Eksploatacja i Niezawodnosc - Maintenance and Reliability
Volume 29 (2027), Issue 1
https://ein.org.pl
The Polish Maintenance Society

Rolling bearing fault diagnosis method based on RSBU-MSCNN under strong background noise

Zhe Zhang[1], Qi Guo[2], Dong Liu[1], Ziyang Chen[1], Zhiying Qin[1], Yuejing Zhao[1], Yanping Cui[1], Zhe Wu[1,*]

[1] School of Mechanical Engineering, Hebei University of Science and Technology, Shijiazhuang, 050018, P.R., China
[2] Center of Research and Development, KINGYEE(Beijing) Co., LTD., Beijing 100024, China
* Corresponding Author: wuzhe@hebust.edu.cn

Received: 25 December 2026
Revised: 17 March 2026
Accepted: 21 April 2026
Online: 3 July 2026

Abstract

Rolling bearings are key elements in rotating machinery, and reliable fault diagnosis is crucial for condition monitoring and maintenance decisions. Under strong background noise, vibration signals are easily distorted, which degrades conventional CNN-based diagnosis. To address this issue, an RSBU-MSCNN-based approach is proposed. First, Gaussian white noise with different signal-to-noise ratios is added to original signals to simulate industrial noise, and one-dimensional vibration signals are transformed into two-dimensional time–frequency representations using CWT. Then, a residual shrinkage module with a soft-threshold function is introduced for adaptive denoising and redundant noise suppression, while multi-channel, multi-scale convolutions enhance robust feature extraction across different receptive fields. Finally, faults are classified using fully connected layers. Experiments on multiple datasets show high accuracy under strong noise, confirming the robustness and applicability of the proposed method for industrial maintenance.

Keywords: *rolling bearings, strong background noise, feature extraction, fault diagnosis*

Article citation:
Zhang Z, Guo Q, Liu D, Chen Z, Qin Z, Zhao Y, Cui Y, Wu Z, Rolling bearing fault diagnosis method based on RSBU-MSCNN under strong background noise, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2027: 29(1) *http://doi.org/10.17531/ein/220967*

Highlights

- An end-to-end RSBU-MSCNN is proposed for bearing fault diagnosis under strong noise conditions.
- A 2D residual shrinkage unit is introduced for adaptive noise suppression.
- A multi-scale CNN with 3×3–7×7 kernels is used for feature fusion.
- Superior accuracy and robustness are achieved under low-SNR environments.

1. Introduction

Rolling bearings, as essential components of rotating machinery transmission devices, have a direct impact on the integrity and safety of mechanical systems [1–4]. Due to the frequent exposure of rolling bearings to harsh environments and complex operating conditions, they are highly prone to failures during operation [5,6]. Some of these failures not only reduce the operational efficiency of equipment but can also, in severe cases, lead to accidents, resulting in significant economic losses and safety risks [6,7].Therefore, fault diagnosis of rolling bearings has become a major focus in both research and industry, receiving extensive exploration and investigation.

Due to the complexity of rolling bearing fault signals, traditional signal processing methods often struggle to accurately extract fault features, as these features exhibit nonlinearity and irregularity [8,9], leading to suboptimal diagnostic accuracy. In recent years, with the rapid development of deep learning [10–14], its powerful automatic feature learning capabilities have enabled the direct utilization of these capabilities, avoiding the limitations caused by the absence of domain knowledge for manual feature extraction [15]. As a result, deep learning has gradually become a research hotspot and has introduced new methods for rolling bearing fault diagnosis. Among these, CNNs have gained widespread attention due to their strong feature extraction capabilities. Dao et al. [16] proposed a fault diagnosis method combining CNN and LSTM, and used a Bayesian optimization algorithm to optimize the model's hyperparameters. This method addresses the high-dimensional hyperparameter optimization issue of

traditional models, improving both diagnostic accuracy and stability. Guo et al. [17] introduced a bearing fault diagnosis method based on motor speed signals and CNN, enhancing diagnostic accuracy by incorporating frequency-domain features. Wang et al. [18] proposed a bearing fault diagnosis method combining SDP representation and convolutional neural networks with a channel attention mechanism. Fault diagnosis visualization was achieved through SDP images, followed by deep feature extraction using a CNN combined with an SE attention mechanism. Gu et al. [19] presented a new rolling bearing fault diagnosis method. This method decomposes the signal using VMD, generates time-frequency images through CWT, then extracts features using CNN, and finally performs fault classification through SVM, achieving a diagnostic accuracy of up to 99.9%, which validates the superiority of their method. Sinitsin et al. [20] proposed a bearing fault diagnosis method combining hybrid input and a hybrid CNN-MLP model. The method extracts deep features from HHT images using CNN, while MLP handles numerical features from angular acceleration signals. This method achieved superior diagnostic accuracy across multiple datasets.

Although the aforementioned methods can effectively diagnose rolling bearing faults and achieve good classification results, traditional CNN models exhibit a decline in feature extraction performance in noisy environments. The interference from noise causes important features in the signal to be overwhelmed, significantly affecting the diagnostic accuracy of the model [21]. To address this issue, researchers have proposed various improvement methods. Wang et al. [22] proposed a bearing fault diagnosis method based on vibration-acoustic data fusion and 1D-CNN. By fusing accelerometer and microphone sensor data, this method significantly improved diagnostic accuracy in different noise environments. Han et al. [23] introduced a multi-task bearing fault diagnosis method combining CNN and Transformer. This method shares global feature information and utilizes a multi-task learning framework to simultaneously address multiple fault tasks, effectively enhancing diagnostic performance in early noise interference environments. Wang et al. [24] presented a bearing fault diagnosis method combining CNN and joint learning (JL-CNN), which integrates fault diagnosis and signal denoising tasks. By sharing global feature information, JL-CNN significantly improved both fault diagnosis and denoising performance in high-noise environments. Han et al. [25] proposed a fault diagnosis method combining Bi-LSTM and capsule networks (BLC-CNN). This method uses Bi-LSTM for denoising and fusion, CNN for feature extraction, and capsule networks to improve diagnostic accuracy under noise and varying speed conditions. Zhao et al. [26] introduced an improved fault diagnosis method combining CEEMDAN and convolutional neural networks. This method uses CEEMDAN for adaptive denoising and decomposition of signals, and combines PCA with fractal dimensions to jointly select and reconstruct the optimal feature sub-signals, which are then input into CNN for feature extraction and fault classification, significantly improving diagnostic accuracy under multi-fault and multi-noise conditions.

In recent years, the development of noise-robust fault diagnosis methods has attracted increasing attention, particularly for rolling bearings operating under strong background interference. Peng et al. [27] fused HPO-VMD with an improved wavelet threshold strategy to enhance vibration signal denoising and improve SNR in complex noise environments. Qiu et al. [28] proposed a multimodal fusion diagnosis framework integrating multiscale denoising autoencoders and dual-branch feature fusion, achieving stable performance even under severe noise interference. Du et al. [29] combined CEEMDAN-based decomposition and phase space reconstruction (PSR) with an adaptive deep echo state network and BiLSTM to improve bearing fault diagnosis in noisy environments, demonstrating strong robustness even under very low SNR conditions. Li et al. [30] proposed a strong-noise bearing diagnosis method by combining SVD denoising, VMD decomposition, and DBO-optimized MCKD for weak fault enhancement. Xiao et al. [31] introduced a noise-robust CNN architecture by integrating global attention and gated convolutional kernels into a multi-scale separable CNN, significantly improving bearing fault detection performance under various noise levels.

Despite the improvements achieved by the aforementioned methods in bearing fault diagnosis under noisy environments, several issues remain: (1) many noise-robust approaches still rely on multi-stage preprocessing procedures (e.g., signal decomposition and thresholding), which increase

implementation complexity and make global optimization difficult; (2) some methods simply cascade or independently design signal enhancement and feature extraction modules, lacking a unified end-to-end collaborative optimization mechanism, thereby limiting the effective alignment between denoising and feature representation; (3) most existing approaches are evaluated under limited noise types or specific signal-to-noise ratio (SNR) settings, and their robustness and generalization ability may degrade when facing non-stationary, structured, or non-Gaussian industrial noise; and (4) conventional CNN backbones with single-scale receptive fields are insufficient to capture complementary fault patterns at different time–frequency resolutions, particularly when weak fault signatures are severely submerged by strong background noise. Therefore, it is crucial to develop an end-to-end diagnostic framework that can jointly perform adaptive denoising and multi-scale robust feature extraction under strong noise interference.

To address the above issues, this paper proposes an end-to-end collaborative optimization framework for rolling bearing fault diagnosis under strong noise conditions. Rather than treating denoising and feature extraction as two independently designed stages, the proposed method tightly couples adaptive shrinkage learning with multi-scale representation modeling within a unified diagnostic pipeline. The primary contributions of this study are summarized as follows:

(1) A unified end-to-end framework is developed for rolling bearing fault diagnosis under strong noise conditions, in which adaptive shrinkage and multi-scale feature extraction are collaboratively optimized rather than independently designed. This framework reduces the reliance on complicated multi-stage preprocessing and improves the consistency between noise suppression and discriminative feature learning.

(2) An RSBU-MSCNN architecture is constructed by extending the conventional residual shrinkage unit to a spatially aware 2D-RSBU and integrating parallel convolution branches with kernel sizes of 3×3, 5×5, and 7×7. This design enables the network to simultaneously capture weak local fault details and broader contextual structures, thereby enhancing robust feature representation under severe noise interference.

The remainder of this paper is organized as follows. Section 2 introduces the relevant basic theories. Section 3 presents the proposed method and the framework of the RSBU-MSCNN. Section 4 outlines the overall technical approach of this study. Section 5 describes the experimental data of rolling bearings and validates the superiority of the proposed model compared to other models. Section 6 concludes the paper and discusses future work.

2. Related theory

2.1. Continuous wavelet transform

Typically, assume that the input signal $x(t) \in L^2(R)$, and the basic wavelet function $\psi(t) \in L^2(R)$. If L^2（R）represents the space of square-integrable real-valued functions, then the Continuous Wavelet Transform (CWT)[32-34] of the signal can be expressed as follows:

$$\mathrm{WT}_x(a,\tau) = \frac{1}{\sqrt{a}} \int x(t)\psi^*\left(\frac{t-\tau}{a}\right) \mathrm{d}t = \langle (x(t), \psi_{a\tau}(t)\rangle \quad (1)$$

Where a and τ represent the scale and translation parameters in the wavelet transform, respectively, and $\psi*(t)$为 denotes the complex conjugate of $\psi(t)$,while ⟨x,y⟩ represents the inner product.

$$\langle x(t), y(t)\rangle = \int x(t)y^*(t)\mathrm{d}t \quad (2)$$

$$\psi_{a\tau}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right) \quad (3)$$

The translation and scale scaling of the wavelet basis in the wavelet transform are represented by (3). The frequency-domain representation of the wavelet transform is expressed as follows:

$$\mathrm{WT}_x(a,\tau) = \frac{1}{\sqrt{a}} \int \hat{x}(\omega)\hat{\psi}^*(a,\omega)e^{jwt}d\omega \quad (4)$$

Where $\hat{x}(\omega)$ represents the Fourier transform of the signal $x(t)$, $\hat{\psi}^*(\omega)$ represents the Fourier transform of the basic wavelet function $\psi(t)$,and is the complex conjugate of $\psi(\omega)$.

The input signal $x(t)$ is decomposed and reconstructed using wavelet transform, which decomposes the signal into a series of basic wavelet functions $\psi(t)$.The basic wavelet is a bandpass function with relatively concentrated amplitude-frequency characteristics, allowing the wavelet transform to represent the local characteristics of the signal in the frequency domain $\hat{x}(\omega)$.

2.2. Convolutional neural networks

CNN [35–37] typically consist of several key components, with the most critical being the convolutional layer, pooling layer, and fully connected layer. These components work together to perform more complex image recognition and processing tasks, as shown in Fig. 1.

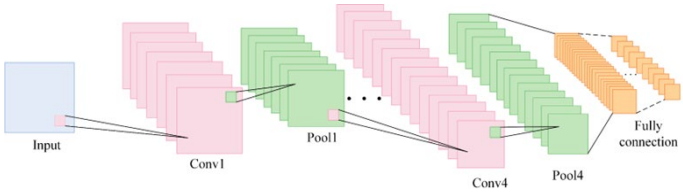


Figure 1. Basic structure of convolutional neural networks.

The convolutional layer is fundamental in CNNs. Multiple kernels convolve the input to produce feature maps, each capturing different patterns. Weight sharing enables rich feature learning with fewer parameters than fully connected layers, improving generalization. The convolution operation is given as follows.

$$g_k = \omega^k \cdot x_i + b^k; k = 1,2,\cdots,q; i = 1,2\cdots,n \quad (5)$$

Where x_i represents the i-th input data, ω^k is the parameter of the k-th convolutional kernel in this convolutional layer, b^k is the corresponding bias parameter, and g_k is the feature extracted by the k-th convolutional kernel from the i-th sample.

Max pooling downsamples feature maps after convolution to reduce dimensionality, parameters, and computation while preserving salient features. It outputs the maximum value in a fixed-size region, formulated as follows:

$$y_i^l = max\text{pooling}(x_i^{l-1}, S_{\text{scale}}, S_{\text{stride}}) \quad (6)$$

Where y_i^l represents the output of the i-th neuron in the current layer, maxpooling $(\cdot)$ is a downsampling function that selects the maximum value within a certain range, S_{scale} is the pooling range, S_{stride} is the stride size of the pooling operation.

The fully connected layer maps extracted features to class labels. The final layer uses Softmax to convert outputs into class probabilities, given by:

$$y^l = f(\omega^l x^{l-1} + b^l) \quad (7)$$

Where l denotes the layer number of the network, y^l represents the output of the fully connected layer, x^{l-1} represents the input to the fully connected layer, ω^l denotes the weight coefficients, b^l represents the bias.

2.3. Deep residual shrinkage network

The Deep Residual Shrinkage Network (DRSN)[38, 39] is built from Residual Shrinkage Building Units (RSBU) and combines soft-thresholding with attention to learn features under strong noise or high redundancy. A threshold-learning subnetwork is embedded in the residual structure to adaptively suppress noise and redundant information. Soft-thresholding removes features below a threshold and shrinks larger magnitudes toward zero, as defined below.

$$y = \begin{cases} x - \tau, & x > \tau \\ 0, & -\tau \le x \le \tau \\ x + \tau, & x < \tau \end{cases} \quad (8)$$

Where x is the input feature, y is the output feature, and τ is the threshold, a positive parameter.

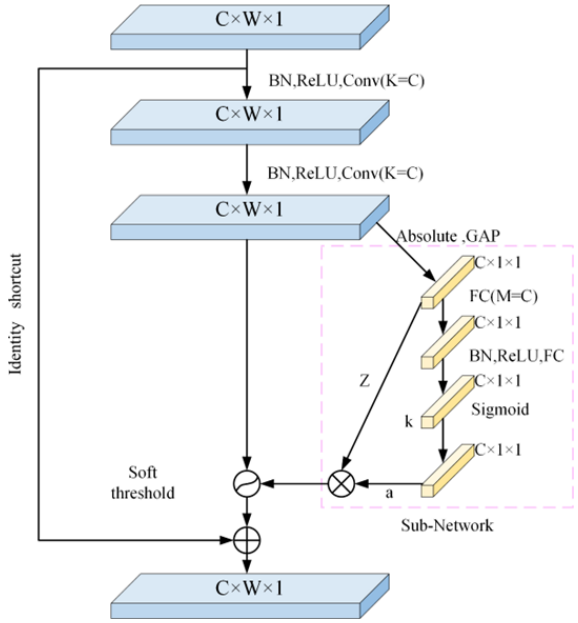


Figure 2. Structure of the RSBU Module.

