achieves 93.35%, exceeding the second-best WKN_Laplace (88.19%) by 5.16 percentage points and the best unsupervised baseline OutlierNet (81.0%) by 12.35 percentage points. At 0 dB, it achieves 98.04%, exceeding the second-best SincNet (91.43%) by 6.61 percentage points and OutlierNet (89.0%) by 9.04 percentage points. The proposed method outperforms both supervised and unsupervised baselines, validating the advantage of the hybrid paradigm of "unsupervised feature extraction plus supervised classification." The ablation results show that introducing the LH, BCSM, and LH + BCSM modules consistently improves accuracy, though the model exhibits selectivity across SNR conditions, as analyzed in Section 4.

To verify cross-machine generalization, the proposed method is further evaluated on the Fan data from the MIMII dataset, which differs from the Slider data in signal characteristics, fault modes, and noise properties. As shown in Table 7, the proposed method achieves 97.11% at 0 dB and 82.88% at −6 dB, exceeding the second-best OutlierNet by 13.61 and 12.63 percentage points, respectively. This result is consistent with the conclusions on the Slider dataset, indicating good cross-machine generalization. The performance advantage is more pronounced under the low SNR condition, further validating the effectiveness of the BCSM sparsity measure under strong noise.

Table 7: Comparison with unsupervised methods on Fan dataset.

Model	0 dB Fan (%)	−6 dB Fan (%)
Base	79.5	66.25
One-ClassDeepSVDD	83.34	64.70
OutlierNet	83.5	70.25
coif1+BCSM+LH (Ours)	97.11	82.88

To evaluate the deployment feasibility on resource-constrained edge devices, Table 8 reports the training time per sample per epoch, the inference time, and the parameter count of the proposed model on the Slider dataset. The total parameter count is 136, the training time is approximately 0.0039 s, and the inference time is approximately 0.61 s, with no significant difference across SNR conditions. Since the filter coefficients and denoising thresholds are learnable parameters of limited quantity, the overall overhead remains controllable despite the full cascade architecture, indicating good deployment feasibility on edge devices.

Table 8: Training time, inference time, and parameter count of the proposed model on the Slider dataset.

SNR (dB)	Train Time (s)	Inference Time (s)	Number of Parameters
−6	0.00386	0.6108	136
0	0.00390	0.6061	136

The comparative analysis of the experimental results presented in Table 6 indicates that the proposed model demonstrates significant advantages in both diagnostic accuracy and robustness as core indicators. Under different signal-to-noise ratio conditions, the diagnostic accuracy of this model not only consistently exceeds the baseline level but also outperforms the three comparison models in all aspects. This result suggests that the model exhibits strong adaptability and stable discriminative performance when facing noise interference, validating its ability to maintain efficient and reliable feature extraction mechanisms even under complex operating conditions.

5.2 *Unsupervised Learning and Interpretability*

The proposed method constructs an unsupervised analysis framework with both high diagnostic performance and interpretability by embedding discrete wavelet filters in neural networks and extracting feature information from decomposition coefficients at each level for fault diagnosis. The interpretability method involves projecting the extracted high-dimensional features into a 2D visualization space through nonlinear dimensionality reduction techniques, thereby intuitively revealing the cluster structures and decision boundaries of fault features.

During the dimensionality reduction process, it was observed that changes in the 3D visualization perspective lead to high variability in distribution patterns, thereby increasing analytical complexity. Therefore, this study uniformly reduced the data to a 2D space and adopted the unsupervised dimensionality reduction method t-SNE for comparative analysis. As shown in Fig. 7, the t-SNE-based dimensionality reduction distribution clearly reveals the intrinsic clustering structure of the data.

As shown in Fig. 7, the characteristic distributions of different datasets exhibit significant spatial separation properties. Even data of the same category may distribute across different regions due to feature variations, demonstrating that the extracted high-dimensional features effectively represent differences between signals.

