Eksploatacja i Niezawodnosc - Maintenance and Reliability
Volume 29 (2027), Issue 3
https://ein.org.pl
The Polish Maintenance Society

Synergistic integration of AdaBoundL2 optimization and DSCNN: a novel framework for rolling bearing fault diagnosis

Xinyue Pei[1], Xiujian Chou[1,*], JiLiang Mu[1]

[1] North University of China, China
* Corresponding Author: chouxiujian@nuc.edu.cn

Received: 6 February 2026
Revised: 17 October 2026
Accepted: 30 October 2026
Online: 20 November 2026

This is an open access article under the CC BY 4.0 license

Abstract

In the extraction of fault data from large-scale mechanical equipment bearings, environmental noise significantly constrains the precision and speed of fault identification. Despite extensive research on convolutional neural networks (CNNs) for intelligent fault diagnosis, several challenges remain, including the difficulty in noise separation, slow convergence due to gradient induction, and a high number of convolutional parameters during bearing fault analysis. This paper introduces a bearing fault diagnosis method based on a depthwise separable convolutional neural network optimized by AdaBoundL2 (AdaBoundL2-DSCNN). The model utilizes a custom regularized AdaBoundL2 optimizer, dynamically adjusting the parameters final_lr and gamma to rapidly explore the parameter space in the early training stages, effectively mitigating noise impact. By synergistically combining depth-wise separable convolutions into the optimization framework, the standard convolution is decomposed into channel-wise spatial convolutions and point-wise 1×1 convolutions, decoupling spatial filtering from channel transformation. Even in complex noise environments (−8 dB ~ 0 dB), it achieves high-precision diagnosis on the CWRU dataset within fewer epochs (epoch < 3), with an average accuracy of 95.71% that surpasses benchmark models. To minimize model complexity, an optimized GAP+FC(Global Average Pooling, GAP; Fully Connected Layer, FC) structure is employed for data dimensionality reduction, resulting in a pa-rameter count of 12.77k, substantially lower than other model architectures, thereby improving generalization capabil-ity.Experimental results demonstrate that the proposed method benefits from such synergistic optimization, significantly surpas-sing traditional time-domain and frequency-domain analysis methods in fault identification rates across various operating condi-tions and exhibits strong noise resistance, with an average macro F1-score exceeding 98.69%.

Keywords: *fault diagnosis, structural optimization, depthwise separable convolution, improved regularized adaptive optimizer*

Article citation:
Pei X, Chou X, Mu J, Synergistic integration of AdaBoundL2 optimization and DSCNN: a novel framework for rolling bearing fault diagnosis, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2027: 29(3) http://doi.org/10.17531/ein/242346

Highlights

- Proposes a novel framework combining AdaBoundL2 and SWDCNN for fault diagnosis.
- AdaBoundL2 optimizer ensures high-accuracy diagnosis in strong noise environments.
- A depthwise separable neural network forms the foundation of our lightweight architecture prototype.
- Superior to traditional methods, achieving a high macro-average F1-score of >98.69% across various operational conditions.

1. Introduction

Rolling bearings play a crucial role in fields such as aer-ospace, transportation, and industrial manufacturing [1]. However, large mechanical systems often face issues like high maintenance costs and prolonged downtime. Conse-quently, bearing fault diagnosis is essential for reducing maintenance costs and enhancing operational safety [2]. In recent years, with the rapid advancement of artificial intel-ligence, numerous convolutional neural network architec-tures based on deep learning [3−6] have been proposed, achieving significant progress in fault diagnosis technolo-gy and reliability engineering [7]. Nevertheless, key tech-nical challenges persist, including difficulties in feature extraction due to noise interference, slow convergence caused by gradient dispersion in deep networks, and the trade-off between model complexity and real-time perfor-mance requirements. Current fault diagnosis methods typi-cally employ a two-stage "noise reduction followed by di-agnosis" architecture, utilizing strategies such as attention mechanisms

[8], probabilistic modeling [9,10], and mul-ti-scale analysis [11] to improve the signal-to-noise ratio. However, these methods often suffer from high model com-plexity and an inherent conflict between noise suppression and feature fidelity. This study proposes a collaborative optimization framework encompassing "noise perception, resource allocation, and feature preservation," which achieves higher diagnostic accuracy within lightweight models. While AdaBound claims to be insensitive to hy-perparameters, it still exhibits sensitivity to final_lr and gamma in hyperspectral image classification and complex network structures [12,13], as these two critical parameters are statically preset and cannot adaptively adjust with training progress. This paper introduces a dynamic adaptive learning paradigm that adjusts the final_lr and gamma pa-rameters. This approach ensures adaptive matching to time-varying noise characteristics while maintaining model lightweightness, optimizing computational resource allo-cation, and preserving feature extraction accuracy. Tradi-tional convolutional neural networks frequently encounter slow convergence due to gradient-induced issues during training, which severely limit the model's performance and deployment efficiency in practical applications. Currently, researchers focus on optimizing the flow of gradients or activation signals to accelerate network convergence. Techniques such as network architecture optimization, the introduction of activation functions [14], and guiding the training process via information entropy mechanisms [15] have been employed to address slow convergence, albeit with varying resource costs and limitations. The light-weight model structure proposed in this paper aims to re-duce parameters and computational complexity, thereby shortening the gradient transmission path, mitigating gra-dient vanishing, and significantly enhancing the model's convergence speed and inference efficiency. This approach offers greater generalizability and deployment value. As network depth increases, the parameter count in convolu-tional neural networks (CNNs) rises substantially. In classi-cal CNN architectures (e.g., VGG, ResNet), each convolu-tional layer generates numerous parameters, posing signifi-cant storage and computational challenges [16]. Existing parameter compression methods (e.g., DEBI [17], three-layer RNN [18], DWSR [19]) can reduce model com-plexity by 22-35%, but they often result in reduced feature representation capability (PSNR decrease of over 1.8dB) or compromised temporal modeling performance. In the con-text of fault diagnosis under strong noise conditions, some existing studies employ a signal preprocessing module for noise reduction. Although this approach mitigates noise interference to a certain extent, it inevitably increases over-all system complexity [20,21]. Another category of meth-ods enhances model representational capacity by incorpo-rating attention mechanisms or graph neural networks, yet often overlooks the influence of optimization stability on diagnostic performance [22,23]. Furthermore, most existing approaches fail to strike a balance between model com-pactness and noise robustness at the architectural level. Nevertheless, vision oriented lightweight networks suffer inherent limitations for vibration fault diagnosis. Excessive channel compression and heavy group convolution readily trigger overfitting and destroy the temporal continuity of fault transients. Differently, DSCNN preserves partial re-dundant channels to maintain fault features under strong noise conditions. This paper proposes a spatial fea-ture compression strategy to eliminate parameter redun-dancy inherent in traditional architectures. An adaptive dimension regulation mechanism is employed to achieve efficient feature channel reorganization. Furthermore, glob-al aggregation and dimensionality reduction of the feature space are performed to significantly lower computational complexity while preserving feature expression capabilities [24].

2. Proposed method

As depicted in Figure 1, the neural network architec-ture proposed in this paper leverages depthwise separable convolution and the AdaBoundL2 dynamic optimizer, tai-lored for fault diagnosis in noisy environments. The model inputs one-dimensional vibration signals augmented with Gaussian white noise and utilizes a one-dimensional con-volutional network to extract pivotal features. By breaking down standard convolutions into channel-wise and point-wise convolutions, deep separable convolution dras-tically reduces parameters and computational burden while boosting structural efficiency, thereby enhancing the per-formance of both the optimizer and activation functions. The AdaBoundL2 optimizer, integrating dynamic learning rate boundaries with L2 regularization, ensures stable tun-ing of lightweight structural weights, thereby accelerating convergence and bolstering generalization

capabilities. The incorporation of the Hard-Swish activation function amplifies nonlinear representation, with its smooth response traits mitigating issues related to inadequate feature representation and refining gradient propagation. Ulti-mately, the GAP + FC structure is employed to condense redundant information and distill discriminative features for classification.

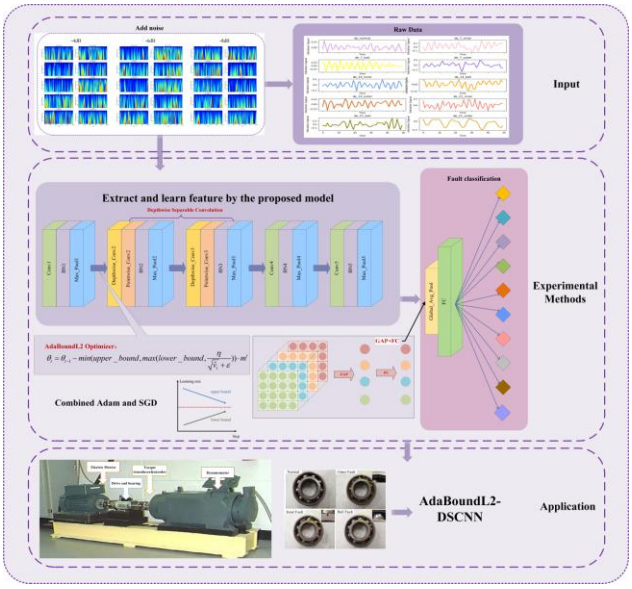


Figure 1. Fault diagnosis structure of AdaBoundL2-DSCNN.

2.1. Depthwise separable convolution

The lightweight structure of depthwise separable convolution is widely used in industrial fault diagnosis due to its efficient computational characteristics [25–27]. Comprising layer-by-layer depthwise convolution and pointwise convolution, it significantly reduces computational load and parameter count [28]. Depthwise separable convolution comprises two key operations: depthwise convolution (DWConv) and pointwise convolution (PWConv), significantly enhancing computational efficiency by decomposing the standard convolution process. As illustrated in Figure 2, for an input feature map of size 12×12×3, depthwise convolution uses three 5×5×1 convolution kernels to independently process each input channel, producing an 8×8×3 feature map with only 75 parameters. Following this, pointwise convolution employs four 1×1×3 convolution kernels for channel dimension transformation, resulting in an 8×8×4 feature map, requiring merely 12 parameters. This decomposition strategy lowers the total parameter count to 87, a 71% reduction compared to the 300 parameters needed for standard convolution. The innovation of depthwise separable convolution lies in decoupling the spatial feature extraction and channel feature fusion functions: depthwise convolution focuses on extracting spatial features from individual channels, while pointwise convolution handles cross-channel feature integration. This design significantly reduces computational complexity and achieves an optimized balance between computational efficiency and model performance by preserving the integrity of feature extraction, offering an effective solution for lightweight neural network architectures. Depthwise separable convolution mathematically decouple spatial convolutions from channel convolutions. This decoupling mechanism suppresses noise propagation across channels, reduces gradient variance, and stabilizes the parameter update process. Meanwhile, the drastically reduced parameter space lowers model complexity, thereby mitigating overfitting induced by noise.

The parameter calculation formulas for the ordinary convolution and depthwise separable convolution are presented as follows:

$$H \times W \times S \times K \times D_f \times D_f \quad (1)$$

$$H \times W \times K \times D_f \times D_f + H \times W \times S \times K \quad (2)$$

H denotes the height of the feature map (degenerate to 1 in the 1D vibration scenario, as the signal possesses only the temporal dimension), W signifies the width of the feature map(corresponding to the sampling length along the time axis), S refers to the size of the convolutional kernel(i.e., the temporal receptive field covering consecutive sampling points), K indicates the number of convolutional kernels(determining the output channels, each learning distinct vibration patterns), and D_f denotes the depth of the convolutional kernel(equal to the number of input sensor channels).

The latest lightweight models and the comparison of model complexity are shown in Table 1.

Table 1. The comparison of model complexity.

Model	Params	FLOPs	Key Feature
WDCNN	2.30 M	0.80 G	Classic fault diagnosis baseline, heavy computation
ResCNN-50	3.80 M	1.20 G	Deep feature extraction, high memory occupation
VGG-16 (1D adapted)	7.20 M	2.10 G	High precision, not suitable for edge devices
CNN-Transformer Hybrid	4.60 M	1.50 G	Strong global feature, huge inference delay
MobileNetV3	5.40 M	0.50 G	Relatively light but still redundant for industrial edge

These classical models generally suffer from excessive parameter redundancy and heavy computational overhead.

Large-kernel standard convolutions and deep stacked structures lead to high memory occupancy and slow inference speed, making them incompatible with edge-end embedded deployment. Hybrid architectures introduce powerful global feature extraction but further exacerbate latency and resource consumption. Even the improved separable convolution schemes of MobileNetV3 still retain redundant channel settings, failing to achieve optimal lightweight efficiency under complex industrial noise and real-time diagnosis requirements.