Figure 2 shows the RSBU, which embeds a subnetwork and a soft-thresholding module. The subnetwork computes the absolute input features and applies global average pooling to obtain Z. In a parallel branch, two fully connected layers and a sigmoid function produce the scaling parameter a ∈ [0,1].

$$a_c = \frac{1}{1+e^{-k_c}} \quad (9)$$

Where k_c is the feature of the c-th channel output from the last fully connected layer, and a_c is the scaling parameter for the c-th channel. The threshold is calculated as follows.

$$\tau_c = a_c \cdot Z \quad (10)$$

Where τ_c is the threshold of the c-th channel of the feature map, and Z is the output feature of the convolutional layer.

Since the noise level varies across samples, RSBU can automatically learn adaptive thresholds from the input data. Each sample is assigned an independent threshold according to its noise content, thereby suppressing noise-related information.

3. Proposed method

3.1. Mechanism analysis of strong noise interference

In practical industrial scenarios, bearing vibration signals are often affected by strong environmental noise, especially under low signal-to-noise ratio (SNR) conditions. When the SNR decreases significantly, noise not only contaminates the raw signal but also fundamentally influences the reliability of feature extraction and model learning. From the perspective of feature representation, the time-frequency images obtained through Continuous Wavelet Transform (CWT) normally exhibit localized energy concentrations corresponding to periodic fault impacts. However, under strong noise interference, noise energy spreads across the time-frequency plane, which submerges weak fault-related components and blurs structural boundaries. As a result, the contrast between different fault categories decreases, and the extracted features from different classes become less distinguishable. Conventional CNNs mainly rely on local convolution operations to capture spatial correlations. When discriminative structures are partially masked by noise, convolutional filters may respond to irrelevant noise patterns, leading to feature confusion and reduced classification accuracy. From the perspective of model learning, strong noise also affects the stability of deep neural network training. Noise-dominant inputs may produce unstable feature activations in early layers, which propagate through the network and influence gradient updates during backpropagation. This may slow convergence or reduce generalization ability under low-SNR conditions. In addition, channel attention or global pooling mechanisms depend on statistical information extracted from feature maps. When these statistics are distorted by noise, the model may assign inappropriate importance weights to noise-dominated regions instead of fault-sensitive areas.

Therefore, strong noise introduces a dual challenge: (1) distortion and submergence of discriminative time-frequency features, and (2) instability in feature extraction and weighting during model training. To achieve reliable fault diagnosis under such conditions, it is necessary to integrate adaptive noise suppression mechanisms and robust multi-scale feature extraction strategies into the network architecture.

Based on the above mechanism analysis, the proposed RSBU-MSCNN framework is developed to simultaneously perform adaptive noise suppression and robust multi-scale feature extraction under low-SNR conditions. The detailed structure of each component is described in the following subsections.

3.2. 2D-RSBU module

This paper uses bearing vibration signals as input and converts one-dimensional time-domain signals into two-dimensional time–frequency images via CWT. Since the original RSBU is designed for one-dimensional time-series processing, it cannot be directly applied to two-dimensional image inputs. Therefore, the RSBU is adapted by replacing one-dimensional convolutions with two-dimensional convolutions to match the time–frequency representation, and the modified unit is termed 2D-RSBU. The structure of 2D-RSBU is shown in Fig. 3. It retains residual connections and soft-threshold-based adaptive feature selection, while improving spatial perception of local image details to enhance fault feature extraction. After ReLU activation, the output is fed to the subsequent network layers, which alleviates vanishing gradients and accelerates convergence. Overall, 2D-RSBU is more suitable for time–frequency images and improves feature extraction and classification for bearing fault diagnosis.

In addition, the threshold-learning subnetwork embedded in each 2D-RSBU is trained end-to-end together with the entire RSBU-MSCNN using the final classification loss only. Specifically, no auxiliary loss is introduced to supervise threshold learning separately. During backpropagation, the gradients of the classification objective are propagated through the soft-thresholding operation and the threshold-generation branch, enabling the model to automatically learn sample-dependent and channel-dependent shrinkage intensities that are most beneficial for fault classification. It should be noted that the learned threshold does not directly represent the physical amplitude of noise in the raw vibration signal. Instead, it acts as a channel-wise suppression boundary in the latent feature space. A larger threshold indicates that more low-magnitude feature

responses in the corresponding channel are likely to be dominated by noise or irrelevant disturbances, and thus stronger shrinkage is required. Therefore, the learned threshold can be interpreted as a feature-space denoising intensity rather than a direct measurement of signal-domain noise power.

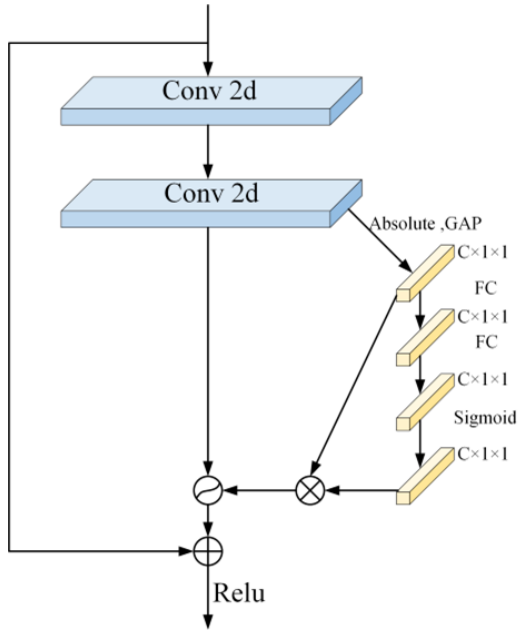


Figure 3. Structure of the 2D-RSBU Module.

3.3. Multi-scale feature fusion

CNN possess strong capabilities for feature extraction and model fitting, enabling them to learn useful information from input data. The basic structure of a CNN includes convolutional layers, activation layers, and pooling layers. Through these combinations, the number of feature channels gradually increases while the size of the feature maps decreases. Throughout this process, the size of the convolutional kernels directly influences the receptive field of the features and the depth of information extraction. By using convolutional kernels of different sizes, CNN can extract feature information at multiple scales, thereby enhancing their ability to represent features.

To better capture multi-scale fault information, this paper proposes a feature integration framework based on parallel convolutional branches with different receptive fields. Specifically, three branches with kernel sizes of 3×3, 5×5, and 7×7 are employed to extract features at different spatial scales from the CWT-based time-frequency images. Among them, the 3×3 branch focuses on fine-grained local textures and edge-like details, the 5×5 branch captures intermediate-scale structural variations, and the 7×7 branch provides broader contextual perception of the time-frequency distribution. Since bearing fault signatures in time-frequency representations may simultaneously involve local fluctuations, regional energy changes, and larger-scale contour patterns, the combination of these three representative kernel sizes enables the network to extract complementary features from local to broader contexts.

In addition, the three branches adopt the same channel dimensions to keep their representation capacities balanced, so that the performance differences can be attributed mainly to scale variation rather than to unequal parameter allocation. Therefore, although the three branches are symmetric in terms of channel numbers, the distinct receptive fields introduced by different kernel sizes enable them to learn features with different semantic emphases, which is beneficial for multi-scale complementarity. Each branch contains two convolutional layers and pooling operations for progressive feature extraction and down-sampling. In the final stage, the features extracted by the three kernel branches are concatenated using the torch.cat(·) operation to form the fused multi-scale representation.

$$F_c = \text{torch.cat}(f_3(x), f_5(x), f_7(x)) \tag{11}$$

In the formula, $f_3(x)$,$f_5(x)$ and $f_7(x)$ represent the outputs of the convolutional kernel channels at different scales, while F_c denotes the multi-scale fused features.

3.4. Theoretical analysis of adaptive shrinkage and multi-scale representation

This section analyzes the intrinsic mechanism by which the combination of 2D-RSBU and multi-scale convolution enhances robustness under strong noise conditions from three perspectives: noise suppression, multi-scale feature extraction, and synergistic gain. The objective of this analysis is to provide a mechanistic interpretation and approximate theoretical derivation rather than a strict statistical optimality proof.

1. Noise Suppression Mechanism of 2D-RSBU

Let the input feature map be denoted as $X \in R^{C\times W\times H}$, which can be conceptually decomposed into a fault-related structural component $S$and a noise-dominant component N.According to the soft-thresholding definition introduced in Section 2.3, the channel-wise output of 2D-RSBU satisfies:

$$y = \begin{cases} x - \tau_c, x > \tau_c \\ 0, \ |x| \le \tau_c \\ x + \tau_c, x < -\tau_c \end{cases} \tag{12}$$

Where τ_c is the adaptively learned threshold for the c-th channel. From (12), two important properties can be derived:

(1) When $|x| \leqslant \tau_c$, the output is forced to zero. Under low-SNR conditions, low-magnitude responses are more likely to be dominated by noise perturbations. Therefore, this mechanism suppresses noise propagation by attenuating weak activations before deeper feature extraction.

(2) For any input x, the following inequality holds:

$$|y - x| \leq \tau_c \tag{13}$$

Equation (13) indicates that although high-magnitude activations are shrunk, the induced distortion is upper-bounded by the adaptive threshold τ_c. Consequently, 2D-RSBU achieves a controllable trade-off between noise suppression and information preservation.

Furthermore, according to Eq. (10), the adaptive threshold is defined as $\tau_c = a_c Z_c$

where $Z_c$represents a channel-wise statistical measure (e.g., global average pooling of absolute feature responses), and $a_c \in [0,1]$is a learnable scaling coefficient. Under additive noise enhancement while the signal energy remains relatively stable, the expected value of $Z_c$increases with noise variance. As a result, $\tau_c$automatically increases, strengthening suppression of noise-dominant activations. When noise weakens, $\tau_c$decreases accordingly, preventing excessive shrinkage of informative fault features. This establishes an adaptive noise-tracking mechanism that dynamically adjusts purification intensity without requiring manually tuned thresholds.

2. Feature fusion of multi-scale branches

After RSBU-based purification, the feature representation typically becomes more structured and sparse. The multi-scale fusion representation can be expressed as:

$$F_c = \mathrm{Concat}(\Phi_3(y), \Phi_5(y), \Phi_7(y)) \tag{14}$$

Where $\Phi_k$denotes a convolution operator with kernel size k, and $Concat$ represents the fusion operation (e.g., concatenation or weighted aggregation).

Assume that discriminative fault textures possess an unknown characteristic receptive-field scale k^*. When the kernel size $k \approx k^*$, spatial correlation between the convolutional filter and the fault structure becomes stronger, resulting in higher effective extraction capability. However, since k^* varies across fault types and operating conditions, a single fixed-scale convolution may suffer from scale mismatch.

The role of parallel multi-scale branches is therefore: (1) To increase the probability that at least one branch approximates the optimal scale k^*; (2) To mitigate degradation caused by scale mismatch; (3) To provide complementary representations for subsequent classification. From a representational capacity perspective, multi-scale fusion enhances coverage of potential texture scales and improves robustness under strong noise, rather than relying on a single receptive field.

3. Collaborative Denoising Mechanism

The combination of 2D-RSBU and multi-scale convolution forms a sequential "purification–reconstruction" pipeline.

(1) Stage I: Threshold-Based Energy Compression

Let the original noise-dominant energy be denoted as E_n.The energy suppressed by RSBU can be expressed as ΔE_{rsbu}. Under strong noise conditions, this low-magnitude energy is more likely to be noise-dominated. Therefore, suppressing ΔE_{rsbu} reduces noise propagation into subsequent layers.

(2) Stage II: Spatial Correlation Smoothing

Multi-scale convolution—particularly branches with larger receptive fields—further reduces random noise variance through spatial correlation enhancement and local averaging effects. Let the variance reduction induced by this stage be denoted as ΔE_{msc}, whose magnitude depends on the smoothing scale and spatial aggregation properties.

Consequently, the effective output SNR can be approximately expressed as:

$$SNR_{out} \approx \frac{P_s}{E_n - (\Delta E_{rsub} + \Delta E_{msc})} \tag{15}$$

Where P_s denotes the power of the fault-related signal component. Equation (15) provides an approximate energy-level interpretation: adaptive magnitude gating and multi-scale spatial reconstruction cumulatively compress noise-dominant energy, thereby improving the effective SNR at the feature level.

3.5. Network architecture based on RSBU-MSCNN

To further enhance the network's feature extraction ability and noise robustness, this paper proposes a multi-scale convolutional neural network (MSCNN) integrated with the RSBU module, as shown in Fig. 4. The input to the network passes through two RSBU modules. The RSBU calculates the channel thresholds through adaptive pooling on the contraction path and applies a soft-threshold operation to remove noise

while preserving important features. The subsequent convolutional layers help expand the receptive field of the network to capture broader global information. The feature maps are then passed into the multi-scale framework for deeper multi-scale feature extraction. Information from different scales is then concatenated, ultimately forming a feature map that fuses multi-scale features. In the following pooling layers and flattening operation, the network retains rich features while reducing the spatial dimensions. Finally, classification is performed through fully connected layers. The specific parameters of the RSBU-MSCNN are shown in Table 1.