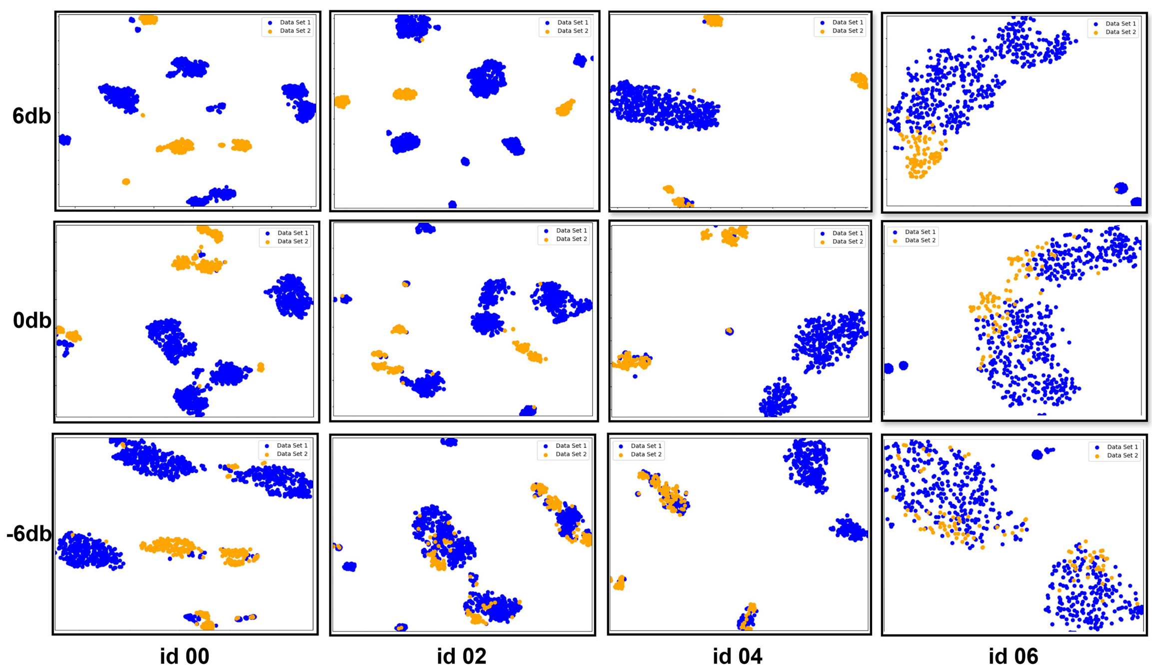

Figure 7: Distribution of Slider data features in low-dimensional space. The figure shows the 2D visualization of high-dimensional features from four types of Sliders under three different signal-to-noise ratios after t-SNE dimensionality reduction. Blue represents normal data, yellow represents abnormal data.

Under the 6 dB signal-to-noise ratio (SNR) condition, the two-dimensional projections of normal and abnormal data cluster in mutually exclusive regions, forming clear decision boundaries that enable direct fault discrimination. However, in complex noise environments with −6 dB SNR, partial overlap occurs between normal and abnormal samples, resulting in reduced separability. This visualization directly reveals the root cause of performance variations: the near-perfect accuracy (approaching 100%) under high SNR (6 dB) conditions stems from strong separability in feature space, while accuracy degradation under low

SNR conditions attributes to interference from feature confusion regions. This phenomenon validates the interpretability of the model's decision-making mechanism. To quantitatively evaluate the clustering quality, two metrics are employed: AKNH, which reflects intra-class compactness (higher is better), and BER, which reflects inter-class intrusion (lower is better). Table 9 reports the values for four machine models on the Slider data under three SNR conditions. At 6 dB, AKNH approaches 1 (minimum 0.9634) and BER approaches 0 (maximum 0.0366), indicating tight intra-class clustering and clear inter-class separation. As SNR decreases, AKNH declines and BER rises; for example, the AKNH of Id_06 drops from 0.9634 at 6 dB to 0.8194 at −6 dB, and BER rises from 0.0366 to 0.1806. This quantitative result is consistent with the t-SNE visualization and further validates the interpretability of the model's decision mechanism.