2.2. AdaBoundL2

AdaBoundL2 inherits the core mechanism of the Adam optimizer, employing exponential decay averaging to dynamically estimate the first and second moments of the gradient. It achieves smooth processing of historical gradients through a time-decay weighted strategy, assigning higher weight coefficients to recent gradients. This algorithm maintains the advantage of Adam's adaptive learning rate while effectively controlling the magnitude of parameter updates by introducing L2 regularization constraints and a dynamic boundary mechanism, thereby striking a better balance between training stability and convergence speed. Compared to traditional Adam, AdaBoundL2 significantly enhances the optimization performance of deep models on non-stationary objective functions through a dual regulation mechanism involving gradient moment estimation and update boundary constraints.

In noisy gradient scenarios, the dynamic boundary mechanism of AdaBoundL2 mitigates the update instability of Adam caused by estimation bias in the second-order moment of gradients, by constraining extreme values of the adaptive step size. Meanwhile, it enables a smooth transition to SGD in the late training phase, driving the model to converge to flat minima with noise robustness. This dual-stage behavior of “adaptive exploration plus stable convergence” is more tolerant to strong noise interference in rolling bearing vibration signals than vanilla Adam or SGD alone.

AdaBoundL2 gradually converges the learning rate to a stable range by dynamically adjusting the final_lr and gamma parameters. The dynamic adjustment equations for final_lr and gamma are as follows: every 1000 steps, final_lr experiences a 10% attenuation, and gamma increases by 10%.

$$final_lr_t = initial_final_lr \cdot (1\text{-}0.1)^{\frac{t}{1000}} \quad (3)$$

$$\gamma_t = initial_gamma \cdot (1{+}0.1)^{\frac{t}{1000}} \quad (4)$$

Note that the 1000-step update interval is a tunable hyperparameter. It requires recalibration when the dataset scale or batch size is changed. The proposed strategy adopts scheduled coupled adjustment of final_lr and gamma, instead of a fully closed-loop data-driven adaptive scheme.

According to the updated final_lr and gamma, the learning rate bounds are calculated:

$$lower_bound = final_lr_t \cdot \left(1 - \frac{1}{\gamma_t \cdot t+1}\right) \quad (5)$$

$$upper_bound = final_lr_t \cdot \left(1 + \frac{1}{\gamma_t \cdot t+1}\right) \quad (6)$$

AdaBoundL2 ingests a constraint on the learning rate based on AdaBound, and the constraint interval is [lower_bound, upper_bound]. The formula to calculate the adaptive learning rate is as follows:

$$step_size = clamp\left(\eta \cdot \frac{\sqrt{\hat{v}_t}+\varepsilon}{\hat{m}_t}, lower_bound, upper_bound\right) \quad (7)$$

Ultimately, the parameter update formula is:

$$\theta_t = \theta_{t-1} - step_size \cdot \hat{m}_t \quad (8)$$

Explanations of relevant parameters are shown in Table 2.

Table 2. Explanations of relevant parameters

Parameters	Definition
η	Initial learning rate
γ_t	controls the hyperparameter of the learning rate, the dynamic boundary convergence speed
ϵ	Smooth constant (typically taking ϵ=1e-8), used for numerical stability
$\hat{m}_t$	First-order moment estimation
$\hat{v}_t$	Second-order moment estimation

To address the issue of setting the final_lr decay rate, experiments were conducted with varying decay rates. As shown in Figure 2, a 10% decay rate resulted in relatively better model convergence.

The AdaBoundL2 proposed in this paper differs from AdamW, which only decouples weight decay from gradient updates, and LAMB, which is specifically designed for large-batch training. Its core innovation lies in an adaptive smooth transition mechanism via dynamic boundaries, enabling the optimizer to retain the adaptive fast convergence of Adam in the early training stage, and smoothly transition to the robust generalization ability of SGD in the later stage. This characteristic precisely satisfies the dual requirements of rapid feature capture and stable noise robustness for fault diagnosis under strong noise.

To further dissect the actual gains of the internal dynamic

adjustment mechanism of the AdaBoundL2 optimizer and to address the limitation that the original ablation study only compared the presence and absence of the complete optimizer, this section adds three sets of controlled experiments, A, B, and C in Table 3, which examine the upper bound of static optimal hyperparameters, the sensitivity of static hyperparameters, and the independent contributions of the trimming and dynamic coupling modules. These experiments quantify the performance advantage of the dynamic adjustment strategy over fixed hyperparameter schemes and distinguish the degree to which different submodules contribute to model performance. The training results are shown in Figure 3.

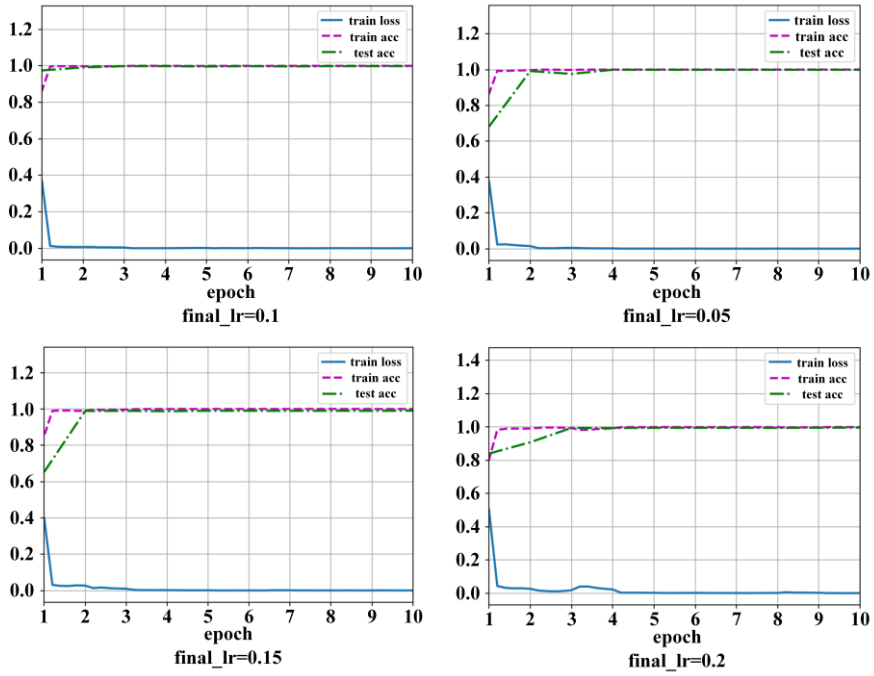


Figure 2. Comparison of training outcomes under different final_lr parameters.

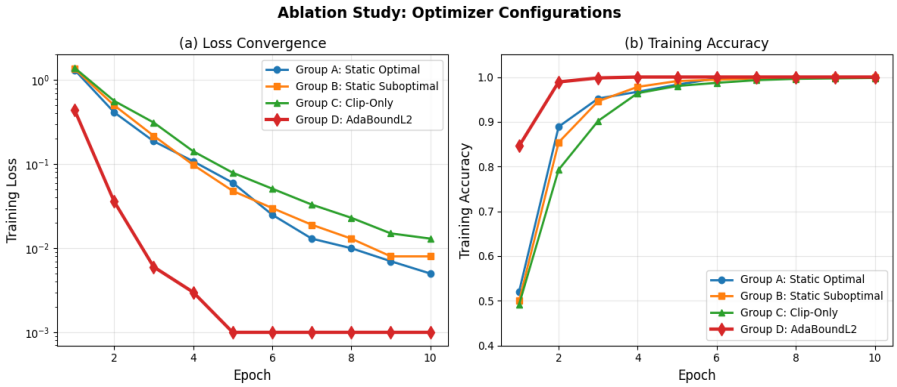


Figure 3. Comparison of training outcomes under different groups.

Table 3. Ablation configurations for analyzing the dynamic adjustment mechanism of AdaBoundL2 optimizer.

Group	Configuration	final_lr	gamma	Purpose
A	Static Optimal	optimal value	optimal value	Upper-bound performance reference
B	Static Suboptimal	0.5× optimal	0.5× optimal	Sensitivity demonstration
C	Clip-Only	optimal value	optimal value	Isolate clipping contribution
D	AdaBoundL2	Adaptive decay	Adaptive decay	Proposed dynamic strategy

2.3. GAP+FC

The Global Average Pooling (GAP) layer and the fully connected layer are essential components of convolutional neural networks, frequently employed in modern architectures like ResNet and Inception. GAP can substitute traditional fully connected layers by performing global averaging on each feature map, thereby producing an output dimension that aligns

with the number of classes, enabling efficient classification connections. The method presented in this paper utilizes a GAP + FC structure, as illustrated in Figure 4. Here, the GAP output y_k represents the average of all elements x_{kpq} in the k feature map, with$|\Re|$denoting the total number of elements in that feature map. This output is then mapped to the classifier via weights$w$and bias$b$.

The global average pooling (GAP) layer compresses feature maps into channel-wise semantic vectors through spatial averaging, thereby reducing parameter dimensionality, introducing translation invariance, and suppressing local noise. The fully connected (FC) layer maps channel semantics to a class-discriminative space, wherein the quality of weight optimization directly determines the robustness of classification boundaries. In contrast to traditional convolutional networks, GAP yields a more compact feature representation without introducing additional parameters, substantially reducing input scale and computational complexity, accelerating training and inference, and proving suitable for resource-constrained environments while mitigating the risk of overfitting in small-sample scenarios. The FC layer adopts a bilinear structure with embedded Dropout for regularization, further enhancing generalization capability. Collectively, these two components establish a dual mechanism characterized by "spatial aggregation for noise reduction and classification constraint for noise resistance" under high-noise conditions, achieving a balance between computational efficiency and diagnostic generalization performance.

Unlike two-dimensional images that require spatial coordinates, one-dimensional fault diagnosis relies on the global spectral energy distribution. Global average pooling (GAP) is selected instead of global max pooling (GMP) because computing the mean over the full time domain acts as an implicit smoothing filter, effectively resisting impulse noise interference in machinery signals (GMP, which is sensitive to peaks, is highly prone to failure in this regard), while preserving translation invariance and mitigating overfitting.

3. Experimental validation

This paper validates the effectiveness of the proposed AdaBoundL2-DSCNN using the CWRU public dataset. Provided by Case Western Reserve University (CWRU) in the United States, this dataset is specifically tailored for research on fault diagnosis and health monitoring of rotat-ing machinery, including motors, pumps, and fans [29].

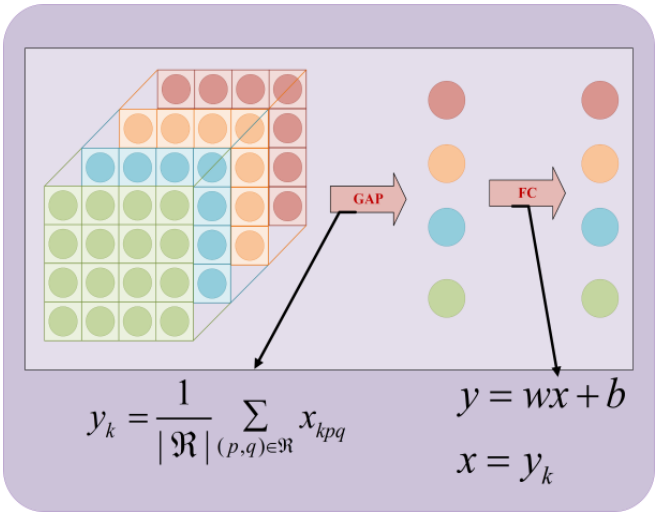


Figure 4. GAP+FC model structure and formulae.

3.1. Datasets

The CWRU (Case Western Reserve University) dataset is a widely utilized standard in rolling bearing fault diagnosis research. Collected via a motor-driven bearing test rig, this dataset employs sensors to capture vibration signals from the equipment under both normal and various fault conditions. During fault simulation, researchers intentionally damaged the inner ring, outer ring, and rolling elements of the bearings, gathering data for diverse fault types across bearings of different sizes (7 inches, 14 inches, and 21 inches). The dataset encompasses samples under various load and speed conditions, ensuring data diversity and representativeness. This collection approach equips the CWRU dataset with abundant standardized fault data, crucial for developing and evaluating fault diagnosis algorithms. This study employs drive-end fault data with a 12 kHz sampling frequency as the experimental focus. Data samples from ten fault states—including normal condition, inner ring fault, outer ring fault, and rolling element fault—are selected, as illustrated in Table 4 and Table 5, to validate the proposed method's effectiveness in fault diagnosis.

Table 4. Normal baseline data.

Motor Load(HP)	Approx Motor Speed(rpm)	Normal Baseline Data
0	1797	Normal_0
1	1772	Normal_1
2	1750	Normal_2
3	1730	Normal_3

Table 5. 12k drive end bearing fault data.