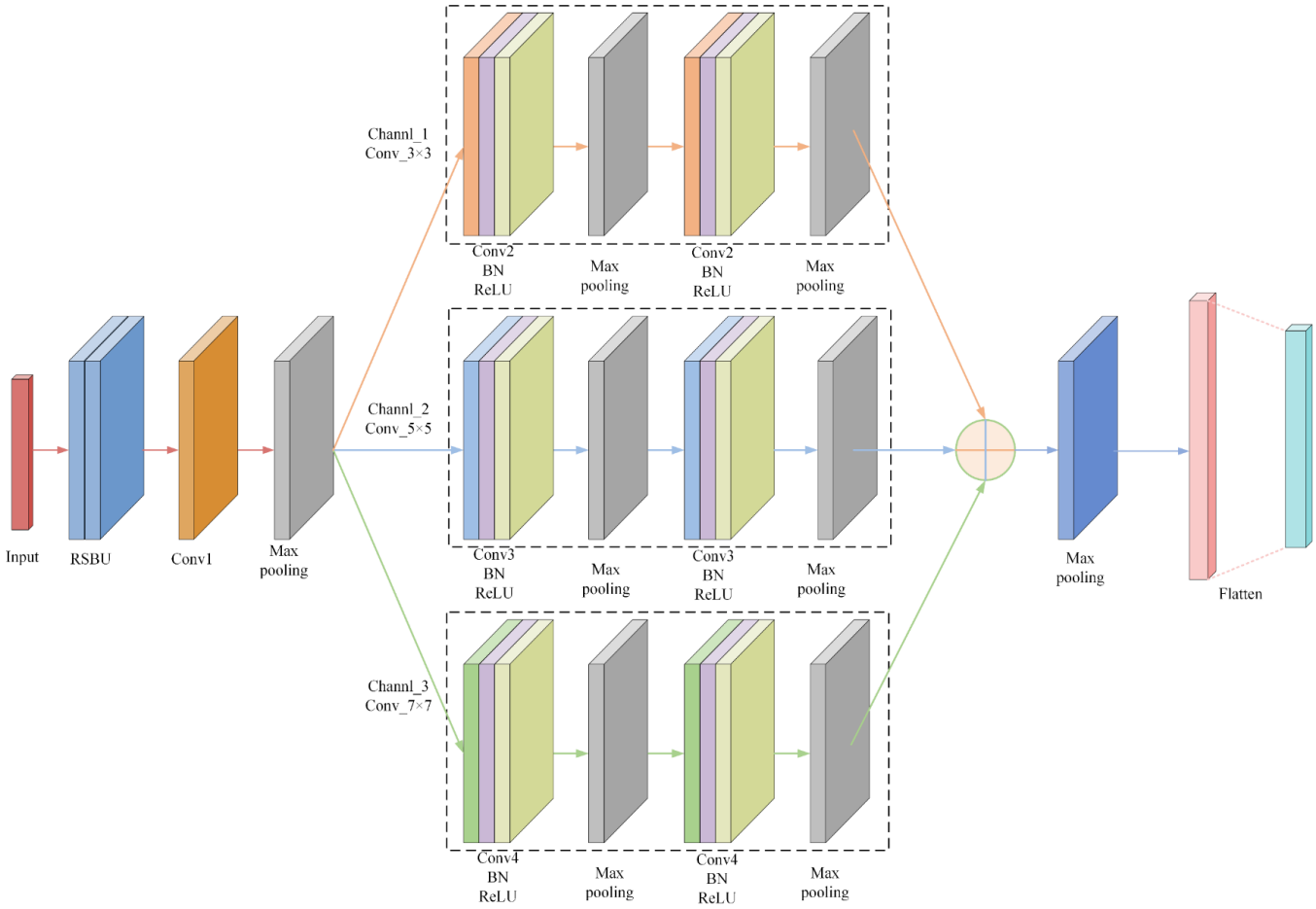


Figure 4. RSBU-MSCNN Network Architecture.

Table 1. RSBU-MSCNN model parameter settings.

Layer	Parameter Settings	Activation Function
Input	Size:(3,128,128)	-
RSBU1(Block)	RSBU Module (Conv1,Conv2,Shortcut,Gap,FC)	ReLU,Sigmoid
RSBU2(Block)	RSBU Module (Conv1,Conv2,Shortcut,Gap,FC)	ReLU,Sigmoid
Conv1	(64,64,32)	-
Pool1	Kernels:2	-
Branch3	Conv_1(64,64,3)	ReLU
	Conv_2(64,128,3)	ReLU
	MaxPool:2	-
Branch5	Conv_1(64,64,3)	ReLU
	Conv_2(64,128,3)	ReLU
	MaxPool:2	-
Branch5	Conv_1(64,64,3)	ReLU
	Conv_2(64,128,3)	ReLU
	MaxPool:2	-
Pool2	MaxPool:2	-
FC	Input:384 Output:num_classes	Softmax

4. Overall Structural Framework of the Proposed Method

The overall framework for bearing fault diagnosis described in this paper is shown in Fig. 5. This framework consists of three distinct steps: signal preprocessing, training using the network model, and result visualization.

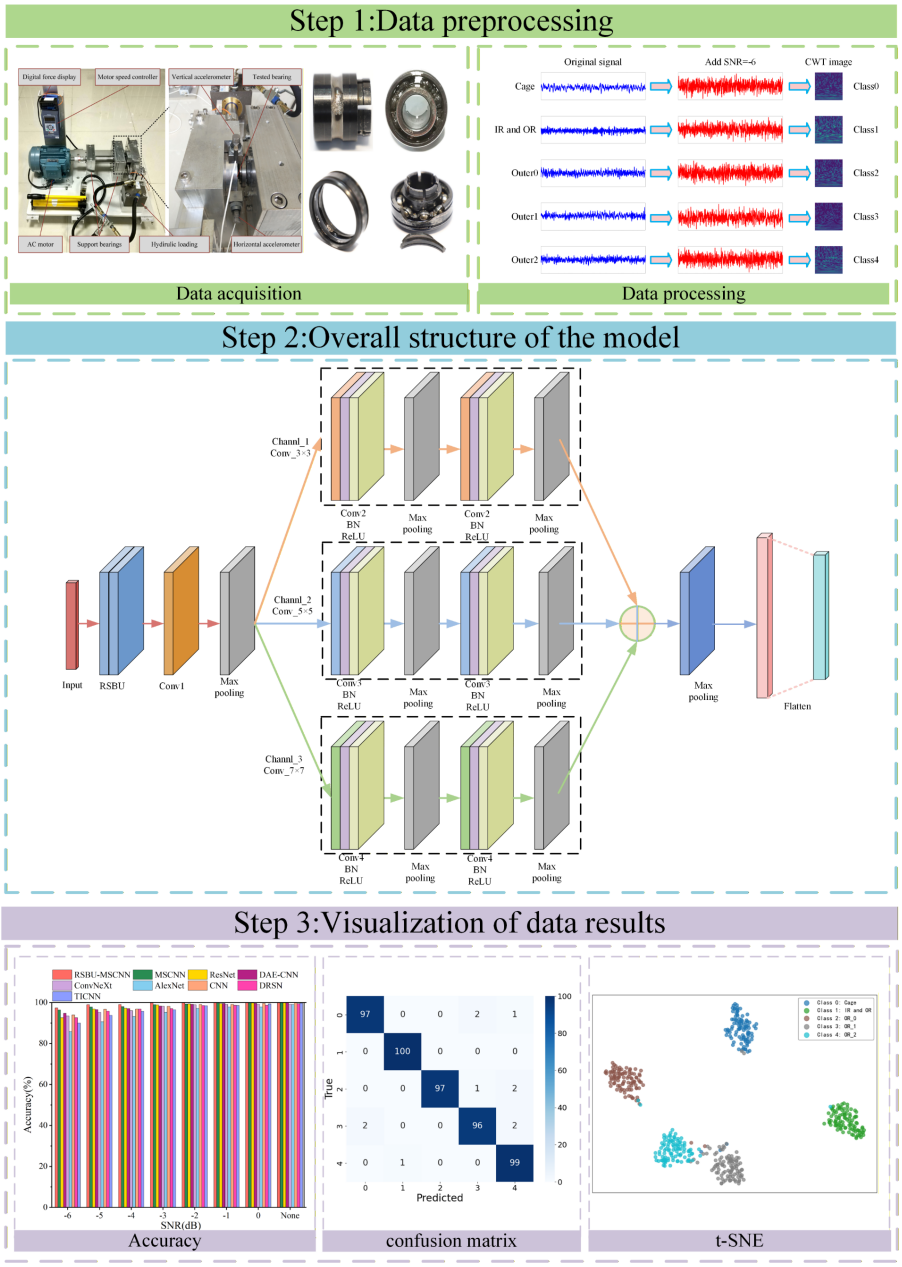


Figure 5. Technical Route Based on RSBU-MSCNN.

The overall framework for bearing fault diagnosis described in this paper is shown in Fig. 5. This framework consists of three distinct steps: signal preprocessing, training using the network model, and result visualization.

Step 1: Obtain the one-dimensional vibration signal of the faulty bearing by testing the bearing at the corresponding position on the test bench. Noise samples are generated by combining the data with noise, and the processed noisy samples are then transformed into a two-dimensional time-frequency image using CWT.

Step 2: Input the processed data into the network model for training. The network utilizes a soft-threshold denoising structure and multi-scale feature fusion to eliminate noise and enhance features, thereby extracting useful feature information.

Step 3: Based on experimental test results, use Confusion Matrix diagrams and t-SNE visualization plots, among others. The results indicate that the model exhibits strong noise robustness in high-noise environments.

5. Experimental verification

The method proposed in this paper is developed based on Python 3.10 within the Pytorch framework, using the Pycharm 2021 IDE editor. The main hardware parameters used in the experiments are: Intel Core2 i7 - 12800HX CPU, paired with 16 GB RAM and an RTX 4070 GPU. The training includes setting the training batch size to 100 and the learning rate to 0.0001.

5.1. CWRU rolling bearing dataset

5.1.1. Dataset introduction

The dataset used in this case study is the Case Western Reserve University (CWRU) bearing dataset, and the test bench is shown in Fig. 6. The drive-end bearing is an SKF6205, and single-point faults were introduced by electrical discharge machining. Fault categories include rolling element, inner race, and outer race faults, each with three fault diameters (0.1778, 0.3556, and 0.5334 mm). Vibration signals were collected at 12 kHz under 1772 r/min and 1 hp load. Using overlapping sampling, 5000 samples were generated, each containing 1024 points. The dataset was split into training and testing sets with an 8:2 ratio (Table 2).

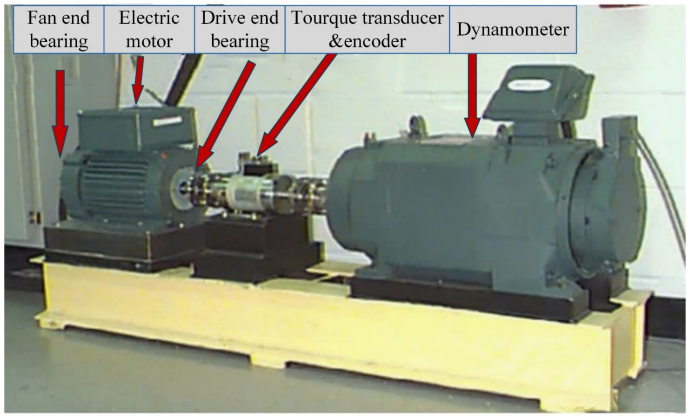


Figure 6. Physical Structure of the CWRU Test Bench.

Table 2. Detailed information of the CWRU experimental dataset.

Fault Label	Fault Type	Fault Size	Train Sample	Test Sample
0	Ball	0.1778	400	100
1	Ball	0.3556	400	100
2	Ball	0.5334	400	100
3	Inner	0.1778	400	100
4	Inner	0.3556	400	100
5	Inner	0.5334	400	100
6	Normal	0	400	100
7	Outer	0.1778	400	100
8	Outer	0.3556	400	100
9	Outer	0.5334	400	100

5.1.2. Raw signal processing

To investigate the impact of noise signals on model diagnostic performance, this study adds Gaussian white noise to the raw vibration signal. Gaussian white noise is chosen as it can simulate the randomness and spectral characteristics of real industrial environments. The signal-to-noise ratio (SNR) formula is shown as follows:

$$SNR = 10 \times log_{10}(\frac{P_S}{P_N}) \tag{16}$$

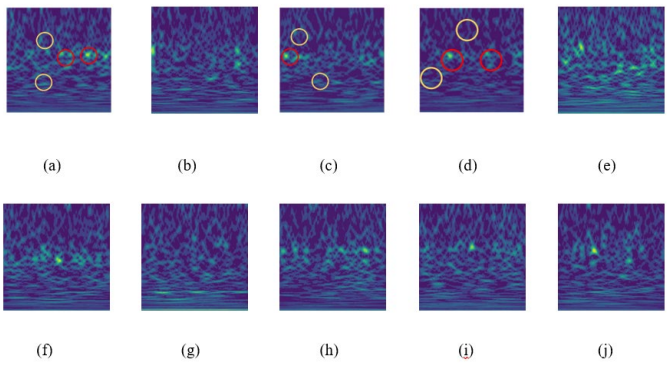


Figure 7. Wavelet time-frequency maps of 10 fault states: (a) B1 (b) B2 (c) B3 (d) IR1 (e) IR2 (f) IR3 (g) N (h) OR1 (i) OR2 (j) OR3.

Here, Ps and P_N denote the effective power of the signal and noise, respectively. The noisy vibration signal is then processed by continuous wavelet transform to generate a $3 \times 128 \times 128$ time–frequency map. At an SNR of −6 dB, the time–frequency maps presented in Fig. 7 exhibit significant noise interference, manifested as scattered and diffuse energy distributions across the time–frequency plane. In representative examples (e.g., Fig. 7(a)–(c)), the red circles highlight localized high-intensity regions corresponding to fault-related characteristic components, typically concentrated in the mid–high frequency bands. In contrast, the yellow circles denote diffuse and irregular background patterns primarily caused by strong noise contamination. The presence of strong noise in a scattered and diffused form leads to local feature distortion and blurring, thereby masking fault-related characteristics and increasing the difficulty of fault pattern recognition and feature extraction.

5.1.3. Evaluation metrics

To evaluate the noise robustness of the proposed model, accuracy, precision, recall, and F1-score [40] are used to assess diagnostic performance and generalization under noise.

Accuracy is defined as the proportion of correctly classified samples, given by:

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{17}$$

Precision and recall are defined as the ratio of true positive samples (TP) to all samples predicted as positive, and the ratio of TP to all actual positive samples, as shown in formulas (18) and (19), respectively.

$$precision = \frac{TP}{TP+FP} \tag{18}$$

$$recall = \frac{TP}{TP+FN} \tag{19}$$

The F1 score is the harmonic mean of precision and recall, calculated as the arithmetic mean divided by the geometric mean. It provides a comprehensive analysis by balancing the trade-off between precision and recall, as shown in formula (20).

$$F1 = \frac{2TP}{2TP+FP+FN} \tag{20}$$

In these formulas, TP represents the count of true positive samples, FN represents the count of false negative samples, FP represents the count of false positive samples, and TN represents the count of true negative samples.

5.1.4. Experimental results analysis

To verify the superiority of the proposed method, the RSBU-MSCNN model is compared with MSCNN and several representative deep learning approaches, including ResNet [41], ConvNeXt [42], AlexNet [43], a traditional CNN model, DAE-CNN, DRSN, and TICNN [44]. These comparison methods represent different network architectures and noise-robust fault diagnosis strategies that are widely used in intelligent fault diagnosis research.

MSCNN extracts multi-scale features through convolutional kernels with different receptive fields. ResNet introduces residual connections to facilitate deep network training and improve representation ability. ConvNeXt is a modern convolutional architecture that integrates several design principles inspired by transformer models. AlexNet is a classical deep CNN architecture that has been extensively used in early deep learning-based fault diagnosis studies. The traditional CNN serves as a baseline model without multi-scale or residual structures. The DAE-CNN model incorporates a denoising autoencoder (DAE)[45] module before the classification network to suppress noise interference and enhance feature robustness.

In addition, two representative noise-robust fault diagnosis models are included for comparison. The Deep Residual Shrinkage Network (DRSN) introduces adaptive soft-thresholding within residual blocks to suppress noise-dominated feature responses. TICNN improves robustness through temporal invariant feature learning, enabling the model to capture stable fault characteristics under noisy conditions. By incorporating these representative noise-robust models, the comparative experiments provide a more comprehensive evaluation of the effectiveness of the proposed RSBU-MSCNN under strong noise environments.

To evaluate model robustness under strong noise interference, experiments are conducted under a −6 dB noise environment, and the classification accuracy results are shown in Fig. 8.