Table 9: AKNH and BER values under different SNR conditions on the Slider dataset.

	Id_00		Id_02		Id_04		Id_06	
SNR	**AKNH**	**BER**	**AKNH**	**BER**	**AKNH**	**BER**	**AKNH**	**BER**
6 dB	0.9983	0.0017	0.9987	0.0013	0.9749	0.0251	0.9634	0.0366
0 dB	0.9908	0.0092	0.9840	0.0160	0.9323	0.0677	0.9024	0.0976
−6 dB	0.9638	0.0362	0.8990	0.1010	0.8334	0.1666	0.8194	0.1806

The "unsupervised" property in this work specifically refers to the training of the wavelet decomposition network, which uses only normal data without label information; the subsequent SVM classifier is trained in a supervised manner using labeled samples (Section 3.4). The overall framework is therefore a hybrid paradigm of unsupervised feature extraction and supervised classification, not a fully unsupervised method. The unsupervised dimensionality reduction techniques (PCA, t-SNE, Isomap) are employed solely as visualization tools to inspect the latent space structure and interpret the SVM decision behavior; they do not participate in the classification decision. Ground-truth labels are used only for computing diagnostic metrics and for SVM training, not for any training procedure of the wavelet decomposition network.

To quantitatively evaluate the clustering quality, two metrics are employed: Average K-Nearest Neighbor Homogeneity (AKNH), which measures the label purity in local neighborhoods and is commonly used to assess the quality of feature representations, dimensionality reduction effects, or clustering reasonableness; and Boundary Misclassification Rate (BER), which reflects the degree of inter-class intrusion (lower is better). Table 9 reports the values for four machine models on the Slider data under three SNR conditions. At 6 dB, AKNH approaches 1 (minimum 0.9634) and BER approaches 0 (maximum 0.0366), indicating tight intra-class clustering and clear inter-class separation. As SNR decreases, AKNH declines and BER rises; for example, the AKNH of Id_06 drops from 0.9634 at 6 dB to 0.8194 at −6 dB, and BER rises from 0.0366 to 0.1806. This quantitative result is consistent with the t-SNE visualization and further validates the interpretability of the model's decision mechanism.

Fig. 8 summarizes the complete framework of this model for fault diagnosis.