Fault Diameter	Motor Load (HP)	Approx Motor Speed	Inner Race	Ball	ORPLZ
0.007	0	1797	IR007_0	B007_0	OR007_@6_0
	1	1772	IR007_1	B007_1	OR007_@6_1
	2	1750	IR007_2	B007_2	OR007_@6_2
	3	1730	IR007_3	B007_3	OR007_@6_3
0.014	0	1797	IR014_0	B014_0	OR014_@6_0
	1	1772	IR014_1	B014_1	OR014_@6_1
	2	1750	IR014_2	B014_2	OR014_@6_2
	3	1730	IR014_3	B014_3	OR014_@6_3
0.021	0	1797	IR021_0	B021_0	OR021_@6_0
	1	1772	IR021_1	B021_1	OR021_@6_1
	2	1750	IR021_2	B021_2	OR021_@6_2
	3	1730	IR021_3	B021_3	OR021_@6_3

Table 6. SEU dataset

Position	Failure Mood	Status Description
gearbox	Health	Healthy operating condition
	Chipped	The gear has developed cracks
	Miss	The gear has broken teeth
	Root	The gear root has developed cracks
	Surface	The gear surface has worn
Rolling Bearing	Health	Healthy operating condition
	Ball Fault	The rolling element has cracks
	Inner Ring Fault	The inner Ring has cracks
	Outer Ring Fault	The outer ring has cracks
	Combination Fault	Inner rolling and outer rolling have cracks

To validate the effectiveness of the proposed fault diagnosis method, experiments were conducted using the publicly available Southeast University (SEU) mechanical fault dataset. This dataset comprises two subsets: the rolling bearing dataset and the gearbox dataset, both of which have been extensively utilized in rotating machinery fault diagnosis research. In this study, data acquired under the operating condition of 20 Hz rotational speed (1200 rpm) with no-load (0 V, 0 Nm) were selected as the experimental subjects. The rolling bearing dataset contains vibration signals corresponding to various health states, including normal operation and multiple fault conditions affecting different bearing components. The gearbox dataset comprises vibration measurements obtained from a two-stage parallel shaft gearbox under both healthy and faulty conditions. Multiple fault types were introduced to simulate common gearbox degradation patterns encountered in industrial applications. The data samples are detailed in Table 6. To further evaluate the robustness of the model when facing non-stationary real industrial noise containing specific frequency components, we conducted additional validation experiments on the Paderborn University (PU) bearing dataset, which contains real vibration signals acquired under actual operating conditions. These tests complement the synthetic-noise-based evaluation and provide more compelling evidence for the generalization capability of the proposed method in practical industrial scenarios, as shown by the experimental results in Figure 7.

3.1.1. Lightweight model architecture design

To achieve lightweight design and reduce computational complexity, this paper proposes a three-level spatial feature compression mechanism based on depthwise separable convolution. At the first level, depthwise separable convolution decouples standard convolution into spatial and pointwise convolutions: spatial convolution applies 3×3 filters independently to each channel, extracting spatial features while avoiding redundant cross-channel computation; pointwise convolution performs linear combinations across channels using 1×1 convolutions, enabling cross-channel interaction and preliminary compression. At the second level, global average pooling (GAP) spatially reduces each channel's feature map to a single scalar. Compared to flattening before a fully connected layer, GAP reduces parameter count while preserving the overall response strength of each channel, thereby achieving efficient feature aggregation. This three-level compression strategy realizes progressive reduction from local feature extraction and channel recombination to global aggregation, laying the structural foundation for the lightweight design.

3.1.2. Dynamic adaptive learning strategy

Hyperparameters include batch size 64, 100 epochs, and weight decay 5×10^{-4}. Grid search yields initial_final_lr=1×10^{-4} and initial_gamma=0.99. The dynamic strategy achieves F1=0.989 under -8dB noise, outperforming fixed variants (0.972–0.975) and reducing convergence by 20 epochs.

The improved method significantly outperforms the baseline in both training efficiency and prediction accuracy. Training time is reduced by 94.4%, achieving a 17.9× speedup. For regression tasks, the mean squared error is reduced by 82.6%, root mean squared error by 58.3%, and mean absolute error by 73.3%. For classification, the F1-score increases from 0.9945 to 0.9980. These results demonstrate that the improved method

substantially enhances computational efficiency and regression accuracy while maintaining near-perfect classification performance.

3.2. Compared approaches and implementation details

To validate the proposed method, comparative experiments were conducted on the Case Western Reserve University (CWRU) bearing fault dataset and the Southeast University (SEU) gearbox dataset. An overlapping sampling strategy was applied with a step size of 28 (overlap rate of 98.6%).To avoid potential overlap of test accuracy caused by extremely high sample similarity under high-overlap sampling, experiments under multiple sampling overlap rates (0%, 50%, 70%, 98.6%) were further carried out, and the corresponding results are reported in Table 7. Collecting 1,000 samples per class at a sampling frequency of 12.8 kHz. The dataset was randomly split into training, validation, and test sets in a 7:2:1 ratio, preserving class balance. Each sample was independently normalized via Z-score standardization to achieve zero mean and unit variance, eliminating the influence of amplitude variations across operating conditions. After preprocessing, the data were reshaped to (N, 2048) and then expanded to (N, 1, 2048) to fit the input format of a one-dimensional convolutional neural network, with a batch size of 128. Four representative fault diagnosis models were selected as benchmark methods:

1) Simplified WDCNN Model: This model uses five layers of depthwise separable convolution to build a lightweight framework. It reduces parameters and computational cost by decomposing standard convolution. Its advantages are simple structure and high efficiency, but its limited capacity makes it less effective for complex fault patterns. It serves as a baseline to set a lower bound for lightweight models.
2) WDCNN Model: This model is based on a standard one-dimensional convolutional neural network [30]. It extracts hierarchical features by stacking convolutional and pooling layers, making it a classic baseline in signal processing. Its structure is general and reliable, but it lacks task-specific optimization. It is used to establish a performance benchmark for standard CNN architectures.
3) ResCNN Model: This model introduces residual connections [31] to address gradient vanishing in deep networks, allowing deeper architectures to extract more discriminative features. It offers high accuracy potential but has a more complex structure. It serves as a baseline to examine the effect of network depth.
4) PaperRepo Model: This model reproduces the multi-layer convolutional network architecture described in reference [22], extracting deep features by stacking one-dimensional depthwise separable convolution layers, batch normalization layers, and spatial dropout layers, while employing a pruning mechanism to reduce network parameters; health state diagnosis is ultimately performed via a Softmax classifier. The original hyperparameter configuration from that work is adopted in this study, with this state-of-the-art scheme serving as the comparison baseline to ensure the fairness of the comparative experiments.

Controlled variable experiments are designed to evaluate performance in terms of diagnostic accuracy, model complexity, and computational efficiency. All comparative experiments adhere to the same training/validation/test set division (7:2:1 ratio) and data preprocessing procedures to ensure fair comparison.

Table 7. Model performance under different sampling overlap rates.

overlap	Step	F1-score	MSE	RMSE	MAE	Training_ Time
0%	2048	0.9950	0.1800	0.4243	0.0300	124.6486
50%	1024	0.9940	0.2160	0.4648	0.0360	63.8455
70%	614	0.9985	0.0110	0.1049	0.0040	96.2461
98.6%	28	0.9970	0.0945	0.3074	0.0165	106.6739

Figure 5 presents a comparative analysis of training performance over 10 epochs. The experimental results reveal that the proposed AdaBoundL2 optimizer achieves notably faster convergence and enhanced training stability with respect to both accuracy metrics and loss functions. The fault diagnosis accuracy comparison is depicted in Figure 8, which demonstrates that the AdaBoundL2-DSCNN model consistently achieves superior classification performance across all diagnostic tasks, significantly outperforming competing methods under equivalent experimental settings.

The proposed AdaBoundL2-DSCNN and comparison models are implemented using PyTorch 2.0 with Python 3.10. The training and inference of these models are conducted on

a standard workstation equipped with an Intel Core i7–9700K CPU and an Nvidia RTX2080 (8GB) GPU. The macro F1-score is used as the evaluation metric, defined as follows:

$$precision_k = \frac{TP}{TP+FP} \tag{9}$$

$$recall_k = \frac{TP}{TP+FN} \tag{10}$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

$$F1_k = \frac{2 \cdot precision_k \cdot recall_k}{precision_k + recall_k} \tag{12}$$

$$score = (\frac{1}{n}\sum F1_k)^2 \tag{13}$$

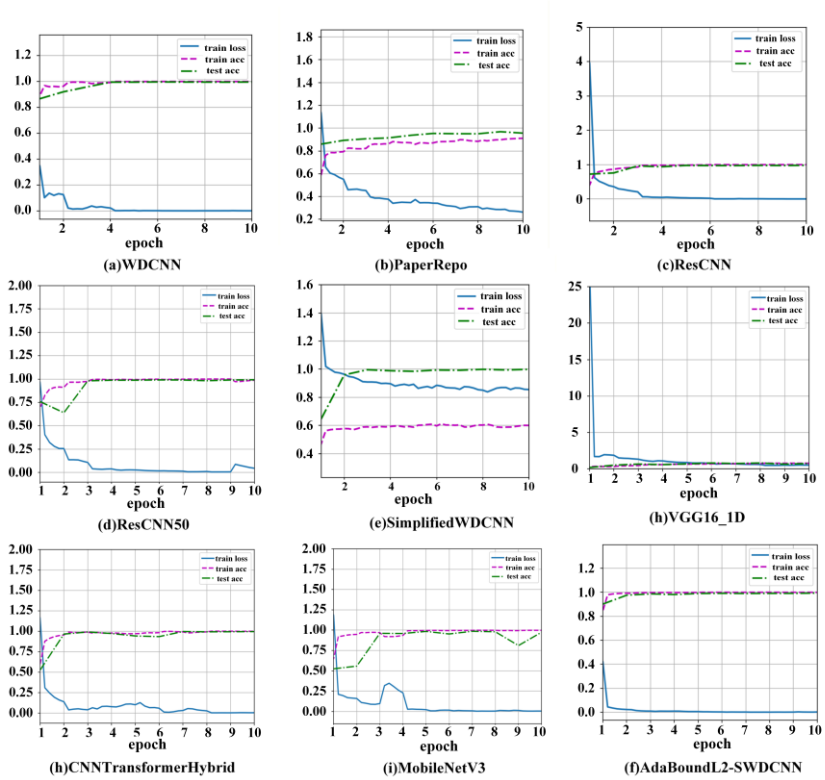


Figure 5. Model training accuracy and loss profile on CWRU.

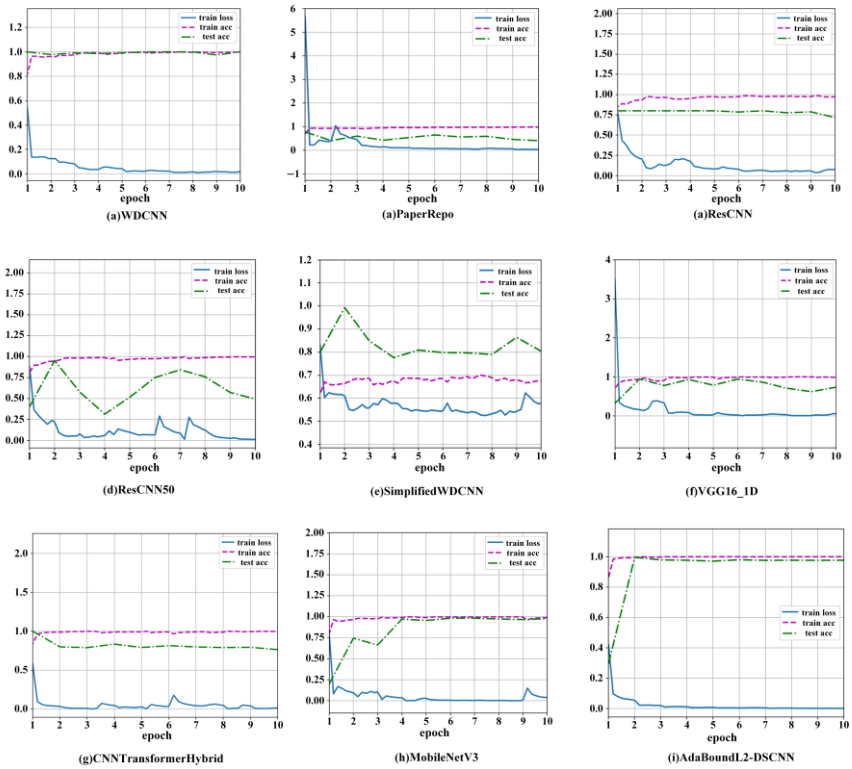


Figure 6. Model training accuracy and loss profile on SEU.