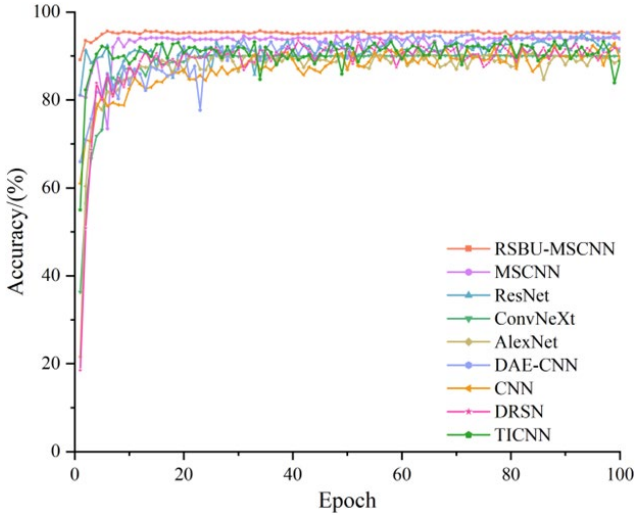


Figure 8. Accuracy curve at SNR = -6 dB.

As shown in Fig. 8, the accuracy of all models rises rapidly in the early training stage and gradually stabilizes with the increase of training epochs. Under the −6 dB noise condition, the proposed RSBU-MSCNN achieves the fastest convergence and consistently maintains the highest accuracy throughout the training process. MSCNN and ConvNeXt also exhibit relatively stable convergence behavior, but their final accuracies are still lower than that of the proposed model. ResNet, DRSN, and TICNN display more significant fluctuations during training, indicating that their feature extraction capabilities are still affected by strong noise interference. AlexNet, DAE-CNN, and the conventional CNN converge more slowly and remain at

relatively low accuracy levels. Overall, these results demonstrate that the proposed RSBU-MSCNN not only converges faster, but also possesses better stability and higher diagnostic accuracy under strong noise environments.

To further observe the superiority of the proposed model's noise robustness, Fig. 9 shows the accuracy variations of each model across noise levels ranging from -6 dB to 0 dB, as well as in the noise-free condition. The analysis indicates that as the noise level increases, the diagnostic accuracy of each method decreases. However, the proposed method demonstrates an average fault diagnosis accuracy of 95.20%, 96.40%, 97.66%, 98.23%, 98.63%, 99.04%, 99.43%, and 99.99% across different noise scenarios. In the same -6 dB noise environment, the RSBU-MSCNN achieves an average accuracy that is 3.06% higher than the 92.14% achieved by MSCNN, while MSCNN outperforms CNN by 4.92%, which proves the effectiveness of the feature fusion strategy and soft threshold denoising proposed in this study.

Furthermore, noise-robust models such as DRSN and TICNN also show improved performance compared with conventional CNN-based architectures. In particular, DRSN benefits from its adaptive soft-threshold shrinkage mechanism for suppressing noise-dominated feature responses, while TICNN improves robustness through temporal invariant feature learning. However, their overall performance remains inferior to that of RSBU-MSCNN under severe noise conditions.

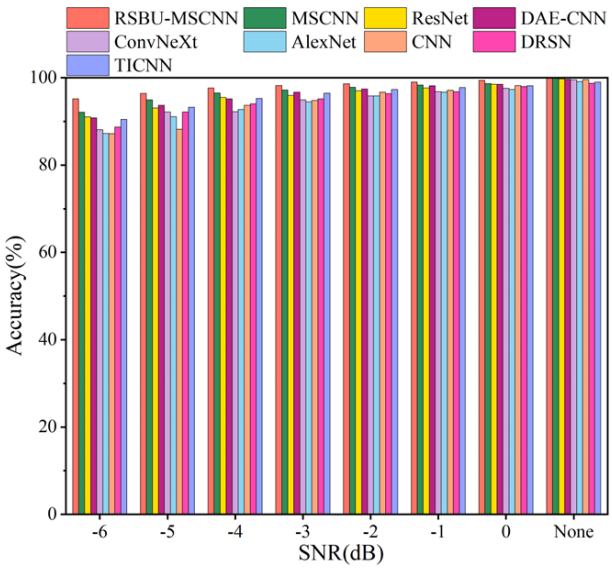
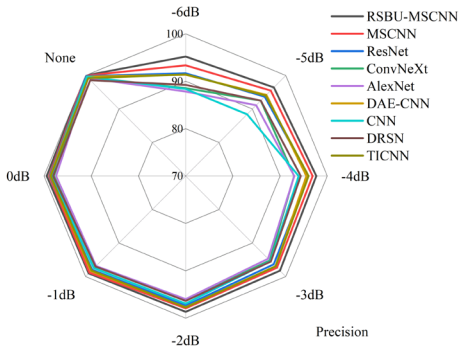
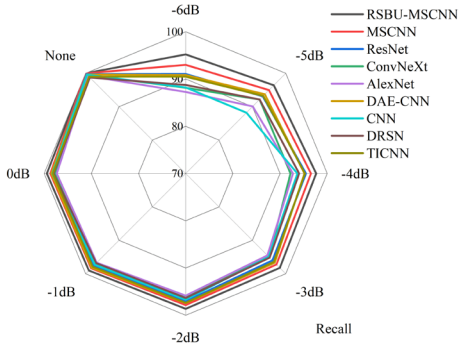


Figure 9. Bar chart of test accuracy for the CWRU dataset.

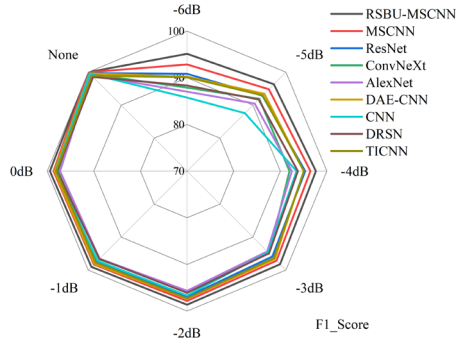
Figure 10 displays the experimental results of three evaluation metrics—Precision, Recall, and F1-Score—in radar chart form. Each axis represents the SNR, and the points along each axis indicate the values of the respective metrics. Points closer to the outer circle indicate higher diagnostic accuracy under the corresponding SNR. The nine methods are represented by different colors, and these points are connected by lines to form polygons. The shape and size of the polygons visually reflect the accuracy of the different metrics for the nine methods at various SNR levels.



(a)



(b)



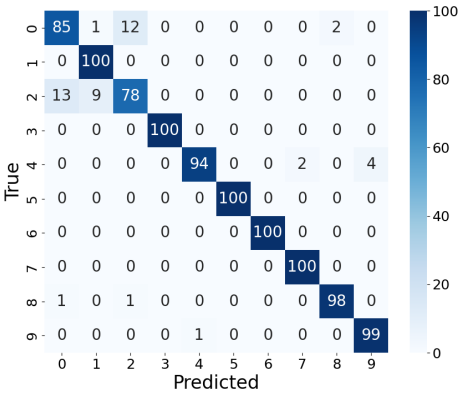
(c)

Figure 10. Radar chart of the three evaluation metrics under the CWRU dataset. (a) Precision (b) Recall (c) F1_Score.

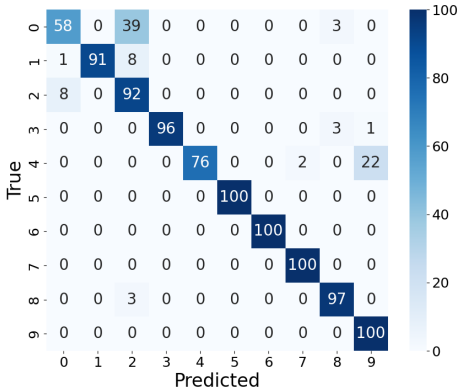
As can be seen from Figure 10, all methods are affected to varying degrees as the noise intensity increases. Obviously, CNN and AlexNet suffer from significant accuracy degradation under low SNR conditions. In contrast, noise-robust models

such as DRSN and TICNN exhibit improved stability compared with models including ConvNeXt. Nevertheless, the proposed RSBU-MSCNN consistently maintains the highest metric values across all SNR levels. Even under heavy noise conditions, its precision, recall, and F1-score remain above 95%, demonstrating stronger robustness against noise interference.

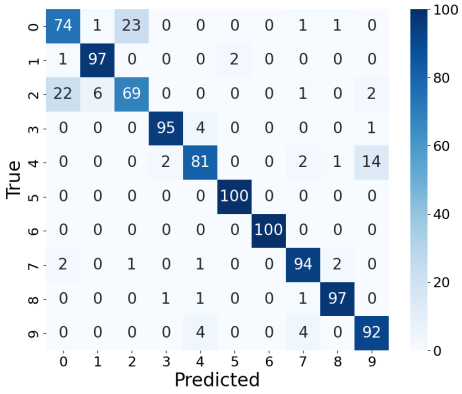
Figure 11 shows the confusion matrix for the proposed method and eight comparative models, illustrating the fault recognition ability for each fault type when SNR=−6 dB. The x-axis represents the predicted categories, while the y-axis represents the true categories. As shown in Fig. 11(a), the proposed RSBU-MSCNN achieves the highest concentration of correct predictions along the main diagonal, indicating superior classification accuracy under strong noise conditions. In contrast, the other eight models exhibit varying degrees of misclassification, with off-diagonal elements becoming more pronounced, reflecting reduced robustness to severe noise interference. Overall, the results demonstrate that the improved model proposed in this study maintains stronger noise resistance and achieves more reliable fault recognition performance compared with the methods.



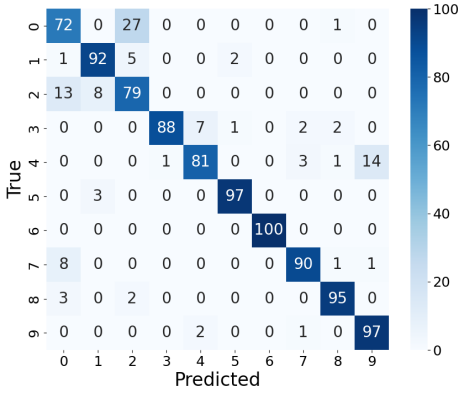
(a)



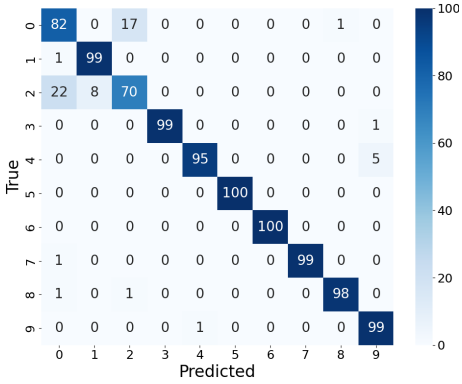
(b)



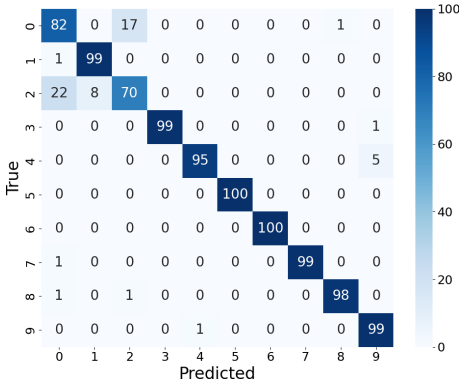
(c)



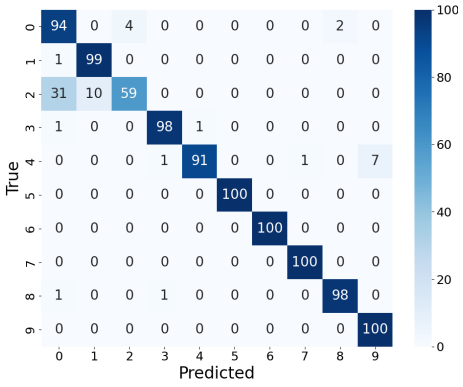
(d)



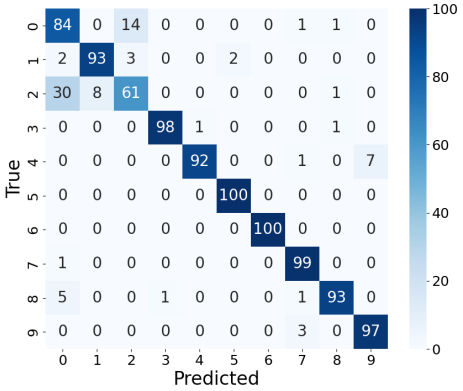
(e)



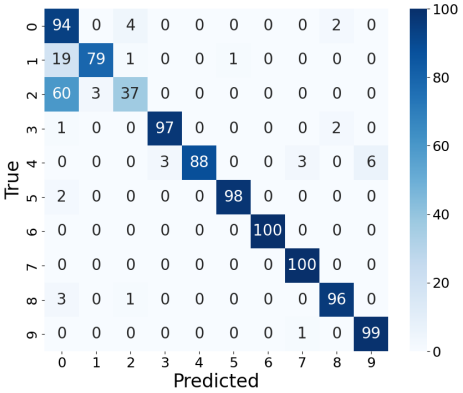
(f)



(g)



(h)

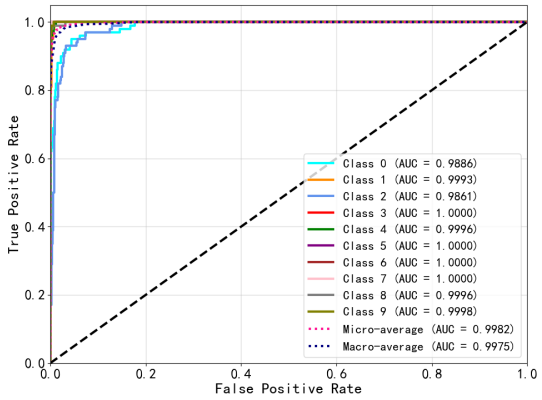


(h)

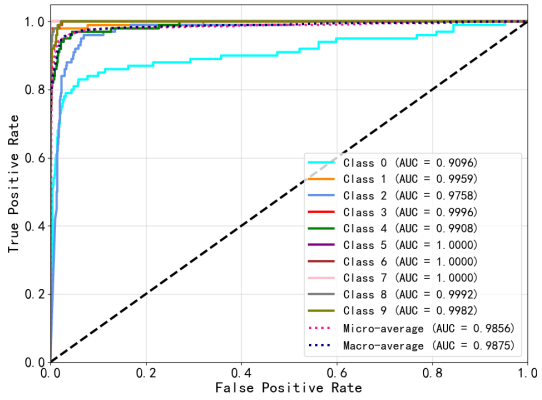
Figure 11. Classification Confusion Matrix for each method at SNR = -6 dB under the CWRU dataset. (a) RSBU-MSCNN (b) ResNet (c) ConvNeXt (d) AlexNet (e) MSCNN (f) CNN (g) DAE-CNN (h) DRSN (i) TICNN.

To comprehensively evaluate classification performance across fault categories under strong noise conditions, ROC curves of nine representative methods are presented in Fig. 12. The AUC values for each health state are shown in the lower right corner of each subfigure.