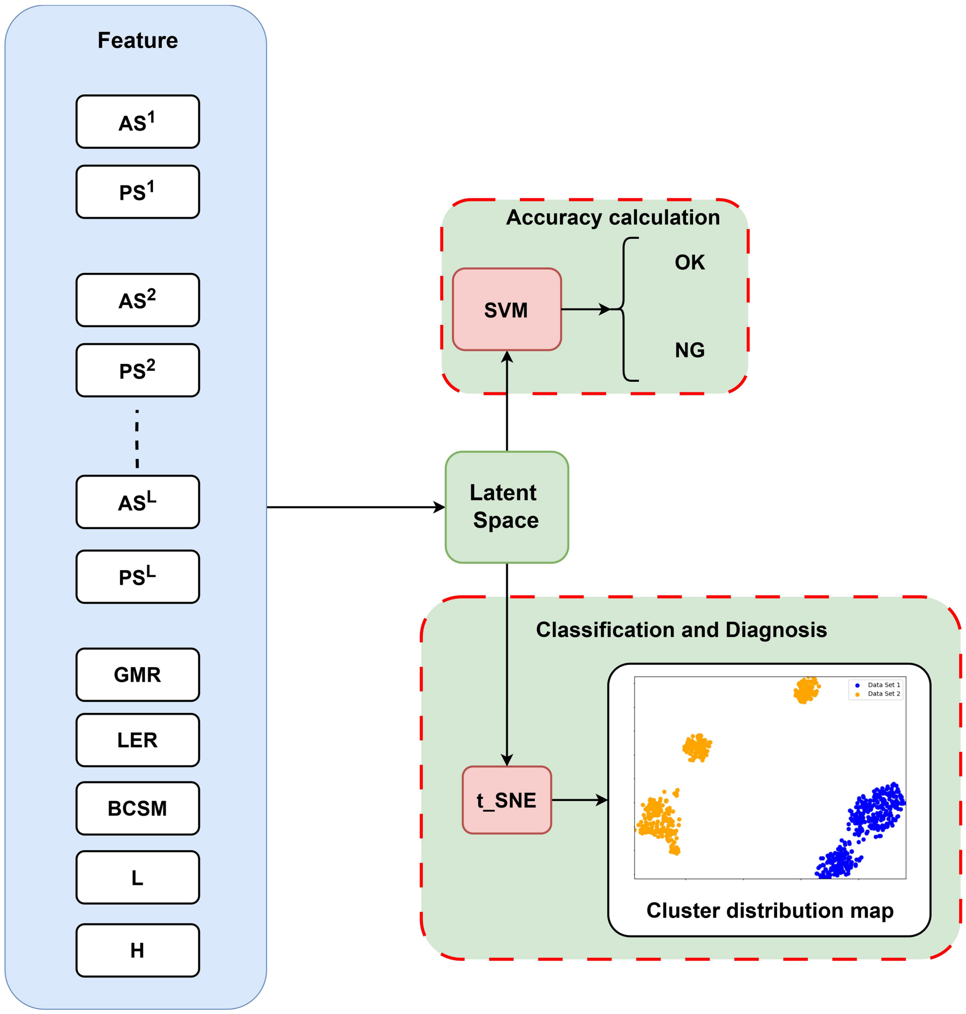

Figure 8: Comprehensive fault diagnosis framework.

6 Conclusion

In this paper, an innovative model that integrates discrete wavelet filters with a neural network is proposed. This model embeds discrete wavelet filters into the convolutional layers of the neural network. By constructing an adaptive learning mechanism, it iteratively optimizes the filter parameters to precisely adapt to data characteristics. Simultaneously, it introduces learnable activation functions to achieve efficient filtering and noise reduction for input signals, significantly enhancing data quality and model performance. The model utilizes the neural network architecture to simulate the time-frequency analysis process of discrete wavelet transform. It extracts features from the detail coefficients at each decomposition layer for fault diagnosis and classification, and introduces a novel sparsity measure method (Box-Cox Sparsity Measure, BCSM) to enhance model performance. The core conclusions of this paper can be summarized as follows:

1. The Robustness and Adaptability of the Model: The proposed model demonstrates excellent adaptability to embedding different types of discrete wavelet filters and achieves high fault diagnosis accuracy under various filter configurations.

2. Effectiveness of BCSM Sparsity Measure: The introduced Box-Cox Sparsity Measure (BCSM) method effectively characterizes the sparsity characteristics of different types of data, significantly enhancing the model's feature extraction and representation capabilities.
3. Advantage in Model Interpretability: Compared to traditional neural network models, the design of embedding discrete wavelet filters endows the model with stronger interpretability, and its processing flow aligns more closely with the physical significance of signal processing.
4. Unsupervised Learning and Resolution of Scarce Sample Problem: Model training relies exclusively on normal-state data, requiring no labeling of any fault samples. This effectively overcomes the challenge of fault sample scarcity while achieving fully unsupervised learning throughout the entire process.

It should be noted that the proposed method may experience performance degradation under conditions such as extremely low signal-to-noise ratios, unseen operating conditions, different sampling frequencies, non-stationary operating states, or when the machine's fault mode differs from those in the training data. So, in future research, deep learning can be deeply integrated with cutting-edge techniques from other signal processing fields to explore multimodal fusion-based fault diagnosis schemes, aiming to improve identification accuracy in extremely complex noise environments. At the same time, the application of unsupervised learning and semi-supervised learning in fault diagnosis can be further expanded to reduce reliance on labeled data and fully unleash the generalization capability of the model. Furthermore, extending the proposed method to more practical engineering scenarios will help promote the large-scale application of the technology.