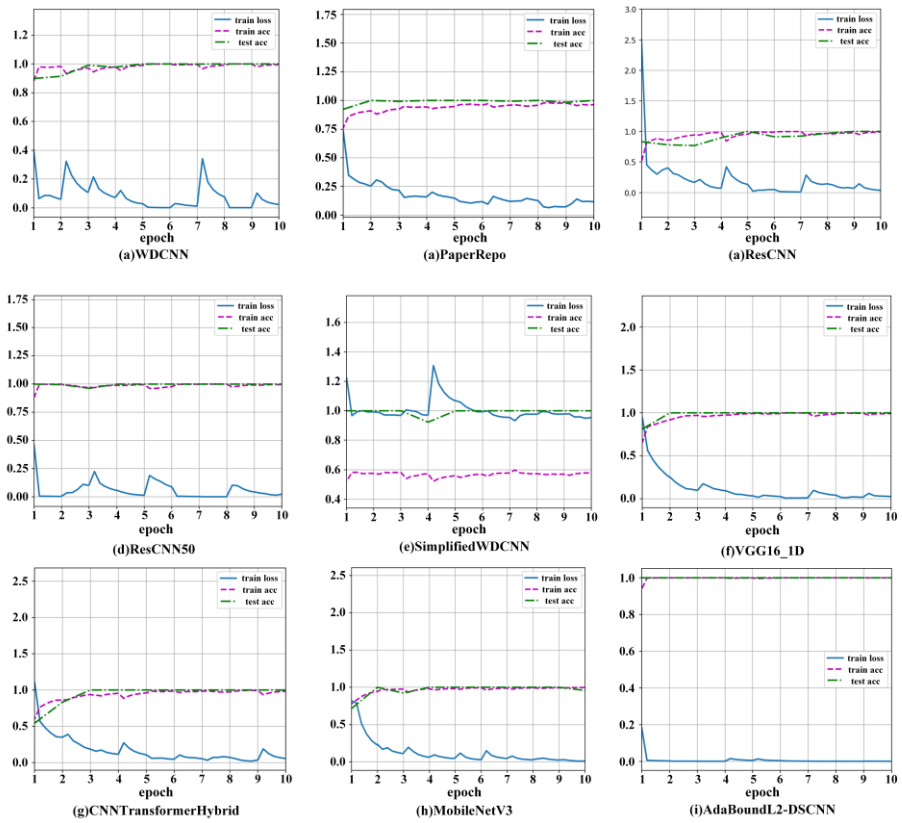


Figure 7. Model training accuracy and loss profile on PU.

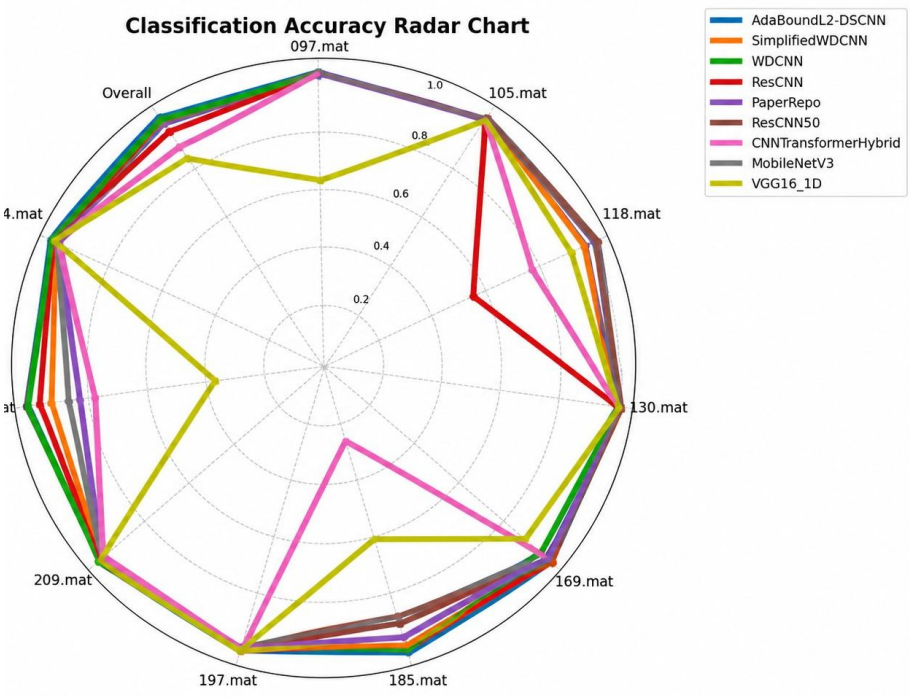


Figure 8. Rader chat of fault diagnosis tasks accuracy.

The F1-score ranges from 0 to 1, with values closer to 1 indicating better recognition performance. An optimizer using an adaptive constraint matrix (AdaBoundL2) is employed to optimize the network's weighted cross-entropy loss. Each set of experiments is conducted using five-fold cross-validation [32], where the original dataset is divided into five parts. For each division, training and evaluation are performed once, and the evaluation results from the five divisions are averaged to report the final classification result. The results of the five-fold cross-validation are shown in Figure 9.

Five-fold cross-validation involves multiple rounds of training and testing, allowing each sample to be part of both the training set and the test set. This approach provides a more comprehensive evaluation of the model's performance and avoids the bias that can result from a single partition. Each test uses 4/5 of the data as the training set and 1/5 as the test set. This method maximizes data utilization and enhances training efficiency. Evaluating the model on different subsets, prevents optimization on a specific partition alone, thereby reducing the risk of overfitting.

To evaluate the generalization capability of the model under different operating conditions, we designed three cross-load experiments with progressively increasing difficulty: same-load (0HP, 7:2:1 split), cross-load (training on 0/1/2HP → testing on 3HP), and extended cross-load (training on 0/1HP → testing on 2/3HP).

Table 8. Definitions of fault diagnosis evaluation metrics.

Related Nouns	Texture density (%)
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
Precision	The classifier determines the weight of positive samples in positive examples
Recall	Accuracy of the model in predicting a positive class
Accuracy	The proportion of the whole sample judged correctly by the classifier
$F1_k$	F1-score under each category
score	The mean value of F1-score under each category

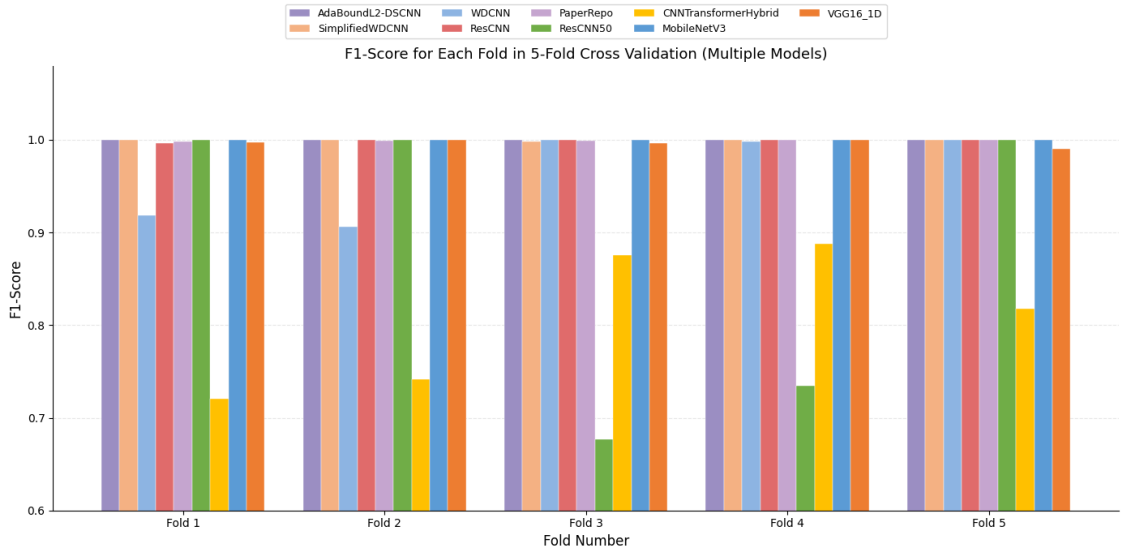


Figure 9. Histogram of 5-fold cross-validation results.

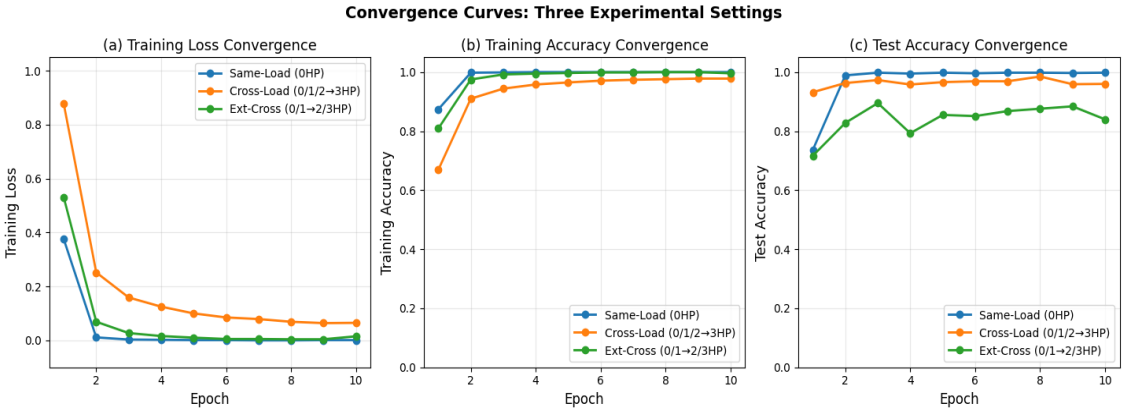


Figure 10. Cross-load generalization performance under three experimental settings.

As shown in Fig. 10, our method achieves a test accuracy of 96.0% on the unseen 3HP load, which is only 3.8% lower than the same-load baseline (99.8%). Even under the more challenging extended cross-load setting, the model maintains an accuracy of 84.0%. These results indicate that our method possesses strong domain adaptation capability and remains robust under significant load variations, demonstrating its feasibility for deployment in real industrial scenarios, even

when the target operating condition is unseen during training.

3.3. Anti-noise performance analysis

To assess the noise resistance of the developed AdaBoundL2-DSCNN, Gaussian white noise, Laplace noise, and pink noise were introduced to the original signal at low signal-to-noise ratios (SNR) for model training and evalua-tion. The proposed AdaBoundL2-DSCNN model and com-parison models were tested with noise signals at −8 dB, −6 dB, −4 dB, −2 dB, and 0 dB. The vibration data image after adding Gaussian white noise at an SNR of −8 dB is shown in Figure 11.

3.3.1. Bearing fault diagnosis at different SNRs

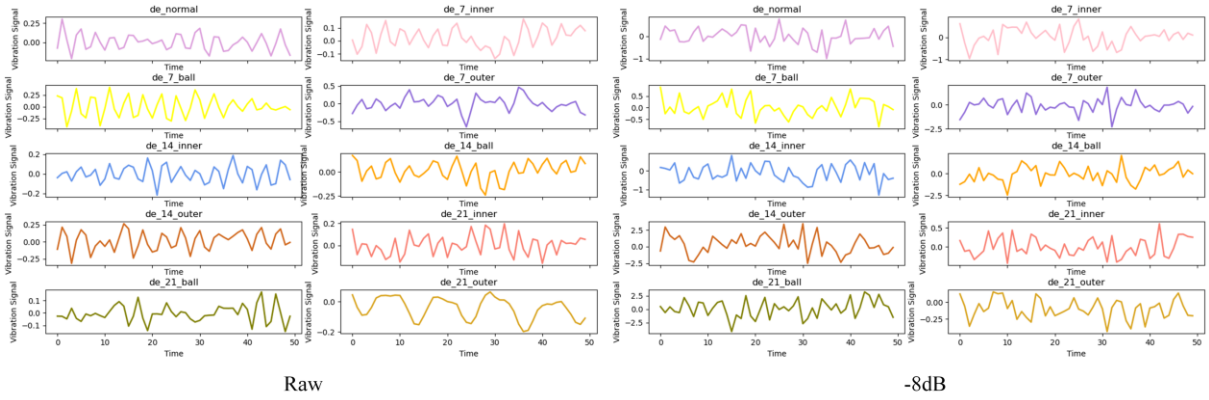


Figure 11. Examples of raw bearing signals with white Gaussian noise under different SNRs.