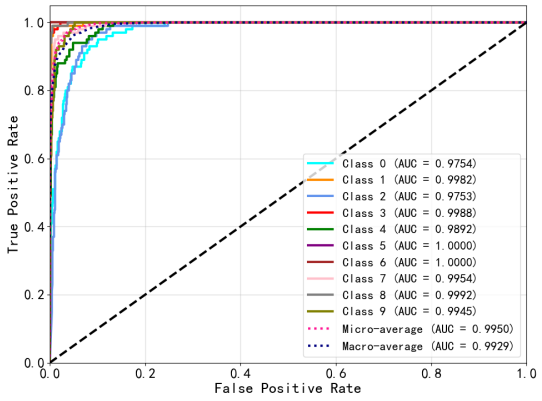
As the noise intensity increases, noticeable performance degradation can be observed in CNN and AlexNet, and to a certain extent in ResNet, indicating that conventional deep learning architectures are more sensitive to strong background interference. Although DAE-CNN introduces an autoencoder-based denoising mechanism and shows improved robustness compared with standard CNN, its AUC values still decline under severe low-SNR conditions. Similarly, models such as DRSN and TICNN incorporate dedicated mechanisms to improve robustness in noisy environments; however, their ROC curves still exhibit slight deviations from the upper-left corner for certain fault categories. ConvNeXt and MSCNN exhibit relatively stronger stability; however, their overall performance remains slightly inferior to that of the proposed RSBU-MSCNN in extremely noisy environments. In contrast, the proposed method consistently achieves higher AUC values across nearly all fault categories, with most ROC curves approaching the upper-left corner. Even under strong noise, RSBU-MSCNN maintains superior separability between classes, demonstrating enhanced discriminative capability and stronger noise robustness.



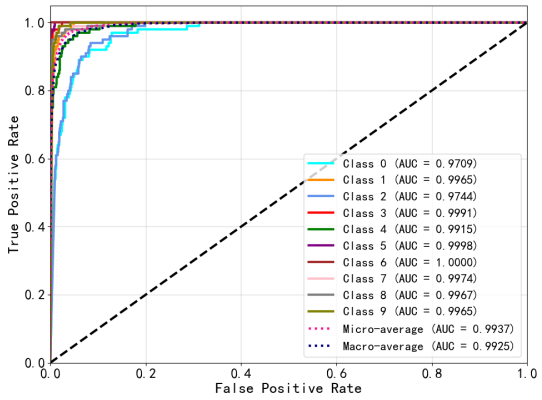
(a)



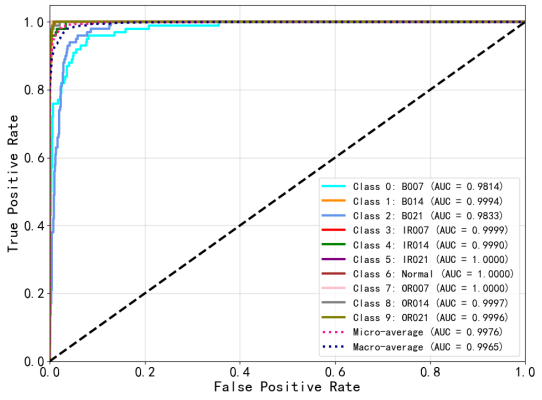
(b)



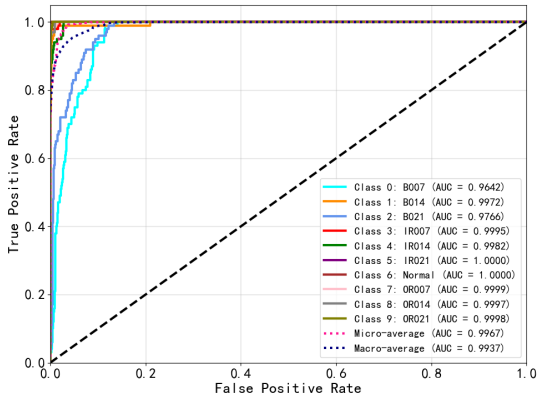
(c)



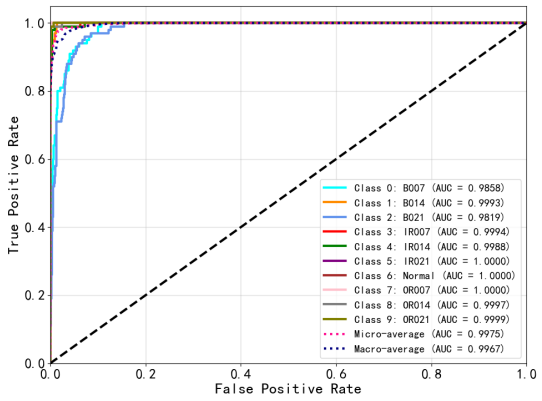
(d)



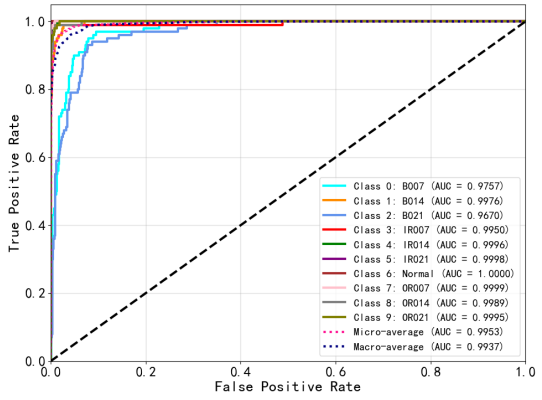
(e)



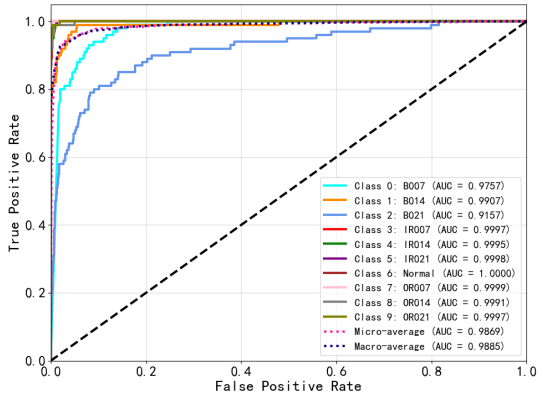
(f)



(g)



(h)



(i)

Figure 12. ROC curves of four methods under the CWRU dataset. (a) RSBU-MSCNN (b) ResNet (c) ConvNeXt (d) AlexNet (e) MSCNN (f) CNN (g) DAE-CNN (h) DRSN (i) TICNN.

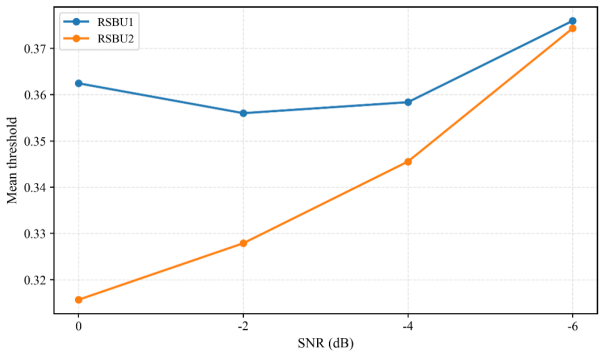
To further verify whether the threshold-learning branch indeed performs adaptive denoising, the learned thresholds generated by the two 2D-RSBU modules were analyzed under different SNR conditions. As shown in Fig. 13(a), the average thresholds of RSBU1 and RSBU2 both vary systematically with the noise level, and remain at relatively higher values under lower-SNR conditions. This indicates that the learned

thresholds are not fixed constants, but can be adaptively adjusted according to the input noise intensity, thereby enhancing feature suppression when stronger noise interference is present. The boxplots in Fig. 13(b) and Fig. 13(c) further illustrate the distribution patterns of the learned thresholds in RSBU1 and RSBU2, respectively. Under decreasing SNR conditions, the threshold distributions of both modules exhibit an overall upward shift, indicating that the adaptive shrinkage mechanism responds consistently to stronger noise contamination. In particular, the medians and interquartile ranges remain at relatively elevated levels under low-SNR conditions, suggesting that the threshold increase is not caused only by a few extreme channels, but reflects a global adjustment of shrinkage intensity across feature channels. Overall, these results demonstrate that the proposed threshold-learning mechanism performs data-dependent adaptive shrinkage rather than static threshold suppression. By dynamically modulating the feature suppression intensity according to the noise level, the 2D-RSBU modules provide effective support for robust fault feature extraction in noisy environments.

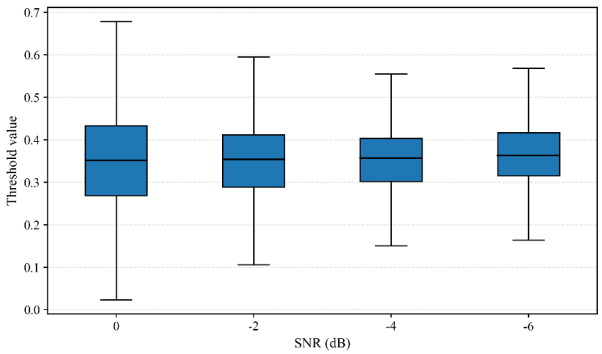
To intuitively illustrate how convolutional kernels with different receptive fields contribute to noise-robust classification, Grad-CAM is employed to visualize the activation regions of each scale-specific branch. The results are presented in Fig. 14. As shown, the 3×3 branch primarily focuses on localized discriminative regions and is capable of capturing subtle fault-related variations. Such small-scale kernels are highly sensitive to local transient feature, which are often critical for early fault detection. However, under severe noise contamination, these localized cues may be easily distorted or even completely obscured. In contrast, the 7×7 branch produces smoother and more spatially extensive activation responses, indicating that larger kernels emphasize global structural patterns and overall energy distributions. Compared with local details, global patterns are less sensitive to stochastic high-frequency noise and therefore provide more stable contextual evidence for classification.

The 5×5 branch exhibits intermediate activation characteristics between local details and global structures, effectively bridging the gap between the two scales. This intermediate-scale representation plays a stabilizing role when local transient cues are weakened or when global contextual information alone is insufficient.

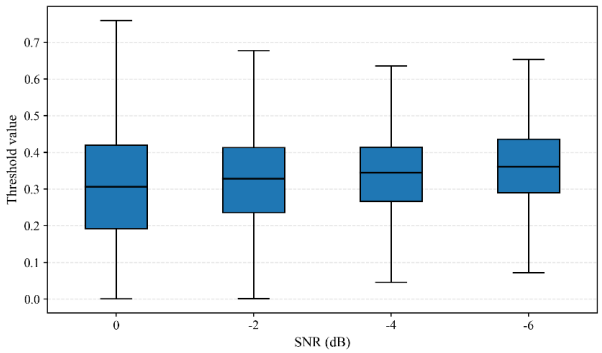
Therefore, the complementarity among multi-scale branches is crucial in the presence of noise: small kernels preserve fine-grained fault signatures when they are available, while large kernels provide robust global context when local features are degraded. By integrating scale-specific representations, the proposed architecture mitigates the risk of misclassification caused by noise-induced feature distortion and enhances overall discriminative reliability.



(a)

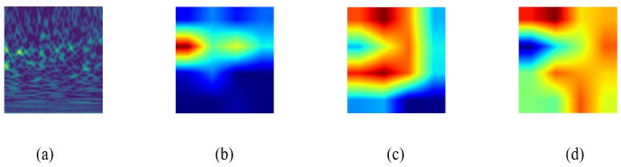


(b)



(c)

Figure 13. Learned threshold analysis under different SNR conditions. (a) Average learned thresholds of RSBU1 and RSBU2 (b) Threshold distribution of RSBU1 (c) Threshold distribution of RSBU2.

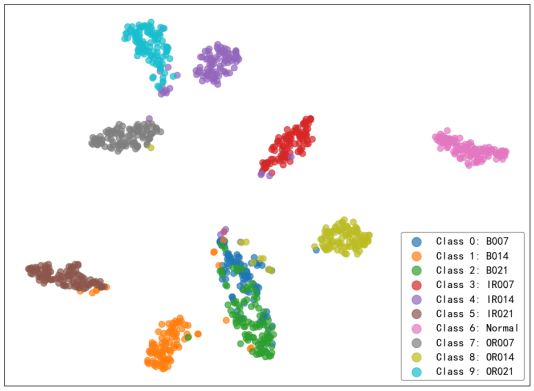


(a) (b) (c) (d)

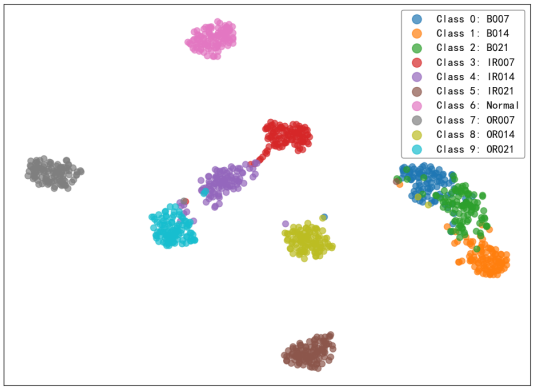
Figure 14. Grad-CAM visualization of scale-specific feature responses under noisy vibration signals. (a) input (b) 3×3 branch (c) 5×5 branch (d) 7×7 branch.

Table 3. Ablation experiment.

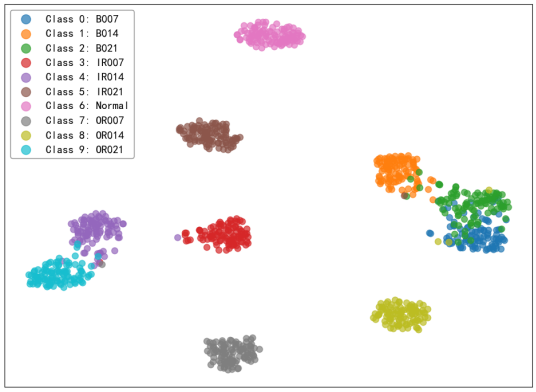
Method	Accuracy
CNN(3×3)	87.22%
MSCNN(3×3+5×5)	90.75%
MSCNN(3×3+5×5+7×7)	92.14%
RSBU-MSCNN	95.20%



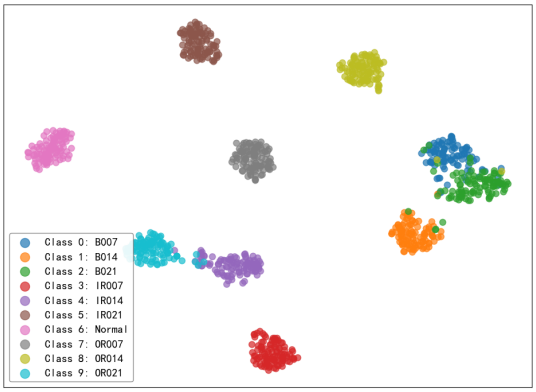
(a)



(b)



(c)



(d)

Figure 15. t-SNE visualization of learned features under different module combinations. (a) CNN(3×3) (b) MSCNN(3×3+5×5) (c) MSCNN(3×3+5×5+7×7) (d) RSBU-MSCNN.

To verify the effectiveness of each key module in enhancing fault diagnosis performance, ablation experiments were conducted on the CWRU dataset under a strong noise environment with an SNR of −6 dB. The average fault classification accuracy on the test set was used as the evaluation metric, and the results are presented in Table 3. As shown in Table 3, the baseline CNN model with a single 3×3 convolution branch achieves an accuracy of 87.22% under strong noise conditions, indicating that severe noise significantly interferes with effective feature extraction. After introducing a multi-scale convolutional structure with two parallel branches (3×3+5×5), the classification accuracy increases to 90.75%, which demonstrates that multi-scale feature extraction can improve the representation capability of the model under noisy conditions. When the 7×7 branch is further incorporated to form the complete MSCNN, the accuracy rises from 90.75% to 92.14%, indicating that the larger receptive-field branch provides additional complementary information beyond the smaller-scale branches and further enhances the multi-scale representation ability of the network. Furthermore, by incorporating the residual shrinkage building unit (RSBU) into the MSCNN framework to construct the RSBU-MSCNN model, the accuracy is further improved to 95.20%. This result demonstrates that the RSBU module can effectively suppress noise-related interference in feature representations while preserving fault-relevant information, thereby further improving the diagnostic performance of the model in strong noise environments.