Acknowledgement: None.

Funding Statement: This work was jointly supported by the National Natural Science Foundation of China (NSFC) (Grant No. 12104153), the China Postdoctoral Science Foundation (Grant No. 2021M701963), and the Ganzhou City Scientific and Technological Innovation Talent Program. Youth Talent Program (Grant No. 2025RCCC0001).

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Haitao Liu, Qiaosheng Guo, Lie Yang; methodology: Haitao Liu, Aichun Wang; software: Haitao Liu; data collection and curation: Xuyang Wang, Qiaosheng Guo; analysis and interpretation of results: Haitao Liu, Xuyang Wang, Lie Yang; draft manuscript preparation: Xuyang Wang, Shengcheng Quan; supervision: Tingfang Zhang, Xiaojian Wu; funding acquisition: Haitao Liu; final version of the manuscript preparation: Haitao Liu, Xuyang Wang, Aichun Wang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are from the publicly available MIMII dataset [26].

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pang B, Tang G, Zhou C, Tian T. Rotor fault diagnosis based on characteristic frequency band energy entropy and support vector machine. Entropy. 2018;20(12):1–22. doi:10.3390/e20120932.
2. Wang Z, Fan J. Fault early recognition and health monitoring on aeroengine rotor system. J Aerosp Eng. 2015;28(2):1–8. doi:10.1061/(asce)as.1943-5525.0000386.
3. Jegadeeshwaran R, Sugumaran V. Fault diagnosis of automobile hydraulic brake system using statistical features and support vector machines. Mech Syst Signal Process. 2015;52–53:436–46. doi:10.1016/j.ymssp.2014.08.007.
4. Rapur J, Tiwari R. On-line time domain vibration and current signals based multi-fault diagnosis of centrifugal pumps using support vector machines. J Nondestruct Eval. 2018;38(1):1–18. doi:10.1007/s10921-018-0544-7.

5. Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK. Applications of machine learning to machine fault diagnosis: a review and roadmap. Mech Syst Signal Process. 2020;138:106587–625. doi:10.1016/j.ymssp.2019.106587.
6. Xu T, Yin Z, Cai D, Zheng D. Fault diagnosis for rotating machinery based on local mean decomposition morphology filtering and least square support vector machine. J Intell Fuzzy Syst. 2017;32(3):2061–70. doi:10.3233/jifs-161676.
7. Chen F, Tang B, Song T, Li L. Multi-fault diagnosis study on roller bearing based on multi-kernel support vector machine with chaotic particle swarm optimization. Measurement. 2014;47(1):576–90. doi:10.1016/j.measurement.2013.08.021.
8. Zhu X, Xiong J, Liang Q. Fault diagnosis of rotation machinery based on support vector machine optimized by quantum genetic algorithm. IEEE Access. 2018;6:33583–8. doi:10.1109/access.2018.2789933.
9. Berredjem T, Benidir M. Bearing faults diagnosis using fuzzy expert system relying on an improved range overlaps and similarity method. Expert Syst Appl. 2018;108(3):134–42. doi:10.1016/j.eswa.2018.04.025.
10. Wu J, Chiang P, Chang Y, Shiao Y. An expert system for fault diagnosis in internal combustion engines using probability neural network. Expert Syst Appl. 2008;34(4):2704–13. doi:10.1016/j.eswa.2007.05.010.
11. Chen Z, Li C, Sanchez RV. Gearbox fault identification and classification with convolutional neural networks. Shock Vib. 2015;2015(2):1–10. doi:10.1155/2015/390134.
12. Peng D, Wang H, Liu Z, Zhang W, Zuo M, Chen J. Multibranch and multiscale CNN for fault diagnosis of wheelset bearings under strong noise and variable load condition. IEEE Trans Ind Inform. 2020;16(7):4949–60. doi:10.1109/tii.2020.2967557.
13. Zhao X, Yao J, Deng W, Ding P, Ding Y, Liu Z. Intelligent fault diagnosis of gearbox under variable working conditions with adaptive intraclass and interclass convolutional neural network. IEEE Trans Neural Netw Learn Syst. 2023;34(9):6339–53. doi:10.1109/tnnls.2021.3135877.
14. Nie X, Xie G. A novel normalized recurrent neural network for fault diagnosis with noisy labels. J Intell Manuf. 2020;32(5):1271–88. doi:10.1007/s10845-020-01608-8.
15. He D, Lao Z, Jin Z, He C, Shan S, Miao J. Train bearing fault diagnosis based on multi-sensor data fusion and dual-scale residual network. Nonlinear Dyn. 2023;111(16):14901–24. doi:10.1007/s11071-023-08638-w.
16. Qin H, Pan J, Li J, Huang F. Fault diagnosis method of rolling bearing based on CBAM_ResNet and ACON activation function. Appl Sci. 2023;13(13):7593. doi:10.3390/app13137593.
17. Wang Z, Chen J, Wang C, Peng C, Xuan J, Zuo M. CNC-VLM: an RLHF-optimized industrial large vision-language model with multimodal learning for imbalanced CNC fault detection. Mech Syst Signal Process. 2026;245:113838.
18. Zhang Q, Zhu S. Visual interpretability for deep learning: a survey. Front Inf Technol Electron Eng. 2018;19(1):27–39. doi:10.1631/fitee.1700808.
19. Li T, Zhao Z, Sun C, Cheng L, Chen X, Gao RX. Waveletkernelnet: an interpretable deep neural network for industrial intelligent diagnosis. IEEE Trans Syst Man Cybern Syst. 2022;52:2302–12.
20. Chen Q, Dong X, Tu G, Wang D, Cheng C, Peng Z. TFN: an interpretable neural network with time-frequency transform embedded for intelligent fault diagnosis. Mech Syst Signal Process. 2024;207:1–21.
21. Ravanelli M, Bengio Y. Interpretable convolutional filters with sincnet. arXiv:1811.09725. 2019.
22. Xu C, Wang Z, Jin Y, Nong W. An adaptive industrial large language model for mechanical fault diagnosis under variable operating conditions. Adv Eng Inform. 2026;74(1):104821. doi:10.1016/j.aei.2026.104821.
23. Michau G, Frusque G, Fink O. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. Proc Natl Acad Sci U S A. 2022;119(8):1–10. doi:10.1073/pnas.2106598119.
24. Wang D, Zhong J, Li C, Peng Z. Box-cox sparse measures: a new family of sparse measures constructed from kurtosis and negative entropy. Mech Syst Signal Process. 2021;160:107930–42.
25. Dozat T. Incorporating nesterov momentum into adam. [cited 2026 Jan 1]. Available from: http://cs229.stanford.edu/proj2015/054_report.pdf.
26. Purohit RH, Ichige K, Endo T, Nikaido Y, Suefusa K, Yohei K. MIMII dataset-sound dataset for malfunctioning industrial. arXiv:1909.09347. 2019.

27. Kilickaya S, Ahishali M, Celebioglu C, Sohrab F, Eren L, Turker I. Audio-based anomaly detection in industrial machines using deep one-class support vector data description. In: Proceedings of the 2025 IEEE Symposium on Computational Intelligence on Engineering/Cyber Physical Systems Companion (CIES Companion); 2025 Mar 17–20; Trondheim, Norway.
28. Abbasi S, Famouri M, Shafiee MJ, Wong A. Outliernets: highly compact deep autoencoder network architectures for on-device acoustic anomaly detection. Sensors. 2021;21(14):1–12.