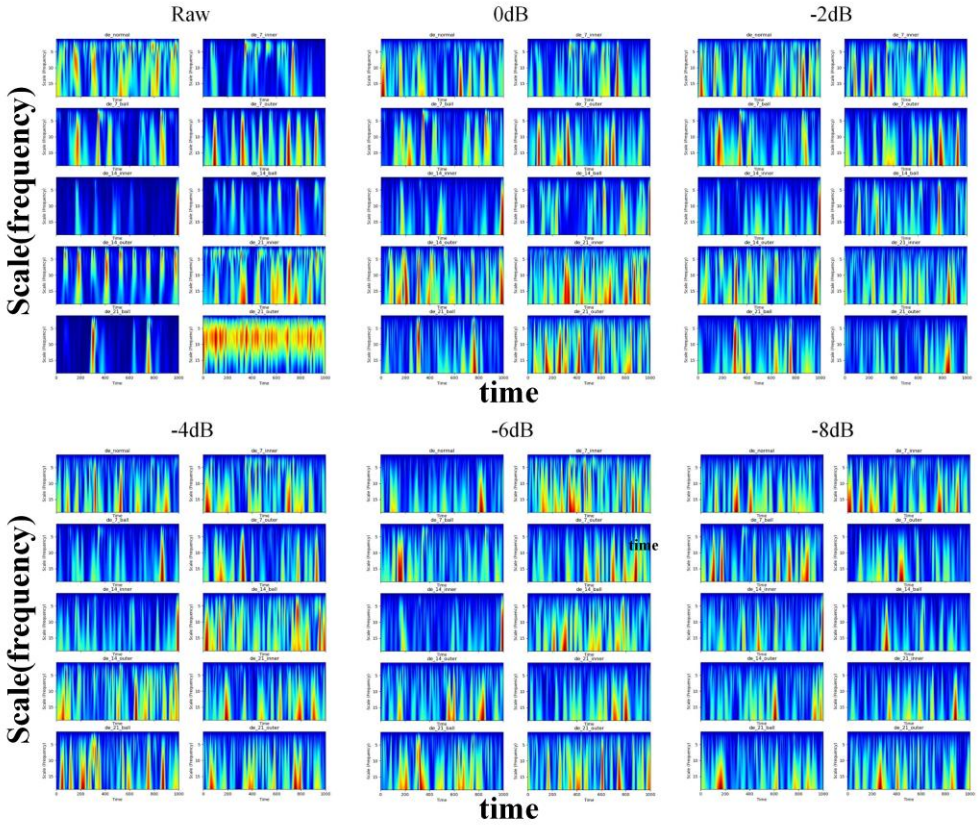


Figure 12. CWT time-frequency analysis of bearing signals examples.

Figure 8 shows an example of the original bearing signal with Gaussian white noise under a -8dB signal-to-noise ratio (SNR) condition. The amplitude characteristics of the signal in the fault state exhibit significant differences. However, as the SNR decreases, distinguishing the key features between normal states and different fault types becomes increasingly difficult. To comprehensively analyze the impact of noise on time-domain signals, the continuous wavelet transform

(CWT) [33] was used to generate time-frequency diagrams of these noisy signals, as shown in Figure 12. The results indicate that the addition of noise, particularly high-frequency noise, significantly blurs the details of fault characteristics, making it gradually difficult to discern the fault-related information in signals from different health states. The F1-score results for bearing fault diagnosis using the original data are shown in Figure 11. The detailed results of different methods for bearing fault diagnosis using noisy signals are also shown in Figure 13, with the best performance highlighted in bold. The proposed method performs the best, with an average F_mac of 95.71% under Gaussian white noise. The ResCNN model failed to accurately classify health conditions in the bearing diagnosis task, with an average F_mac of 78.54% under Gaussian white noise.

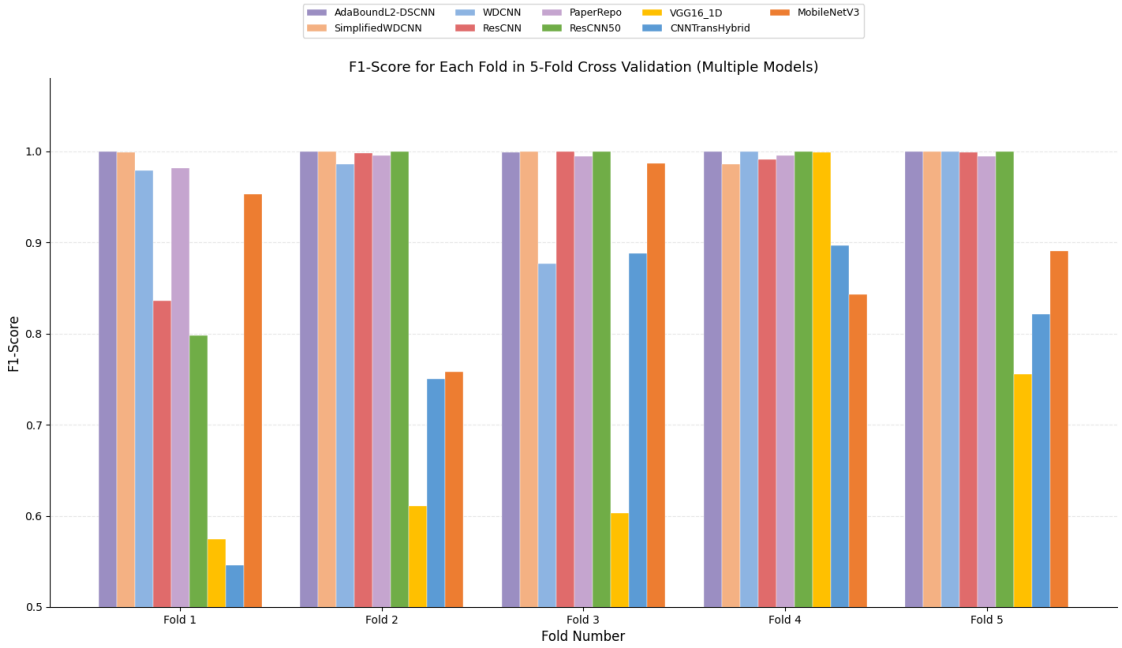
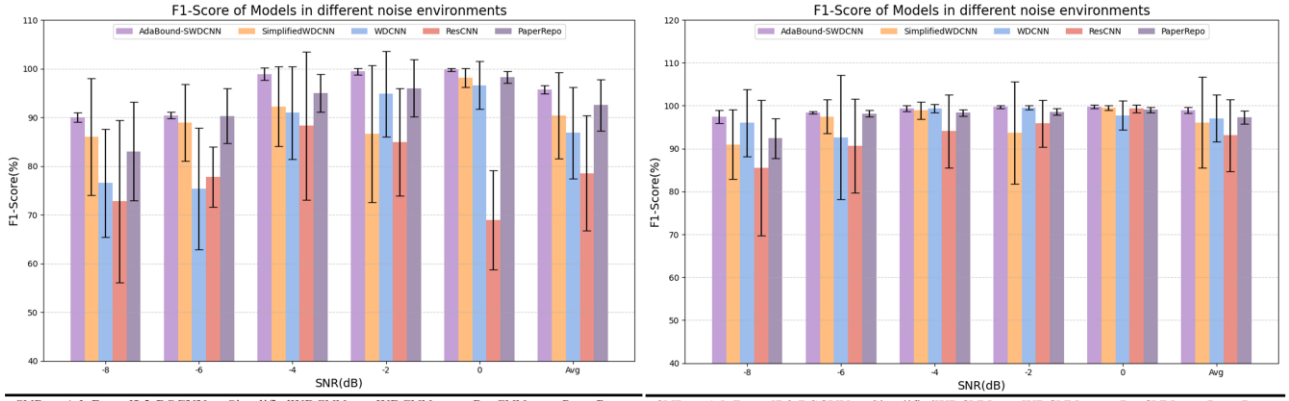


Figure 13. F1-score for each fold in 5-fold cross-validation (multiple models).

In the experiment shown in Figure 14(a), all data were processed with fewer iterations (epoch=3). This setting allows us to observe the model's convergence in a short period. The experimental results indicate that the proposed model can achieve relatively ideal performance with fewer training iterations, and its convergence speed is significantly faster compared to other models. This phenomenon demonstrates that the model proposed in this paper can effectively learn the latent features of the data early in the training process, achieving high accuracy with fewer training cycles and showing superior convergence efficiency and training speed compared to other models. Figure 14(b) illustrates the training performance of various models after 12 iterations (epoch=12). The experimental results indicate that AdaBoundL2-DSCNN significantly outperforms other models, achieving an average F_mac of 98.94% under Gaussian white noise. When the SNR decreases from 0 dB to -8dB, the average F_mac under Gaussian noise for SimplifiedWDCNN, WDCNN, ResCNN, and PaperRepo decreases by 8.57%, 5.14%, 13.78%, and 6.62%, respectively. In contrast, the evaluation metric for AdaBoundL2-DSCNN only decreases by 2.38%. The proposed AdaBoundL2-DSCNN achieves extremely low p-values in both Levene's test ($1.03e\times10^{-317}$) and pairwise t-tests (2.38×10^{-286}), confirming its superior fine-grained discriminative capability and feature separability for similar faults. Figure 15 shows

the confusion matrices of different methods on the orientation dataset under -8dB Gaussian noise. The columns of the matrix represent the true labels of health conditions, while the rows represent the labels predicted by the model. From the figure, it is evident that the method proposed in this paper excels in fault diagnosis accuracy, significantly outperforming other comparative models. This advantage is reflected not only in the overall improvement of diagnostic accuracy but also in maintaining a high degree of consistency across various diagnostic tasks. Whether in assessing health conditions or identifying various fault modes, the method in this paper demonstrates strong robustness and balance. The stability of this performance further confirms the superiority and practical value of the method proposed in this paper for complex noise environments, offering a reliable solution for fault diagnosis in high-noise industrial settings.



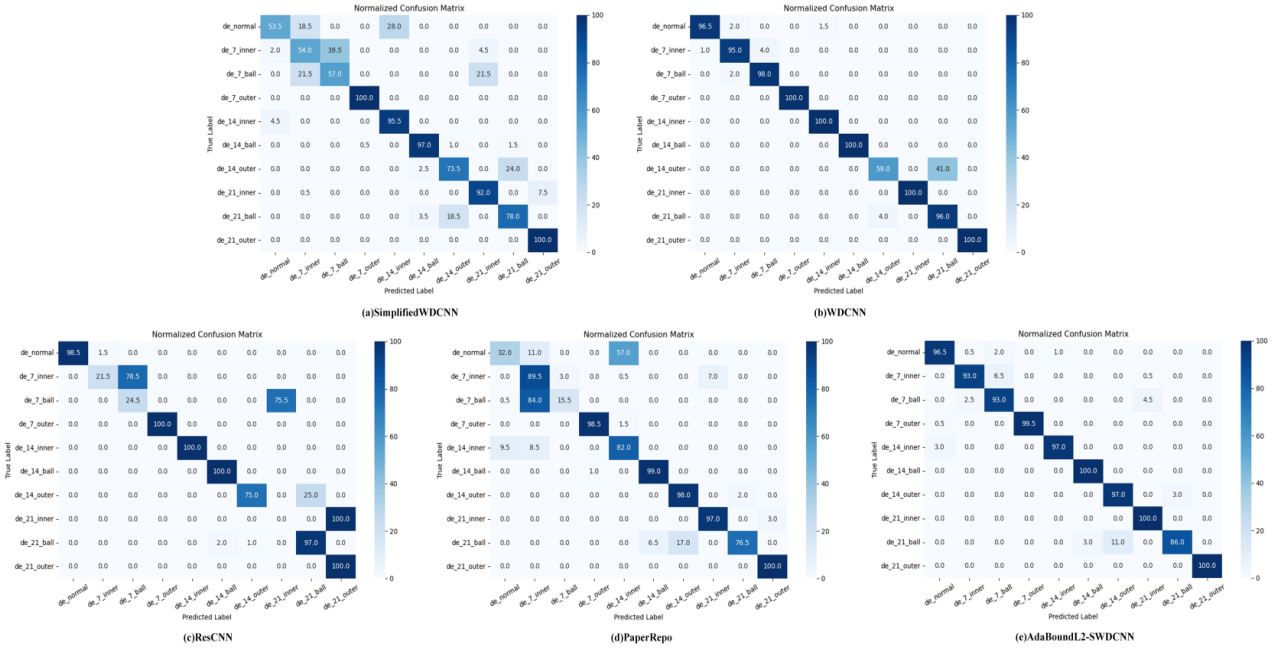
SNR	AdaBoundL2-DSCNN	SimplifiedWDCNN	WDCNN	ResCNN	PaperRepo
-8dB	**90.00±0.96**	86.02±12.01	76.50±11.12	72.76±16.70	83.02±10.11
-6dB	**90.42±0.69**	88.94±7.85	75.38±12.50	77.78±6.16	90.30±5.60
-4dB	**98.92±1.26**	92.21±8.18	90.96±9.53	88.27±15.18	94.98±3.86
-2dB	**99.42±0.71**	86.62±14.05	94.84±8.87	84.95±11.05	95.98±5.81
0dB	**99.77±0.33**	98.17±1.96	96.59±4.90	68.94±10.15	98.24±1.20
Avg	**95.71±0.79**	90.39±8.18	86.85±9.40	78.54±11.85	92.50±5.33

(a)Gaussian noise (epoch=3)

SNR	AdaBoundL2-DSCNN	SimplifiedWDCNN	WDCNN	ResCNN	PaperRepo
-8dB	**97.39+1.53**	90.97+8.13	95.99+7.81	85.53+15.83	92.41+4.61
-6dB	**98.42+0.27**	97.46+3.93	92.61+14.46	90.64+10.99	98.17+0.73
-4dB	**99.40+0.72**	98.89+2.07	99.40+1.00	94.07+8.53	98.40+0.76
-2dB	**99.74+0.34**	93.65+11.89	99.56+0.44	95.87+5.44	98.55+0.76
0dB	**99.77+0.42**	99.54+0.54	97.75+3.43	99.31+0.93	99.03±0.60
Avg	**98.94+0.66**	96.10+10.62	97.06+5.43	93.08+8.34	97.31+1.49

(b)Gaussian noise (epoch=12)

Figure 14. Diagnosis performance for the bearing dataset.