To further illustrate the effectiveness of different module combinations, t-SNE was employed to visualize the learned feature distributions. As shown in Fig. 15, with the gradual introduction of multi-scale convolution kernels and the RSBU module, the intra-class compactness is progressively improved and the inter-class separation becomes clearer. Compared with the baseline CNN, the RSBU-MSCNN model exhibits more discriminative feature representations, which is consistent with the quantitative results in Table 3.

5.2. XJTU rolling bearing dataset

5.2.1. Dataset introduction

To further validate the effectiveness of the proposed method, this section uses the XJTU-SY bearing dataset for verification. The physical structure of the test bench is shown in Fig. 16. The experimental conditions for sampling are set at 35 kHz/12 kN. This case involves data division for five types of faults, with the specific details provided in Table 4.

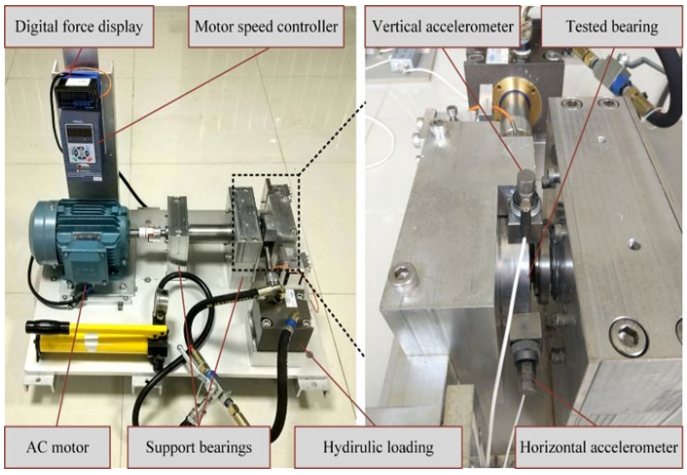


Figure 16. Physical structure of the XJTU-SY test bench.

Table 4. Detailed information of the XJTU-SY experimental dataset.

Fault Label	Fault Type	Train Sample	Test Sample
0	Cage	400	100
1	Inner race and outer race	400	100
2	Outer race_0	400	100
3	Outer race_1	400	100
4	Outer race_2	400	100

5.2.2. Raw signal processing

In this case, Gaussian white noise is added to the raw vibration signal to simulate a real industrial environment. The noisy vibration signal is then subjected to continuous wavelet transform to generate a two-dimensional time-frequency map. Fig. 17 shows the wavelet time-frequency maps of the five fault types at SNR = -6 dB.

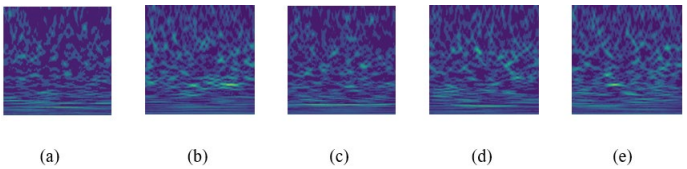


Figure 17. Wavelet time-frequency maps of 5 fault states under the XJTU-SY dataset. (a) Ca (b) Io (c) Or_0 (d) Or_1 (e) Or_2.

5.2.3. Experimental results analysis

Figure 18 shows the accuracy transformation curves of each model under the noise condition of SNR = −6 dB. It is evident from the figure that the proposed model achieves the fastest convergence and superior stability under these conditions. In contrast, ConvNeXt also demonstrates effective convergence but shows inferior noise robustness compared with the proposed model. Although DAE-CNN and DRSN benefit from their denoising or shrinkage mechanisms and achieve relatively stable performance during training, their overall accuracy remains lower than that of RSBU-MSCNN.

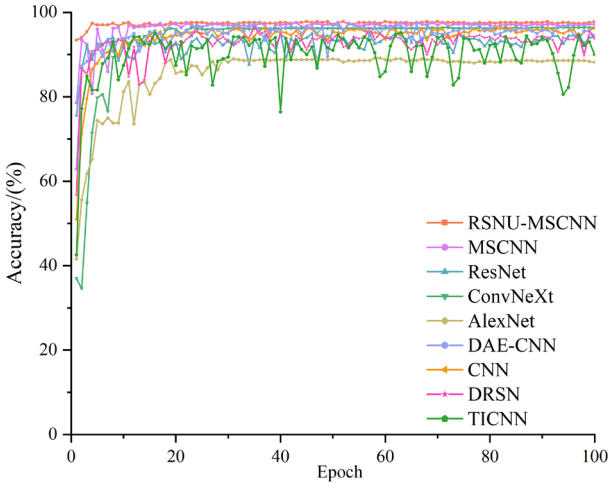


Figure 18. Accuracy curve at SNR = -6 dB for the XJTU-SY dataset.

Figure 19 shows the accuracy variations of each model across noise levels from -6 dB to 0 dB, as well as in the noise-free condition. Under the influence of noise, each model shows a gradual decline in accuracy. However, the proposed method demonstrates an accuracy range of 97.38% to 99.99% across different noise scenarios, achieving over 97% accuracy even in the -6 dB environment. This indicates that the proposed method has high noise robustness.

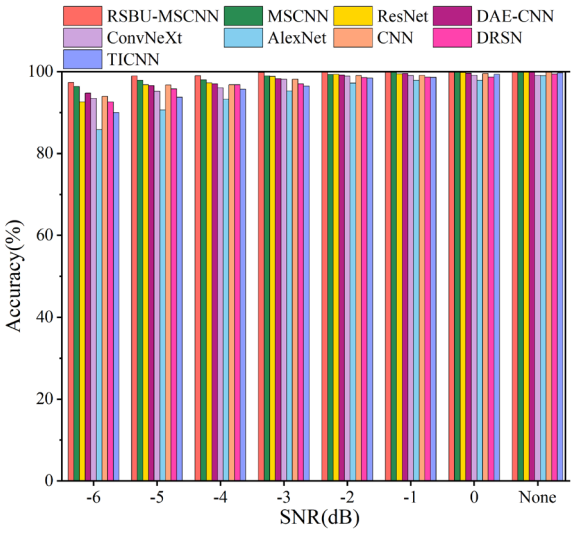
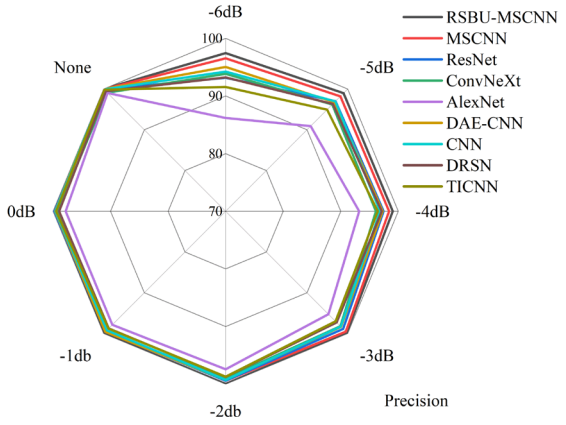


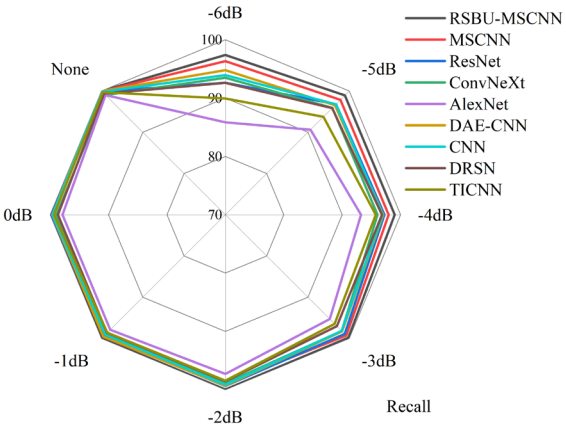
Figure 19. Bar chart of test accuracy for the XJTU-SY dataset.

To further validate the superiority of the proposed method, Fig. 20 shows the radar charts for Precision, Recall, and F1-Score values under varying levels of noise, and it can be observed that the performance of other methods significantly declines. The most noticeable decline is observed in AlexNet, which experiences a rapid performance drop due to the higher impact of noise. In contrast, the fluctuations in the proposed model are minimal, and each metric consistently achieves over 97% under all conditions. This validates that the proposed method can maintain high-level performance even under substantial noise interference.

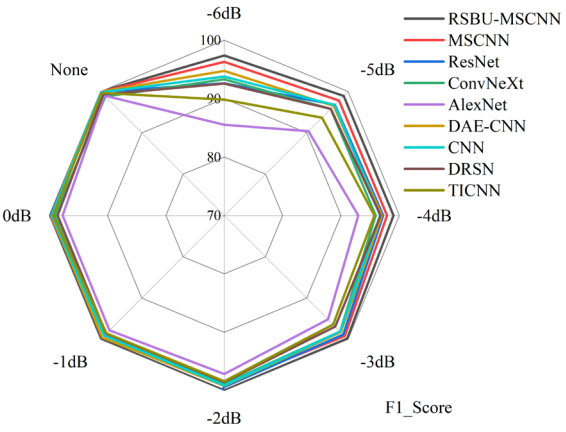
Figure 21 presents the confusion matrices of the proposed model and eight comparative methods under an SNR of −6 dB on the XJTU-SY dataset. The matrices illustrate the classification performance of all nine methods across different fault categories under severe noise conditions. From the results, it can be observed that most comparative models exhibit varying degrees of misclassification, particularly for the outer ring fault categories, where feature similarity and strong noise interference make discrimination more challenging. Among them, AlexNet and CNN show relatively pronounced confusion between fault types, indicating limited robustness in low-SNR environments. ResNet, ConvNeXt, MSCNN, and DAE-CNN improve the classification results to some extent, but misclassified samples are still visible in several categories. DRSN yields a relatively more concentrated distribution along the main diagonal, suggesting that its shrinkage mechanism helps suppress part of the noise interference; however, some confusion remains in the later fault categories. By comparison, TICNN shows more obvious off-diagonal errors, especially for several outer ring fault classes, indicating that its discriminative ability is still affected under severe noise conditions. In contrast, the proposed RSBU-MSCNN maintains a higher concentration of correct predictions along the main diagonal, with fewer off-diagonal errors across almost all fault types.



(a)

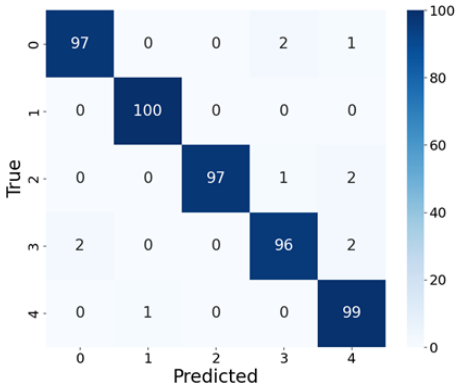


(b)

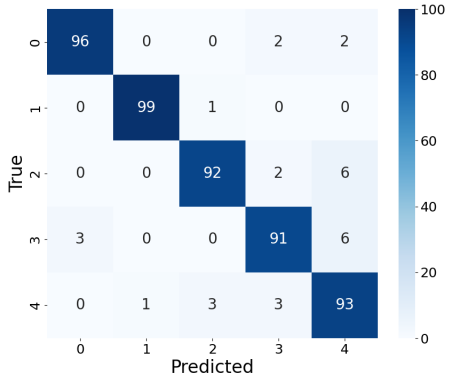


(c)

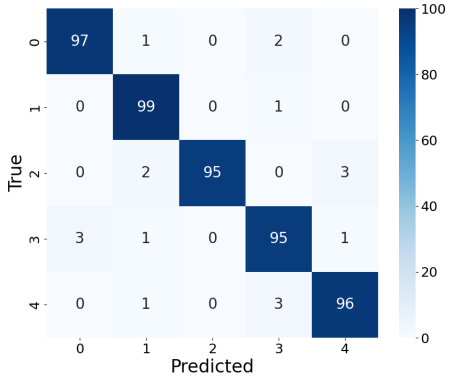
Figure 20. Radar chart of the three evaluation metrics under the XJTU-SY dataset. (a) Precision (b) Recall (c) F1_Score.


True
Predicted
0 1 2 3 4
97 0 0 2 1
0 100 0 0 0
0 0 97 1 2
2 0 0 96 2
0 1 0 0 99
100 80 60 40 20 0

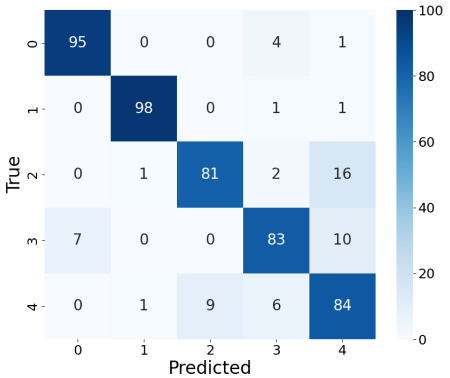
(a)


True
Predicted
0 1 2 3 4
96 0 0 2 2
0 99 1 0 0
0 0 92 2 6
3 0 0 91 6
0 1 3 3 93
100 80 60 40 20 0

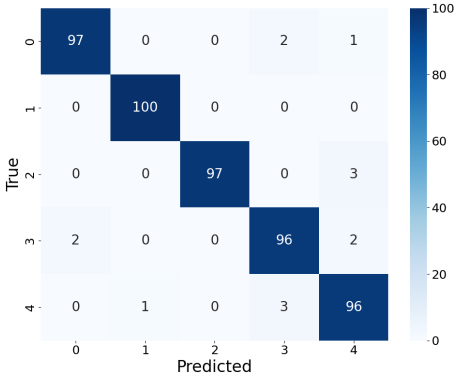
(b)


True
Predicted
0 1 2 3 4
97 1 0 2 0
0 99 0 1 0
0 2 95 0 3
3 1 0 95 1
0 1 0 3 96
100 80 60 40 20 0

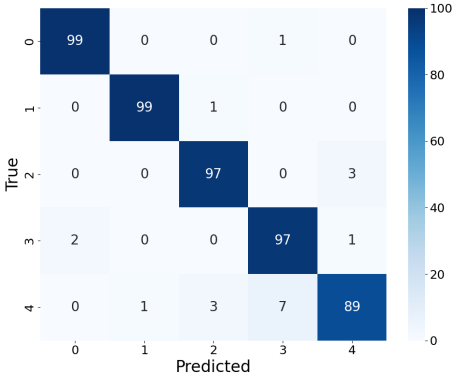
(c)


True
Predicted
0 1 2 3 4
95 0 0 4 1
0 98 0 1 1
0 1 81 2 16
7 0 0 83 10
0 1 9 6 84
100 80 60 40 20 0

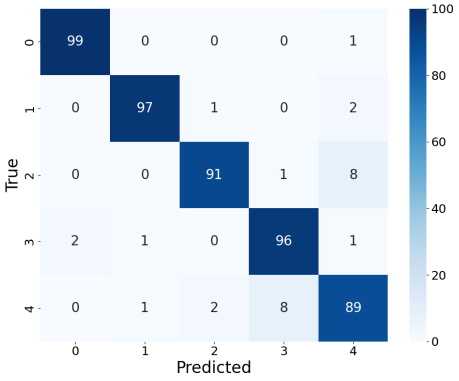
(d)


True
Predicted
0 1 2 3 4
97 0 0 2 1
0 100 0 0 0
0 0 97 0 3
2 0 0 96 2
0 1 0 3 96
100 80 60 40 20 0

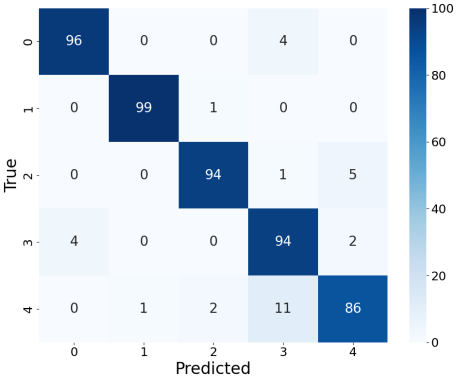
(e)


True
Predicted
0 1 2 3 4
99 0 0 1 0
0 99 1 0 0
0 0 97 0 3
2 0 0 97 1
0 1 3 7 89
100 80 60 40 20 0

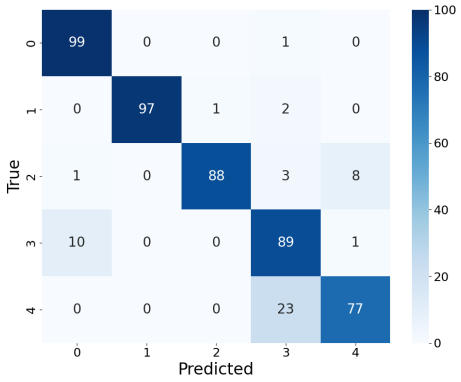
(f)


True
Predicted
0 1 2 3 4
99 0 0 0 1
0 97 1 0 2
0 0 91 1 8
2 1 0 96 1
0 1 2 8 89
100 80 60 40 20 0

(g)

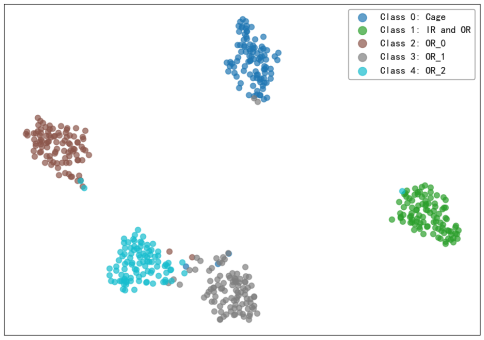

True
Predicted
0 1 2 3 4
96 0 0 4 0
0 99 1 0 0
0 0 94 1 5
4 0 0 94 2
0 1 2 11 86
100 80 60 40 20 0

(h)

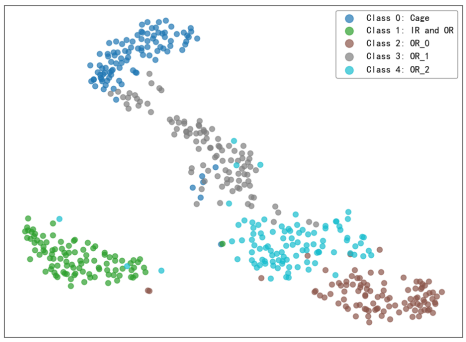


(i)

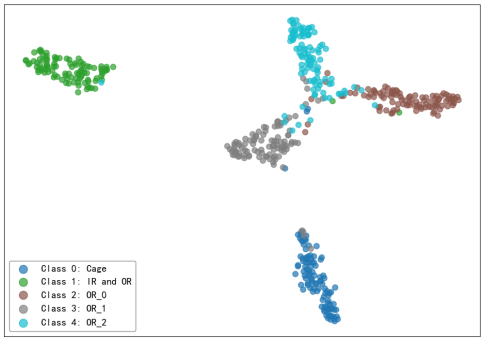
Figure 21. Classification confusion matrices of each method under an SNR of -6 dB on the XJTU-SY dataset. (a) RSBU-MSCNN (b) ResNet (c) ConvNeXt (d) AlexNet (e) MSCNN (f) CNN (g) DAE-CNN (h) DRSN (i) TICNN.



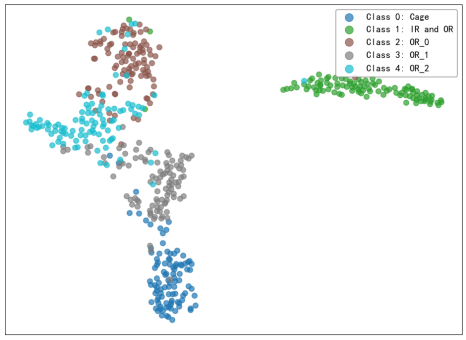
(a)



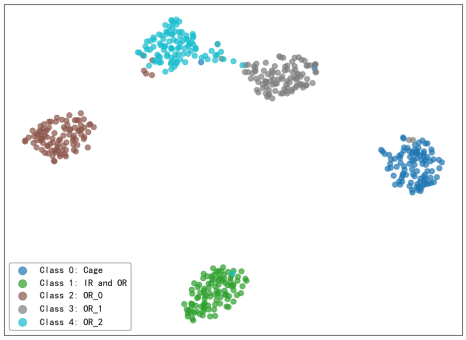
(b)



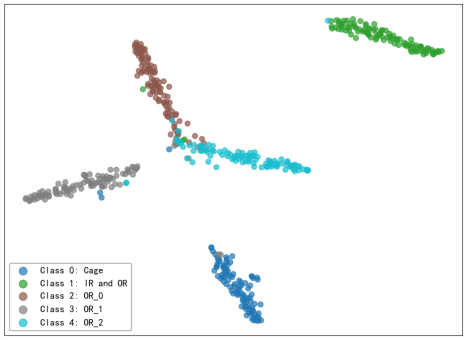
(c)



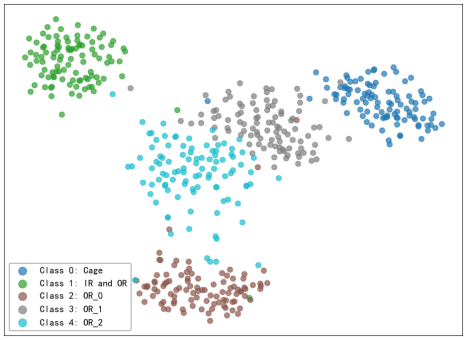
(d)



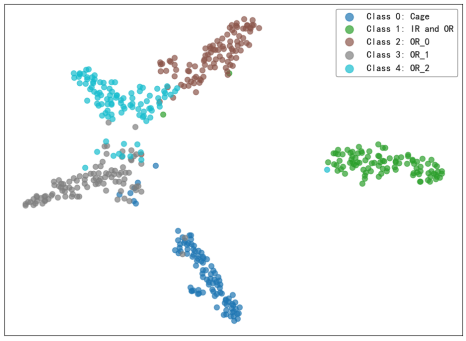
(e)



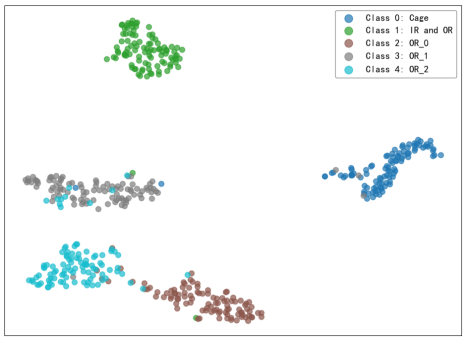
(f)



(g)



(h)



(i)

Figure 22. t-SNE visualization of four methods at SNR = -6 dB.(a) RSBU-MSCNN (b) AlexNet (c) ResNet (d) ConvNeXt (e) MSCNN (f) CNN (g) DAE-CNN (h) DRSN (i) TICNN.

Figure 22 presents the t-SNE visualization results of nine methods under an SNR of −6 dB. Each subfigure illustrates the distribution of learned feature representations for different fault categories in the low-SNR scenario. As shown in Fig. 21, AlexNet and CNN exhibit substantial inter-class overlap, indicating limited discriminative capability under severe noise interference. ConvNeXt and MSCNN demonstrate improved clustering structures; however, partial mixing between certain outer ring fault categories can still be observed. ResNet shows relatively clearer cluster separation, yet noticeable overlap remains among OR_0, OR_1, and OR_2, suggesting that noise still affects feature separability. Although DAE-CNN benefits from its denoising mechanism and achieves more compact clusters than shallow architectures, boundary ambiguity persists for several categories under strong noise conditions. DRSN produces comparatively clearer class boundaries, indicating that its shrinkage mechanism contributes to noise suppression; however, several categories still remain partially close to each other. TICNN also forms relatively compact clusters, but slight overlap and reduced inter-class margins can still be observed for some fault categories under severe noise interference. In contrast, the proposed RSBU-MSCNN forms more compact intra-class clusters with clearer inter-class boundaries. Only slight overlap between OR_1 and OR_2 can be observed, while other categories remain well separated.

Based on the above, the model presented in this paper has demonstrated superior performance across different datasets, proving the generalization and effectiveness of the proposed method. It exhibits strong robustness even in noisy and disturbed working environments, providing solid support for practical industrial applications.

5.3. Generalization under different noise distributions

In practical industrial environments, vibration signals are typically corrupted by structured and non-Gaussian disturbances rather than ideal white Gaussian noise. To assess the robustness of the proposed RSBU-MSCNN under noise distribution variations beyond the Gaussian assumption, two additional noise types that better reflect realistic industrial conditions are introduced at a fixed signal-to-noise ratio (SNR) of −6 dB: (i) harmonic interference noise and (ii) pink (1/f) noise. The same SNR definition as in Eq. (12) is employed to ensure fair and consistent evaluation.

(i) Harmonic Interference Noise

Harmonic interference is used to simulate power-line electromagnetic interference commonly observed in industrial environments. The noise is modeled as a superposition of sinusoidal components at the fundamental frequency and its harmonics:

$$n_h(t) = \sum_{k=1}^{K} a_k \sin(2\pi k f_0 t + \phi_k) \quad (21)$$

Where f_0 denotes the fundamental frequency (randomly selected from 50 Hz or 60 Hz), $K$is the number of harmonics, ϕ_k represents uniformly distributed random phases, and a_k follows an inverse proportional decay $a_k \propto 1/k$. Random amplitude perturbations are additionally introduced to emulate practical electromagnetic interference variations.

(ii)Pink Noise

Pink noise, also known as 1/f noise, is used to emulate colored background noise commonly encountered in mechanical systems. Its power spectral density satisfies:

$$S(f) \propto \frac{1}{f} \quad (22)$$

Compared with white Gaussian noise, pink noise exhibits stronger low-frequency components and more realistic spectral characteristics in industrial scenarios.

All noises are scaled according to the target SNR (−6 dB) using the same energy-based definition

Figure 23 shows the average classification accuracy of different models under harmonic interference noise and pink noise at an SNR of −6 dB. It can be observed that the proposed RSBU-MSCNN achieves the highest accuracy under both noise conditions.

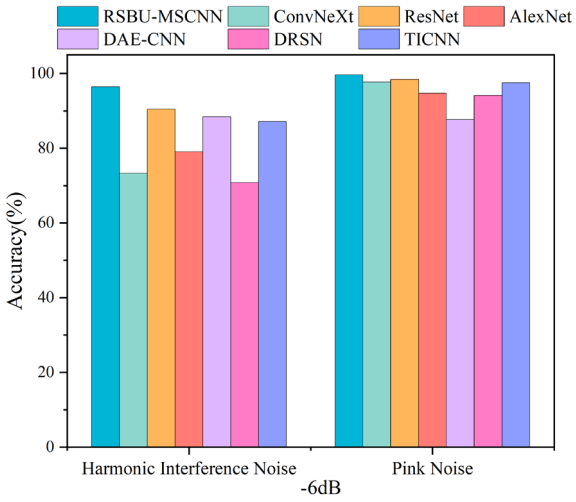


Figure 23. Performance comparison under different non-Gaussian noise distributions at −6 dB SNR.

Under harmonic interference noise, the performance gap among different models becomes more significant. The proposed RSBU-MSCNN achieves an accuracy of above 96%, significantly outperforming ConvNeXt, ResNet, AlexNet, DAE-CNN, as well as the representative noise-robust baseline models DRSN and TICNN. In particular, ConvNeXt exhibits an obvious performance degradation under harmonic interference, indicating its sensitivity to structured periodic disturbances. Under this challenging condition, the robustness of DRSN and TICNN is also inferior to that of the proposed method.

Under pink noise conditions, all models show improved performance compared with the harmonic interference scenario. Nevertheless, RSBU-MSCNN still achieves the best performance among all compared methods, slightly surpassing TICNN and clearly outperforming DRSN, DAE-CNN, AlexNet, and ConvNeXt.

Overall, the results indicate that the proposed method maintains stable performance across different non-Gaussian noise distributions under the same SNR, demonstrating enhanced robustness and improved generalization to various noise types.

To further evaluate the generalization ability of the proposed method under unknown noise intensity conditions, a cross-SNR experiment was conducted in this paper. Specifically, the model was trained using vibration signals contaminated by Gaussian white noise with an SNR of −6 dB, while the test data were collected at a lower SNR of −10 dB. This setting simulates a practical scenario where the noise level during model deployment is more severe than that during training.

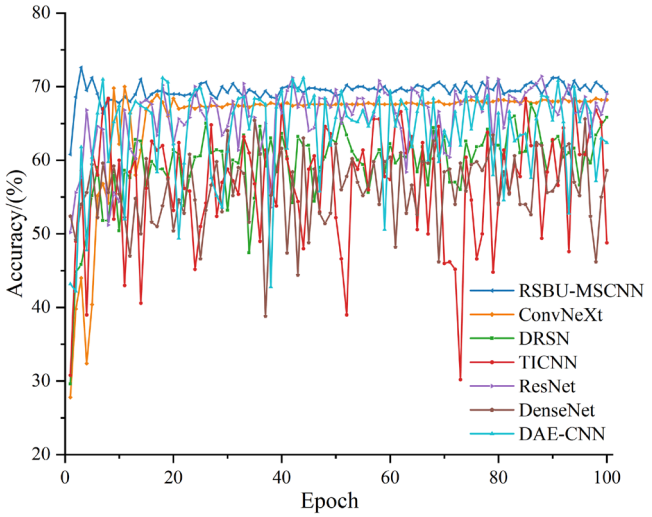


Figure 24. Cross-SNR Generalization Test (-6 dB→-10 dB).

Figure 24 shows the test accuracy curves of different models during the cross-SNR evaluation. It can be observed that most comparison methods suffer from obvious performance degradation and unstable convergence when tested under the unknown −10 dB noise condition. In contrast, the proposed RSBU-MSCNN maintains relatively stable accuracy fluctuations across all training epochs and consistently achieves high recognition accuracy among all compared models. The cross-SNR experimental results demonstrate that the proposed RSBU-MSCNN still exhibits stronger generalization ability when the noise level in the test environment exceeds the range observed during the training phase, further verifying the strong robustness of the proposed model.