(a)SimplifiedWDCNN (b)WDCNN

(c)ResCNN (d)PaperRepo (e)AdaBoundL2-SWDCNN

Figure 15. Confusion matrices for bearing fault diagnosis under −8 dB Gaussian noise.

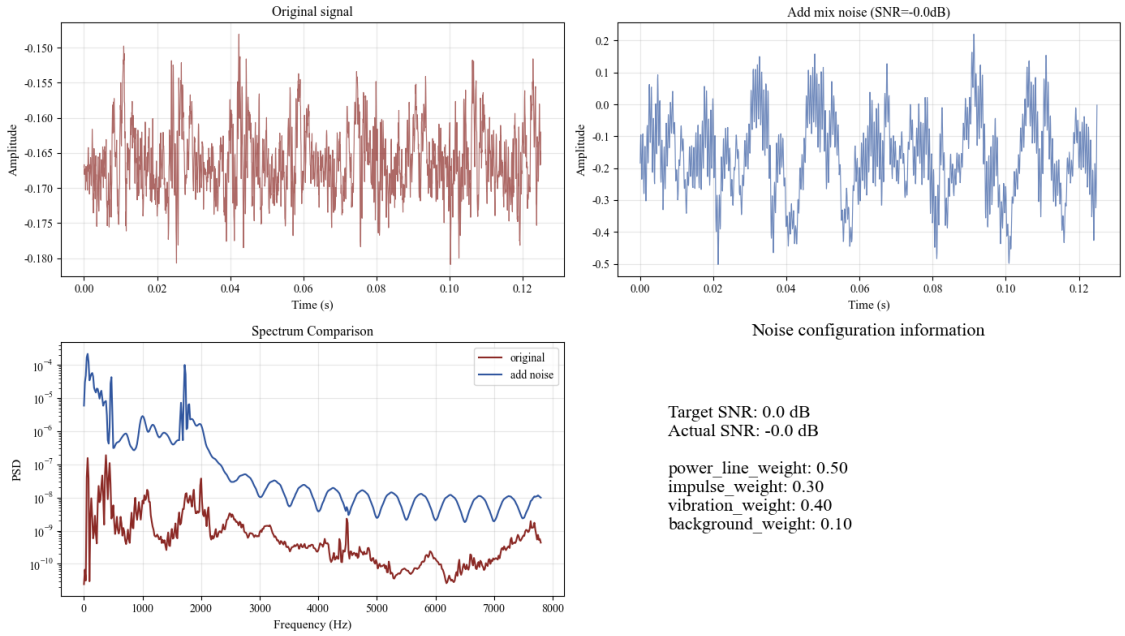


Figure 16. SEU dataset with multi-source aliased noise.

Table 9. Computational efficiency comparison of different methods.

Models	Total training time(s)	MSE	RMSE	MAE	F1-score
LiConvFormer	84.23	0.9590	0.9793	0.4850	0.6867
EverAdapt	72.95	1.8770	1.3700	0.7570	0.6510
SqueezeNet_DLCNN	85.61	1.2640	1.1243	0.6320	0.6067

To validate the effectiveness of AdaBoundL2-DSCNN in fault detection (FD), multi-source aliased noise was added to the vibration signals from the Southeast University gearbox dataset, as illustrated in Figure 16. The signal-to-noise ratio (SNR) was set to 0 dB, where the noise components included power line interference, random pulses, vibration aliasing, and background noise. The weighting coefficients of each noise component were configured as follows: power line interference = 0.5, random pulses = 0.3, vibration aliasing = 0.4, and background noise = 0.1. Two representative lightweight models proposed in recent literature, LiConvFormer (2025), EverAdapt (2025), and SqueezeNet_DLCNN (2025) [34–36], were employed as benchmark models for performance comparison. A quantitative evaluation was conducted to assess the training accuracy and noise robustness of all models under complex noise conditions, and the evaluation results are shown in Table 9. All comparative methods are implemented following their original code or paper architectures with consistent parameter settings. Results are evaluated under heavy-noise conditions to assess algorithm robustness in extreme industrial environments.

3.3.2. Analysis of anti-noise experimental results

As mentioned above, this study evaluated the performance of the compared methods under different noise interference conditions using a dataset. The radar chart in Figure 13 clearly illustrates the differences in accuracy for various diagnostic tasks under multiple noise interferences. All experiments were conducted with a fixed epoch of 3. With a relatively small number of iterations, the AdaBoundL2-DSCNN method proposed in this paper demonstrated significant advantages, achieving rapid convergence, high test accuracy, and low loss values. Compared with the previous charts, it is evident that after adding noise during the training process, the accuracy of other comparative models

generally declined significantly. However, the AdaBoundL2-DSCNN maintained high accuracy, and the accuracy in diagnostic tasks under different fault conditions remained stable without significant differences.

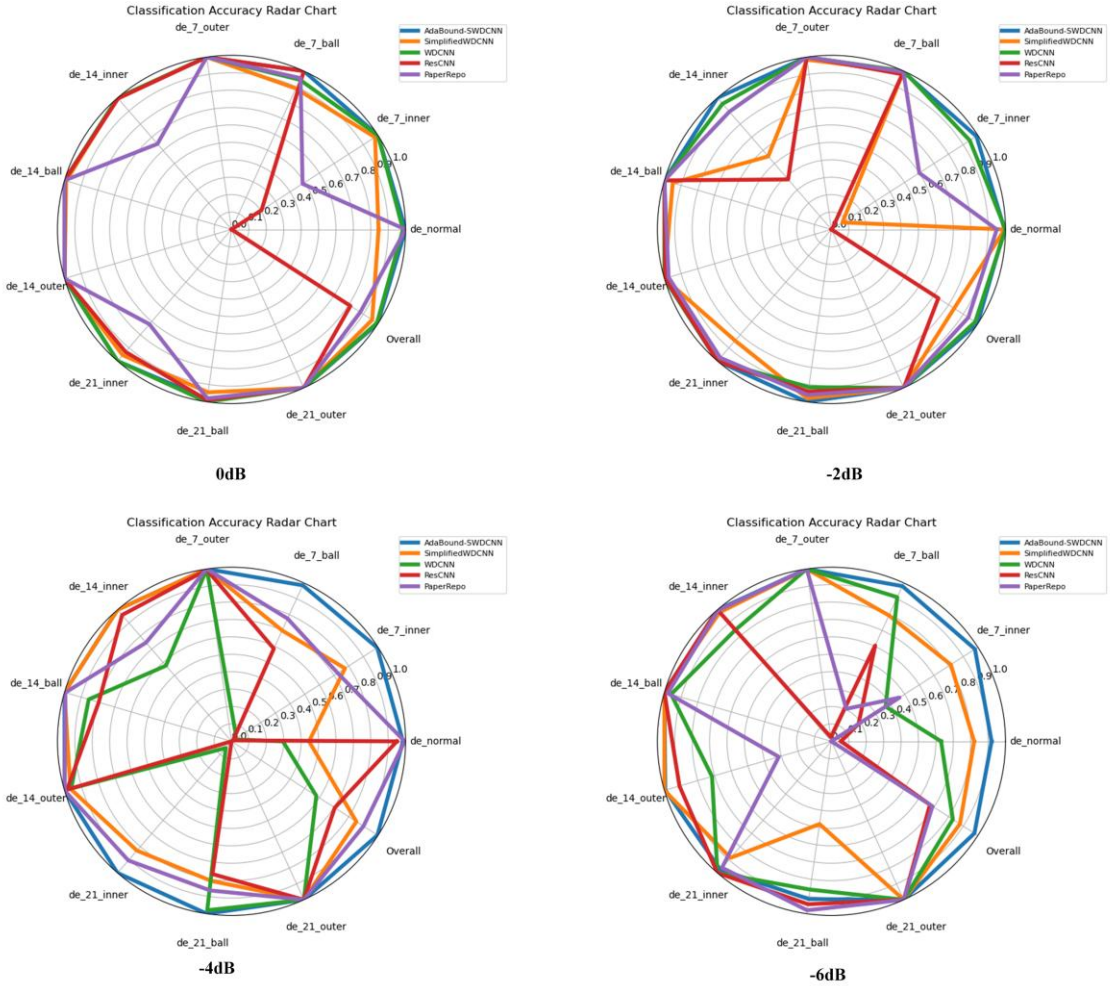


Figure 17. Radar charts of the fault diagnosis results under different Gaussian noise.

The results in Figure 17 clearly show that the evaluation metrics of the proposed AdaBoundL2-DSCNN method consistently occupy the outermost region of different diagnostic tasks, further proving that the proposed method has superior health state recognition capabilities compared to other methods under high noise interference conditions. Specifically, these results indicate that AdaBoundL2-DSCNN can effectively capture long-term dependencies in the signal and eliminate redundant information, thereby significantly reducing the impact of noise interference and ensuring accurate identification of health state types in complex environments. In this way, the proposed model not only performs excellently in theory but also demonstrates strong robustness and practicality in real-world applications, particularly in high-noise conditions, offering more reliable predictive performance for diagnosing various faults.

3.4. Computational efficiency analysis

To evaluate the computational efficiency of various methods, Table 10 presents several efficiency measurements for the dataset diagnostic tasks. As demonstrated in Table 10, AdaBoundL2-DSCNN achieves the lowest peak memory (0.588 GB) with competitive inference speed, offering optimal edge-deployment efficiency. FLOPs, which denote the number of floating-point operations required per second,

are commonly used to assess a model's computational complexity. In the table, AdaBoundL2-DSCNN has a FLOP count of 0.29 Mmac, whereas other models like WDCNN, ResCNN, and PaperRepo have FLOPs of 0.75 Mmac, 1.78 Mmac, and 1.54 Mmac, respectively. This indicates that AdaBoundL2-DSCNN has the lowest FLOPs, signifying the highest computational efficiency and the least demand for computational resources, making it ideal for environments with limited computational capabilities. The term "Params" refers to the number of trainable parameters in the model, which directly impacts the model's complexity and the storage and computational resources needed for training and inference. AdaBoundL2-DSCNN has 12.77k parameters, significantly fewer than ResCNN's 260.96k, WDCNN's 54.31k, and PaperRepo's 12.3k. Despite having fewer parameters, AdaBoundL2-DSCNN's performance remains robust, maintaining high efficiency while delivering strong performance. The reduced number of parameters enhances the model's training speed and decreases memory consumption, making it particularly suitable for resource-constrained scenarios. FPs measure how many frames a model processes per second on a given hardware, with higher values indicating faster performance. In terms of inference speed, WDCNN and ResCNN are in the lead with 1166.52 fps and 1160.62 fps, respectively. The inference speed of AdaBoundL2-DSCNN is 669.51 fps, which is relatively lower. Despite this, its extremely low computational complexity and parameter count allow AdaBoundL2-DSCNN to achieve efficient real-time inference, making it suitable for applications requiring rapid response. Total Params indicate the total number of parameters in the model, directly affecting the difficulty of training and deployment. AdaBoundL2-DSCNN has a total of 12,774 parameters, significantly lower than ResCNN's 260,960, indicating reduced memory usage and storage requirements. In comparison, WDCNN and PaperRepo have total parameters of 54,310 and 12,299, respectively, highlighting AdaBoundL2-DSCNN's advantage in model size and storage needs. EFS and AFS represent the estimated and actual file sizes, impacting the model's storage requirements. AdaBoundL2-DSCNN has an EFS and AFS of 0.05MB and 0.06MB, respectively, which are significantly lower than those of other models. This indicates that AdaBoundL2-DSCNN excels in storage space and memory usage, making it highly suitable for embedded systems and low-storage environments.

Table 10. Computational efficiency comparison of different methods.

Models	peak memory(GB)	inference time	Flops(Mmac)	Params(k)	FPs(frames/s)	Total Params	EFS(MB)	AFS(MB)
WDCNN	0.607	0.0016s	0.75	54.31	1166.52	54310	0.21	0.22
ResCNN	0.615	0.0008s	1.78	260.96	1160.62	260960	1.00	1.01
PaperRepo	0.632	0.0028s	1.54	12.3	789.99	12299	0.05	0.06
AdaBoundL2-DSCNN	0.588	0.0015s	0.29	12.77	669.51	12774	0.05	0.06

Table 11. Design parameters of the surface textures.