6. Conclusion

This paper proposes a collaborative optimization framework, termed RSBU-MSCNN, for rolling bearing fault diagnosis under strong background noise conditions. Unlike conventional approaches that treat denoising and feature extraction as separate stages, the proposed method integrates these processes into a unified end-to-end architecture, enabling collaborative modeling of noise suppression and discriminative feature learning. The main innovations and findings of this study can be summarized as follows:

(1) A synergistic denoising and feature representation mechanism is established for strong-noise fault diagnosis. The core innovation of the proposed method lies in the collaborative integration of 2D residual

shrinkage building units (2D-RSBU) and multi-scale convolutional feature fusion within a unified framework. Specifically, the 2D-RSBU performs channel-wise adaptive soft-thresholding, which suppresses noise-dominated feature responses while preserving weak fault-related components. In parallel, the multi-scale convolution branches extract complementary fault patterns at different receptive fields, enabling the model to capture both local discriminative textures and broader structural information. Through this cooperation, the proposed framework forms a sequential "adaptive purification–multi-scale reconstruction" mechanism, which effectively improves feature separability and robustness under low-SNR conditions.

(2) A strong-noise-oriented network architecture is designed for time-frequency fault representations. To adapt residual shrinkage learning to CWT-based time-frequency images, the conventional RSBU is extended to a 2D spatially aware shrinkage unit, allowing adaptive threshold learning in the latent feature space of two-dimensional representations. On this basis, a parallel multi-branch structure with convolution kernels of sizes 3×3, 5×5, and 7×7 is introduced to alleviate the scale mismatch problem under strong noise conditions. This design enhances the ability of the network to jointly model fine-grained local details and robust global contextual information, thereby improving the representation quality of weak fault signatures submerged in noise.

(3) Extensive experiments verify the robustness and generalization capability of the proposed method. Experimental results on the CWRU and XJTU-SY datasets demonstrate that RSBU-MSCNN consistently outperforms multiple comparative models, including conventional CNN-based methods and representative noise-robust approaches such as DRSN and TICNN, under severe noise conditions. At an SNR of −6 dB, the proposed model achieves diagnostic accuracies of 95.20% and 97.38% on the two datasets, respectively. In addition, the model maintains stable performance under non-Gaussian disturbances (harmonic interference and pink noise) and in cross-SNR evaluation, which further confirms its robustness and generalization ability in complex industrial noise environments.

Overall, by tightly coupling adaptive shrinkage mechanisms with multi-scale representation learning within a unified optimization framework, this study establishes a robust fault diagnosis approach for strong-noise environments, providing a theoretically grounded and practically applicable solution for intelligent condition monitoring in complex industrial scenarios.

This study primarily focuses on rolling bearing fault diagnosis under strong noise environments and validates the effectiveness of the proposed RSBU-MSCNN framework on balanced datasets under constant-speed operating conditions. Compared with representative noise-robust methods such as DRSN and TICNN, the proposed method shows superior robustness and discriminative capability under severe noise interference, which can be attributed to the joint design of adaptive shrinkage and multi-scale feature extraction. However, the proposed framework still has some limitations. Due to the introduction of multiple convolution branches, the model architecture is relatively more complex than that of some conventional backbone networks, which may increase computational cost in practical deployment. In addition, the current study is conducted under relatively controlled conditions, while real industrial scenarios are often more complex, involving variable speeds, fluctuating loads, and data imbalance issues. Therefore, future work will further investigate lightweight model design and evaluate the adaptability and robustness of the proposed framework under variable-speed and load-changing conditions, as well as on imbalanced datasets, to enhance its practical applicability.

References

1. Zhang Y, Ji J, Ren Z, Ni Q, Gu F, Feng K, Yu K, Ge J, Lei Z, Liu Z. Digital twin-driven partial domain adaptation network for intelligent fault diagnosis of rolling bearing. *Reliability Engineering & System Safety* 2023; 234: 109186. https://doi.org/10.1016/j.ress.2023.109186.
2. Wu H, Li J, Zhang Q, Tao J, Meng Z. Intelligent fault diagnosis of rolling bearings under varying operating conditions based on domain-adversarial neural network and attention mechanism. *ISA Transactions* 2022; 130: 477-489. https://doi.org/10.1016/j.isatra.2022.04.026.
3. Xu Y, Li Z, Wang S, Li W, Sarkodie-Gyan T, Feng S. A hybrid deep-learning model for fault diagnosis of rolling bearings. *Measurement*

2021; 169: 108502. https://doi.org/10.1016/j.measurement.2020.108502.

4. Wu D, Guan H, Zhao H. Parameterized Iterative Time–Frequency-Multisqueezing Transform for Bearing Fault Diagnosis. *IEEE Transactions on Instrumentation and Measurement* 2025; 74: 1-11. https://doi.org/10.1109/tim.2025.3561399.
5. Zhong J, Lin C, Gao Y, Zhong J, Zhong S. Fault diagnosis of rolling bearings under variable conditions based on unsupervised domain adaptation method. *Mechanical Systems and Signal Processing* 2024; 215: 111430. https://doi.org/10.1016/j.ymssp.2024.111430.
6. Cheng J, Yang Y, Shao H, Pan H, Zheng J, Cheng J. Enhanced periodic mode decomposition and its application to composite fault diagnosis of rolling bearings. *ISA Transactions* 2022; 125: 474-491. https://doi.org/10.1016/j.isatra.2021.07.014.
7. Gao Y, Yu D. Intelligent fault diagnosis for rolling bearings based on graph shift regularization with directed graphs. *Advanced Engineering Informatics* 2021; 47: 101253. https://doi.org/10.1016/j.aei.2021.101253.
8. Yu M, Zhang Y, Yang C. Rolling bearing faults identification based on multiscale singular value. *Advanced Engineering Informatics* 2023; 57: 102040. https://doi.org/10.1016/j.aei.2023.102040.
9. Cai B, Zhang L, Tang G. Encogram: An autonomous weak transient fault enhancement strategy and its application in bearing fault diagnosis. *Measurement* 2023; 206: 112333. https://doi.org/10.1016/j.measurement.2022.112333.
10. Ni Q, Ji J C, Halkon B, Feng K, Nandi A K. Physics-Informed Residual Network (PIResNet) for rolling element bearing fault diagnostics. *Mechanical Systems and Signal Processing* 2023; 200: 110544. https://doi.org/10.1016/j.ymssp.2023.110544.
11. Chen S, Zheng W, Xiao H, Han P, Luo K. A residual convolution transfer framework based on slow feature for cross-domain machinery fault diagnosis. *Neurocomputing* 2023; 546: 126322. https://doi.org/10.1016/j.neucom.2023.126322.
12. Zhao H, Liu C, Dang X, Xu J, Deng W. Few-Shot Cross-Domain Fault Diagnosis of Transportation Motor Bearings Using MAML-GA. *IEEE Transactions on Transportation Electrification* 2026; 12(1): 1165-1174. https://doi.org/10.1109/tte.2025.3625779.
13. Li J, Deng W, Ding J, Zhao H. IBN-MixStyle Network With Dynamic Weighted Invariant Risk Minimization for Domain-Generalized Bearing Fault Diagnosis. *IEEE Transactions on Consumer Electronics* 2025; 71(4): 9929-9939. https://doi.org/10.1109/tce.2025.3607134.
14. Huang C, Peng Y, Deng W. A dendrite net learning multi-objective artificial bee colony algorithm for UAV. *Applied Soft Computing* 2026; 189: 114449. https://doi.org/10.1016/j.asoc.2025.114449.
15. Deng W, Li X, Sun Y, Zhao H. Privacy Protection-Enhanced Vertical-Horizontal Federated Learning Secure Sharing for Multisource Heterogeneous Data. *IEEE Transactions on Industrial Informatics* 2026; 1-10. https://doi.org/10.1109/tii.2025.3649540.
16. Dao F, Zeng Y, Qian J. Fault diagnosis of hydro-turbine via the incorporation of bayesian algorithm optimized CNN-LSTM neural network. *Energy* 2024; 290: 130326. https://doi.org/10.1016/j.energy.2024.130326.
17. Guo Z, Yang M, Huang X. Bearing fault diagnosis based on speed signal and CNN model. *Energy Reports* 2022; 8: 904-913. https://doi.org/10.1016/j.egyr.2022.08.041.
18. Wang H, Xu J, Yan R, Gao R X. A New Intelligent Bearing Fault Diagnosis Method Using SDP Representation and SE-CNN. *IEEE Transactions on Instrumentation and Measurement* 2020; 69(5): 2377-2389. https://doi.org/10.1109/tim.2019.2956332.
19. Gu J, Peng Y, Lu H, Chang X, Chen G. A novel fault diagnosis method of rotating machinery via VMD, CWT and improved CNN. *Measurement* 2022; 200: 111635. https://doi.org/10.1016/j.measurement.2022.111635.
20. Sinitsin V, Ibryaeva O, Sakovskaya V, Eremeeva V. Intelligent bearing fault diagnosis method combining mixed input and hybrid CNN-MLP model. *Mechanical Systems and Signal Processing* 2022; 180: 109454. https://doi.org/10.1016/j.ymssp.2022.109454.
21. Deng W, Li H, Zhao H. Antinoise Bearing Fault Diagnosis Using Time-Reassigned Multisynchrosqueezing Transform and Complex Sparse Learning Dictionary. *IEEE Transactions on Instrumentation and Measurement* 2025; 74: 1-10. https://doi.org/10.1109/tim.2025.3604987.
22. Wang X, Mao D, Li X. Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. *Measurement* 2021; 173: 108518. https://doi.org/10.1016/j.measurement.2020.108518.
23. Han Y, Zhang F, Li Z, Wang Q, Li C, Lai P, Li T, Teng F, Jin Z. MT-ConvFormer: A Multitask Bearing Fault Diagnosis Method Using a Combination of CNN and Transformer. *IEEE Transactions on Instrumentation and Measurement* 2025; 74: 1-16. https://doi.org/10.1109/tim.2024.3502821.
24. Wang H, Liu Z, Peng D, Cheng Z. Attention-guided joint learning CNN with noise robustness for bearing fault diagnosis and vibration signal denoising. *ISA Transactions* 2022; 128: 470-484. https://doi.org/10.1016/j.isatra.2021.11.028.
25. Han T, Ma R, Zheng J. Combination bidirectional long short-term memory and capsule network for rotating machinery fault diagnosis.

Measurement 2021; 176: 109208. https://doi.org/10.1016/j.measurement.2021.109208.

26. Zhao K, Xiao J, Li C, Xu Z, Yue M. Fault diagnosis of rolling bearing using CNN and PCA fractal based feature extraction. *Measurement* 2023; 223: 113754. https://doi.org/10.1016/j.measurement.2023.113754.
27. Peng S, Xing J, Liu X. A Rolling Bearing Vibration Signal Noise Reduction Processing Algorithm Using the Fusion HPO-VMD and Improved Wavelet Threshold. *Symmetry* 2025; 17(8): 1316. https://doi.org/10.3390/sym17081316.
28. Qiu Z, Fan S, Liang H, Liu J. Multimodal fusion fault diagnosis method under noise interference. *Applied Acoustics* 2025; 228: 110301. https://doi.org/10.1016/j.apacoust.2024.110301.
29. Du Y, Geng X, Zhou Q, Cheng S. A fault diagnosis method for offshore wind turbine bearing based on adaptive deep echo state network and bidirectional long short term memory network in noisy environment. *Ocean Engineering* 2024; 312: 119101. https://doi.org/10.1016/j.oceaneng.2024.119101.
30. Li D, Li M, Yang L, Wang X, Zhang F, Liang Y. Rolling bearing fault diagnosis in strong noise background based on vibration signals. *Signal, Image and Video Processing* 2023; 18(2): 1295-1303. https://doi.org/10.1007/s11760-023-02846-y.
31. Xiao B, Zhao Y, Zhou C, Ou J, Huang G. A noise-robust CNN architecture with global attention and gated convolutional Kernels for bearing fault detection. *Measurement Science and Technology* 2024; 35(8): 086142. https://doi.org/10.1088/1361-6501/ad4d16.
32. Chen Z, Wang Y, Wu J, Deng C, Hu K. Sensor data-driven structural damage detection based on deep convolutional neural networks and continuous wavelet transform. *Applied Intelligence* 2021; 51(8): 5598-5609. https://doi.org/10.1007/s10489-020-02092-6.
33. Cheng Y, Lin M, Wu J, Zhu H, Shao X. Intelligent fault diagnosis of rotating machinery based on continuous wavelet transform local binary convolutional neural network. *Knowledge-Based Systems* 2021; 216: 106796. https://doi.org/10.1016/j.knosys.2021.106796.
34. Belaid K, Miloudi A, Bournine H. The processing of resonances excited by gear faults using continuous wavelet transform with adaptive complex Morlet wavelet and sparsity measurement. *Measurement* 2021; 180: 109576. https://doi.org/10.1016/j.measurement.2021.109576.
35. Huang Y J, Liao A H, Hu D Y, Shi W, Zheng S B. Multi-scale convolutional network with channel attention mechanism for rolling bearing fault diagnosis. *Measurement* 2022; 203: 111935. https://doi.org/10.1016/j.measurement.2022.111935.
36. Tang S, Zhu Y, Yuan S. Intelligent fault diagnosis of hydraulic piston pump based on deep learning and Bayesian optimization. *ISA Transactions* 2022; 129: 555-563. https://doi.org/10.1016/j.isatra.2022.01.013.
37. Li J, Lin M, Li Y, Wang X. Transfer learning network for nuclear power plant fault diagnosis with unlabeled data under varying operating conditions. *Energy* 2022; 254: 124358. https://doi.org/10.1016/j.energy.2022.124358.
38. Zhao M, Zhong S, Fu X, Tang B, Pecht M. Deep Residual Shrinkage Networks for Fault Diagnosis. *IEEE Transactions on Industrial Informatics* 2020; 16(7): 4681-4690. https://doi.org/10.1109/tii.2019.2943898.
39. Lv X, Wang J, Qin R, Bao J, Jiang X, Zhang Z, Han B, Jiang X. Self-learning guided residual shrinkage network for intelligent fault diagnosis of planetary gearbox. *Engineering Applications of Artificial Intelligence* 2025; 139: 109603. https://doi.org/10.1016/j.engappai.2024.109603.
40. Zhan F, Hu L, Huang W, Dong Y, He H, Wu G. Category knowledge-guided few-shot bearing fault diagnosis. *Engineering Applications of Artificial Intelligence* 2025; 139: 109489. https://doi.org/10.1016/j.engappai.2024.109489.
41. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016; 770-778. https://doi.org/10.1109/CVPR.2016.90.
42. Liu Z, Mao H, Wu C Y, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2022; 11976-11986. https://doi.org/10.1109/CVPR52688.2022.01167.
43. Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 2017; 60(6): 84-90. https://doi.org/10.1145/3065386.
44. Zhang W, Li C, Peng G, Chen Y, Zhang Z. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical Systems and Signal Processing* 2018; 100: 439-453. https://doi.org/10.1016/j.ymssp.2017.06.022.
45. Vincent P, Larochelle H, Bengio Y, Manzagol P A. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning* 2008; 1096-1103. https://doi.org/10.1145/1390156.1390294.