Layer	Output Shap	Value
Input	(batch_size, 1, 2048)	0
convolution	(batch_size, 16, 128)	1072
Batch normalization	(batch_size, 16, 128)	0
Depthwise separable convolution	(batch_size, 32, 128)	672
Batch normalization	(batch_size, 32, 128)	0
Depthwise separable convolution	(batch_size, 64, 128)	2368
Batch normalization	(batch_size, 64, 128)	0
Depthwise separable convolution	(batch_size, 64, 128)	576
Batch normalization	(batch_size, 64, 128)	0
convolution	(batch_size, 64, 128)	576
Batch normalization	(batch_size, 64, 128)	0
Global Average Pool	(batch_size, 64, 1)	0
Fully Connected1	(batch_size, 64, 100)	6500
Fully Connected2	(batch_size, 100, 10)	1010

Although the inference frame rate of AdaBoundL2-DSCNN on a desktop GPU (669.51 FPS) is lower than that of WDCNN and ResCNN, this is a common phenomenon among lightweight depthwise separable convolutional models on high-computing-power platforms. Its core advantage of achieving comparable diagnostic accuracy with extremely low parameter counts and FLOPs makes it more suitable for resource-constrained edge deployment scenarios, such as aerospace equipment.

The AdaBoundL2-DSCNN network structure and the number of network parameters at each layer are specified in Table 11.

3.5. Training ratio analysis

The scarcity of training samples significantly hinders the advancement of diagnostic technology based on AdaBound2 depthwise separable convolutional neural networks in practical applications. To further assess the effectiveness and robustness of the proposed AdaBoundL2-DSCNN method, comparative experiments were conducted using different training ratios in stratified ten-fold cross-validation. The bearing fault diagnosis task under -8dB Laplacian noise was selected as the test benchmark. The number of test samples is consistent with Section 3.3.1. The experimental results are illustrated in Figure 18, where each data point represents the F1-score at the corresponding training ratio. The results demonstrate that AdaBoundL2-DSCNN can achieve an F1-score of over 0.7 with a training ratio of 0.1. The training outcomes are significantly superior to the comparative models, and there is potential for rapid improvement in subsequent training sessions. The AdaBoundL2-DSCNN model maintains satisfactory diagnostic performance even at extremely low training rates, while other comparative methods exhibit a marked decline in performance when the training rate is reduced from 50% to 10%.

Experimental results under other noise conditions are shown in Table 12.

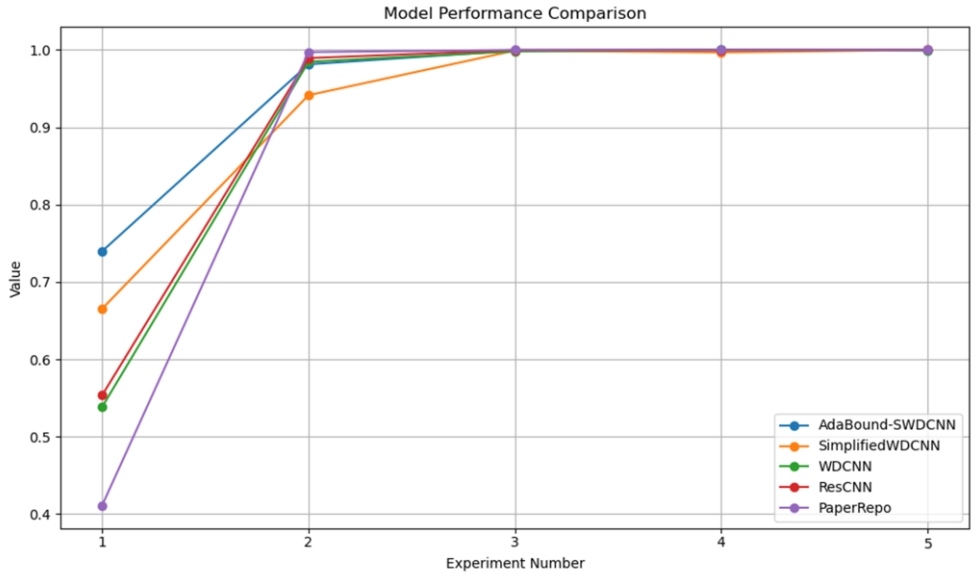


Figure 18. F1-score of models at different training ratios.

Table 12. Experimental results under other noise conditions.

SNR	AdaBoundL2-DSCNN	SimplifiedWDCNN	WDCNN	ResCNN	PaperRepo
0dB	**0.882442**	0.876008	0.848841	0.852136	0.836415
-2dB	**0.848479**	0.772196	0.829331	0.847194	0.816083
-4dB	**0.807742**	0.793594	0.791950	0.769499	0.785536
-6dB	**0.785386**	0.727798	0.699044	0.713733	0.701676
-8dB	**0.726987**	0.655592	0.688996	0.719401	0.686949

3.6. Ablation experiments

To further verify the impact of the AdaBoundL2 filter and the depthwise separable convolutional neural network on fault diagnosis under severe noise levels in the proposed model, a comparative study was conducted to confirm the necessity and effectiveness of each component. The specific configurations for gearbox diagnosis tasks under varying Laplacian noise are as follows:

(1) Model 1: The AdaBoundL2 optimizer is removed.

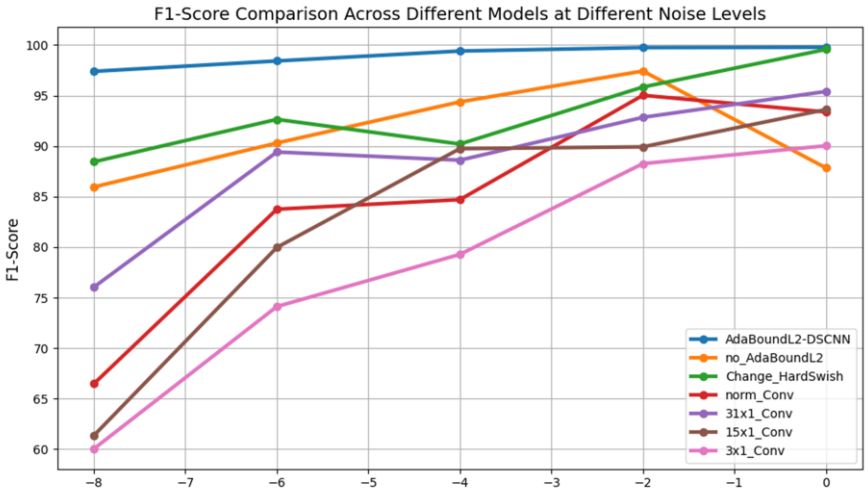
(2) Model 2: The Hard-Swish activation function is replaced with the ReLU activation function.

(3) Model 3: DW-CConvs are replaced with regular continuous convolutions.

(4) Model 4: The large 64*1 convolutional kernel in the first layer is replaced with a smaller 31*1 convolutional kernel, without any continuous kernel parameterization.

(5) Model 4: The large 64*1 convolutional kernel in the first layer is replaced with a smaller 15*1 convolutional kernel, without any continuous kernel parameterization.

(6) Model 4: The large 64*1 convolutional kernel in the first layer is replaced with a smaller 3*1 convolutional kernel, without any continuous kernel parameterization.



SNR	AdaBoundL2-DSCNN	No_AdaBoundL2	Change_HardSwish	Norm_Conv	31×1_Conv	15×1_Conv	3×1_Conv
0dB	**97.49**	85.95	88.44	66.48	76.04	61.36	60.04
-2dB	**98.42**	90.30	92.64	83.75	89.41	79.97	74.12
-4dB	**99.40**	94.37	90.20	84.70	88.60	89.74	79.28
-6dB	**99.74**	97.42	95.84	95.02	92.84	89.92	88.26
-8dB	**99.77**	87.84	99.56	93.37	95.40	93.61	90.03

Ablation experiments epoch=12

Figure 19. Testing macro F1-scores with limited training ratios of different methods on dataset under different SNRs white Gaussian noise.

Figure 19 shows the results of the previously mentioned model and the proposed AdaBoundL2-DSCNN for fault diagnosis tasks under Gaussian white noise. The results clearly indicate that the 3×1 depthwise separable convolution significantly affects the noise resistance performance of the diagnostic model. Model 4 achieved an average test macro F1-score of only 60.04% under -8dB Gaussian white noise interference, demonstrating that large convolution kernels (such as 64×1) can capture a broader range of features, specifically long-term dependencies. This is particularly important in signal sequences where the effective information of features spans a longer range, as larger convolution kernels can help the model capture these global features. In contrast, small convolution kernels (such as 3×1)

can only capture local short-term features, which limits the model's ability to effectively capture global information in signals with long-term dependencies, leading to a decline in performance. The line graph in the figure indicates that changes in the structure of the depthwise separable convolutional neural network significantly impact the F1-score. When SNR is ⩽-4 dB, the F1-score of Model 3 consistently remains below 85%. When SNR is >-4 dB, there is some improvement in the F1-score, but it still falls considerably short of the method proposed in this paper. The main reason is that after modifying the model structure, the parameter count and computational load of standard convolutions increase significantly, making the model prone to overfitting and unstable training. Depthwise separable convolutions can extract features more efficiently and avoid redundant information, whereas the global computation approach of standard convolutions may lack flexibility in feature extraction. Furthermore, the high computational cost of standard convolution can result in a slow training process, which in turn affects the model's convergence and performance. Under constraints of limited data and computational resources, standard convolution is less effective than depthwise separable convolution, leading to reduced generalization ability and ultimately impacting the F1-score.

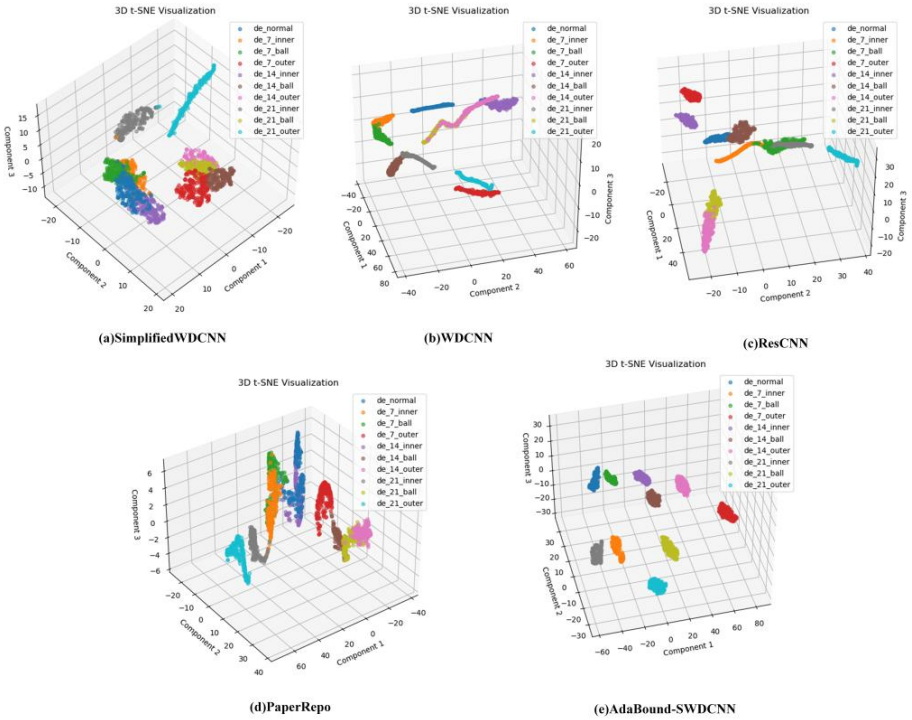


Figure 20. T-SNE 3D visualization of different methods for bearing fault diag-nosis under -8dB white Gaussian noise.

3.7. Visual analysis

3.7.1. Visualization of feature distributions

To further illustrate the feature learning of various methods, stochastic neighborhood embedding (t-SNE) [37] is employed under Gaussian white noise with an SNR of -8dB to visualize the features learned by classifiers in the gearbox diagnostic task. The aim of the t-SNE technique is to reduce the dimensionality of high-dimensional data and capture the similarity between data points in a lower-dimensional space. Figure 20 presents the three-dimensional t-SNE representations of different methods. As shown in Figure 20(a), some input data points from different health conditions exhibit significant overlap and appear quite chaotic, making them indistinguishable. In Figure 20(b), there

is a clear clustering effect among the same fault types, yet significant overlap of data points persists between different fault diagnosis tasks. From Figure 20(b) to Figure 20(d), the same fault types are still not the closest in different methods. This may be due to highly noisy features introducing additional variations and inconsistencies in the data, resulting in multiple clusters rather than a single coherent cluster. From Figure 20(b) to Figure 20(d), it is evident that some samples are misclassified as others. However, the results visualized in Figure 20(e) clearly demonstrate that the proposed AdaBoundL2-DSCNN exhibits strong representation learning capability for high-noise signals. Experiments in Table 13 show that the proposed AdaBoundL2-DSCNN achieves superior feature discrimination and separation, effectively suppressing inter-class overlap under strong noise. It attains the highest F1-score (0.9482) and inter/intra distance ratio (11.73), outperforming the baseline by 8.5% and 17.2%, respectively. These results indicate a favorable balance between feature compactness and separability, leading to significantly enhanced diagnostic robustness.

Table 13. Comparison of key performance metrics between AdaBoundL2-DSCNN and benchmark models.

Models	F1-Score	Intra-class Distance	Inter-class Distance	Ratio (Inter/Intra)
AdaBoundL2-DSCNN	**0.9482**	0.1137	1.3328	**11.7251**
SimplifiedWDCNN	0.5587	0.1855	0.9076	4.8914
WDCNN	0.4329	0.1510	1.2535	8.2993
ResCNN	0.4983	0.0939	0.9403	10.0140
PaperRepo	0.8737	0.1903	1.0834	5.6943

4. Conclusion

The AdaBoundL2-DSCNN model proposed in this paper maintains stable and accurate predictions under high noise, providing a reliable technical method for vibration monitoring and fault diagnosis. However, there are some limitations to be explored in future work:

(1) This study only adopts single-modal vibration data for equipment health identification, which suffers from low reliability and incomplete perception due to single-sensor constraints. To address these drawbacks, future research will explore multimodal heterogeneous data fusion by integrating infrared thermography, visual images, current, temperature, and stress signals. Constructing deep learning-driven fusion frameworks will achieve cross-modal feature complementation and effectively enhance the feature representation and anti-interference performance of health monitoring systems.

(2) Although the AdaBoundL2-DSCNN model performs well in computational and memory efficiency, it has obvious performance bottlenecks in complex task processing and large-scale dataset handling. Therefore, follow-up research will focus on model structure optimization and adaptive training strategy improvement so as to further enhance its computational efficiency, cross-domain generalization ability, and real-time monitoring capacity.

(3) In industrial practice, failure samples of rotating machinery are extremely scarce, and severe data imbalance imposes significant constraints on model training. Existing research predominantly relies on conventional data splitting strategies (70% training set) and fails to adequately address fault diagnosis under extremely low data ratios with only a small number of fault samples. Correspondingly, there is a lack of in-depth investigation and exploration of fault diagnosis methods suitable for such scenarios. Future research will focus on model optimization mechanisms under low-sample and data-imbalanced conditions, and further explore the synergistic integration of optimizer design and few-shot learning techniques.

In conclusion, targeting the limitations of single-modal vibration analysis, subsequent research will prioritize multimodal information fusion to improve model generalization, robustness, and real-time performance, providing reliable technical support for the intelligent operation and maintenance of industrial systems.

References

1. Matania O, Dattner I, Bortman J, Kenett RS, Parmet Y. A systematic literature review of deep learning for vibration-based fault diagnosis of critical rotating machinery: Limitations and challenges. *Journal of Sound and Vibration* 2024; 590: 118562. doi:10.1016/j.jsv.2024.118562.
2. Zhang H, Che W, Cao Y, Guan Z, Zhu C. Condition Monitoring and Fault Diagnosis of Rotating Machinery Towards Intelligent

Manufacturing: Review and Prospect. *Iranian Journal of Science and Technology, Transactions of Mechanical Engineering* 2024; 49(1): 1-34. doi:10.1007/s40997-024-00783-w.

3. 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. doi:10.1016/j.measurement.2023.113754.
4. Mohammadi A, Westny T, Jung D, Krysander M. Analysis of Numerical Integration in RNN-Based Residuals for Fault Diagnosis of Dynamic Systems. *IFAC-PapersOnLine* 2023; 56(2): 2909-2914. doi:10.1016/j.ifacol.2023.10.1411.
5. Yu JB. Evolutionary manifold regularized stacked denoising autoencoders for gearbox fault diagnosis. *Knowledge-Based Systems* 2019; 178: 111-122. doi:10.1016/j.knosys.2019.04.022.
6. Li C, Mo L, Kwoh CK, Li X, Chen Z, Wu M, et al. Noise-robust multi-view graph neural network for fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing* 2025; 224: 112025. doi:10.1016/j.ymssp.2024.112025.
7. Xie T, Huang X, Choi SK. Intelligent Mechanical Fault Diagnosis Using Multisensor Fusion and Convolution Neural Network. *IEEE Transactions on Industrial Informatics* 2022; 18(5): 3213-3223. doi:10.1109/tii.2021.3102017.
8. 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. doi:10.1016/j.isatra.2021.11.028.
9. Shao Y. Denoising Noisy Neural Networks: A Bayesian Approach with Compensation_supp1-3290327.pdf. *Institute of Electrical and Electronics Engineers (IEEE)*. doi:10.1109/tsp.2023.3290327/mm1.
10. Lou S, Deng J, Lyu S. Chaotic signal denoising based on simplified convolutional denoising auto-encoder. *Chaos, Solitons & Fractals* 2022; 161: 112333. doi:10.1016/j.chaos.2022.112333.
11. Zhong T, Cong Z, Wang H, Lu S, Dong X, Dong S, et al. Multiscale Encoder–Decoder Network for DAS Data Simultaneous Denoising and Reconstruction. *IEEE Transactions on Geoscience and Remote Sensing* 2023; 61: 1-15. doi:10.1109/tgrs.2023.3323527.
12. Munishamaiaha K, Rajagopal G, Venkatesan DK, Arif M, Vicoveanu D, Chiuchisan I, et al. Robust Spatial–Spectral Squeeze–Excitation AdaBound Dense Network (SE-AB-Densenet) for Hyperspectral Image Classification. *Sensors* 2022; 22(9): 3229. doi:10.3390/s22093229.
13. Liu X, Zhao T, Cao J, Li P. Design and prediction of self-organizing interval type-2 fuzzy wavelet neural network. *Information Sciences* 2024; 661: 120157. doi:10.1016/j.ins.2024.120157.
14. Li W, Chiu PWY, Li Z. An Accelerated Finite-Time Convergent Neural Network for Visual Servoing of a Flexible Surgical Endoscope With Physical and RCM Constraints. *IEEE Transactions on Neural Networks and Learning Systems* 2020; 31(12): 5272-5284. doi:10.1109/tnnls.2020.2965553.
15. Meni MJ, White RT, Mayo ML, Pilkiewicz KR. Entropy-based guidance of deep neural networks for accelerated convergence and improved performance. *Information Sciences* 2024; 681: 121239. doi:10.1016/j.ins.2024.121239.
16. Islam MN, Shrestha R, Chowdhury SR. Energy-Efficient and High-Throughput CNN Inference Engine Based on Memory-Sharing and Data-Reusing for Edge Applications. *IEEE Transactions on Circuits and Systems I: Regular Papers* 2024; 71(7): 3189-3202. doi:10.1109/tcsi.2024.3392807.
17. Papp L, Haberl D, Ecsedi B, Spielvogel CP, Krajnc D, Grahovac M, et al. DEBI-NN: Distance-encoding biomorphic-informational neural networks for minimizing the number of trainable parameters. *Neural Networks* 2023; 167: 517-532. doi:10.1016/j.neunet.2023.08.026.
18. Romero Ugalde HM, Carmona JC, Alvarado VM, Reyes-Reyes J. Neural network design and model reduction approach for black box nonlinear system identification with reduced number of parameters. *Neurocomputing* 2013; 101: 170-180. doi:10.1016/j.neucom.2012.08.013.
19. Chu SC, Dou ZC, Pan JS, Kong L, Snášel V, Watada J. DWSR: an architecture optimization framework for adaptive super-resolution neural networks based on meta-heuristics. *Artificial Intelligence Review* 2024; 57(2). doi:10.1007/s10462-023-10648-4.
20. HU M, HE Y, LIN X, LU Z, JIANG Z, MA B. Digital twin model of gas turbine and its application in warning of performance fault. *Chinese Journal of Aeronautics* 2023; 36(3): 449-470. doi:10.1016/j.cja.2022.07.021.
21. Lin H, Huang X, Chen Z, He G, Xi C, Li W. Matching Pursuit Network: An Interpretable Sparse Time-Frequency Representation Method Toward Mechanical Fault Diagnosis. *IEEE Transactions on Neural Networks and Learning Systems* 2025; 36(7): 12377-12388. doi:10.1109/tnnls.2024.3483954.
22. Qi J, Chen Z, Kong Y, Qin W, Qin Y. Attention-guided graph isomorphism learning: A multi-task framework for fault diagnosis and

remaining useful life prediction. *Reliability Engineering & System Safety* 2025; 263: 111209. doi:10.1016/j.ress.2025.111209.

23. Zhi S, Niu Y, Ma L, Wu H, Shen H, Wang T. Local Entropy Selection Scaling-extracting Chirplet Transform for Enhanced Time-Frequency Analysis and Precise State Estimation in Reliability-Focused Fault Diagnosis of Non-stationary Signals. *Eksploatacja i Niezawodność – Maintenance and Reliability* 2026; 28(1): 205977. doi:10.17531/ein/205977.
24. ZHANG J, KONG X, LI X, HU Z, CHENG L, YU M. Fault diagnosis of bearings based on deep separable convolutional neural network and spatial dropout. *Chinese Journal of Aeronautics* 2022; 35(10): 301-312. doi:10.1016/j.cja.2022.03.007.
25. Jia X, Qin N, Huang D, Zhang Y, Du J. A clustered blueprint separable convolutional neural network with high precision for high-speed train bogie fault diagnosis. *Neurocomputing* 2022; 500: 422-433. doi:10.1016/j.neucom.2022.05.056.
26. Guo Y, Zhou D, Chen H, Yue X, Cheng Y. Fault Intelligent Diagnosis for Distribution Box in Hot Rolling Based on Depthwise Separable Convolution and Bi-LSTM. *Processes* 2024; 12(9): 1999. doi:10.3390/pr12091999.
27. Liu W, Zhang Z, Ye Z, He Q. A novel intelligent fault diagnosis method for gearbox based on multi-dimensional attention denoising convolution. *Scientific Reports* 2024; 14(1). doi:10.1038/s41598-024-75522-x.
28. Li Z, Liu F, Yang W, Peng S, Zhou J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems* 2022; 33(12): 6999-7019. doi:10.1109/tnnls.2021.3084827.
29. Saxena A, Goebel K. CWRU bearing data set. Case Western Reserve University. https://engineering.case.edu/bearingdatacenter, 2008 [accessed 20 November 2026].
30. Gong W, Chen H, Zhang Z, Zhang M, Gao H. A Data-Driven-Based Fault Diagnosis Approach for Electrical Power DC-DC Inverter by Using Modified Convolutional Neural Network With Global Average Pooling and 2-D Feature Image. *IEEE Access* 2020; 8: 73677-73697. doi:10.1109/access.2020.2988323.
31. Huang X, Xie T, Wu J, Zhou Q, Hu J. Deep continuous convolutional networks for fault diagnosis. *Knowledge-Based Systems* 2024; 292: 111623. doi:10.1016/j.knosys.2024.111623.
32. Lee D, Jeong J. Few-Shot Learning-Based Light-Weight WDCNN Model for Bearing Fault Diagnosis in Siamese Network. *Sensors* 2023; 23(14): 6587. doi:10.3390/s23146587.
33. Zhao J, Lu P, Du C, Cao F. Active Fault-Tolerant Strategy for Flight Vehicles: Transfer Learning-Based Fault Diagnosis and Fixed-Time Fault-Tolerant Control. *IEEE Transactions on Aerospace and Electronic Systems* 2024; 60(1): 1047-1059. doi:10.1109/taes.2023.3333763.
34. Yan S, Shao H, Wang J, Zheng X, Liu B. LiConvFormer: A lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention. *Expert Systems with Applications* 2024; 237: 121338. doi:10.1016/j.eswa.2023.121338.
35. Edward, Ragab M, Wu M, Xu Y, Chen Z, Alseiari A, et al. EverAdapt: Continuous adaptation for dynamic machine fault diagnosis environments. *Mechanical Systems and Signal Processing* 2025; 226: 112317. doi:10.1016/j.ymssp.2025.112317.
36. Li J, Wang N, Wang D, Gao G, Pan W. Lightweight neural networks with anti-colored noise for bearing fault diagnosis using deep separable convolution and transfer learning. *Scientific Reports* 2025; 15(1). doi:10.1038/s41598-025-28332-8.
37. Suh S, Rey VF, Lukowicz P. TASKED: Transformer-based Adversarial learning for human activity recognition using wearable sensors via Self-KnowledgE Distillation. *Knowledge-Based Systems* 2023; 260: 110143. doi:10.1016/j.knosys.2022.110143.