



ARTICLE

Meta-Learning Method with Improved Relation-Distribution Module for Industrial Equipment Fault Diagnosis Using Small-Sample Acoustic Signals

Jialin Li*, Yihong Liu, Yang Yang and Shirong Li

School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing, China

*Corresponding Author: Jialin Li. Email: jialinli_neu@163.com

Received: 01 July 2026; Accepted: 24 August 2026; Published: 22 October 2026

ABSTRACT: Compared to traditional machine learning and signal processing techniques, deep learning has achieved remarkable results in industrial fault diagnosis. However, the difficulty of obtaining sufficient labeled fault acoustic data in real industrial environments has led to increased interest in few-shot learning methods. Existing methods are typically constrained by their reliance on convolutional architectures that mainly emphasize local patterns, limiting their ability to capture task-specific relationships across classes. To address this challenge, a novel meta-learning method with an improved relation-distribution module is proposed for industrial acoustic signals under small-sample data conditions. The framework integrates a hierarchical convolutional feature backbone to extract multi-scale spectral representations, improving robustness to variations in operating conditions and noise. A relation-aware metric alignment module further models task-specific semantics and cross-sample dependencies, while a distribution-aware Wasserstein metric module refines class separability by emphasizing local discriminative structures. Finally, a classification module combines contextual and distribution-aware cues to ensure reliable decision making with very few training examples. Extensive experiments and ablation studies on the MIMII and PT Gear Fault datasets, together with additional validation on the CUMTB dataset, demonstrate that the proposed framework consistently outperforms compared methods under different signal-to-noise ratios and working conditions, confirming its effectiveness, robustness, and generalization capability in practical industrial scenarios.

KEYWORDS: Meta-learning; few-shot learning; acoustic signal; fault diagnosis

1 Introduction

Recent advances in machine learning and deep learning have enabled intelligent fault diagnosis for industrial equipment [1–3], allowing automatic extraction of discriminative features from acoustic and vibration signals and shifting diagnosis from experience-driven to data-driven, improving accuracy and efficiency [4]. Audio-based methods capture subtle sound variations, reveal localized defects and global anomalies, and offer a non-invasive, robust, and easily deployable solution suitable for harsh environments and continuous monitoring. However, applying deep-learning-based acoustic diagnosis remains challenging because models require large labeled datasets, while industrial acoustic data are often scarce, imbalanced, and difficult to annotate [5,6].

Although data augmentation and meta-learning have been widely explored for few-shot fault diagnosis [7–10], and recent prototypical variants have also been developed [11,12], existing methods still face difficulties under noisy industrial acoustic signals with time-varying spectra and overlapping fault patterns.

To clarify the novelty of the proposed framework, Table 1 compares it with recent closely related few-shot fault diagnosis methods.

Table 1: Novelty-oriented comparison with closely related few-shot fault diagnosis methods.

Method	Main Mechanism	Difference from this Work
MSN [13]	Multi-path Siamese CNN with pairwise similarity learning	No explicit support-query relation alignment or distribution-aware Wasserstein metric.
PTFM [14]	Prototypical time-frequency mixer	No acoustic HCF backbone or relation-distribution joint metric.
TURL-ML [15]	Task-adaptive meta-regularization for cross-domain diagnosis	Focuses on regularization rather than acoustic relation and class-distribution modeling.
TETARPN [16]	Transformer-enhanced task-adaptive prototypical network	Lacks the HCF-RAMA-DAWM feature-relation-distribution chain.
Proposed	HCF + RAMA + DAWM with weighted fusion	Integrates multi-scale acoustic features, support-query alignment, and Wasserstein distribution modeling.

As shown in Table 1, the proposed framework differs from existing methods by jointly considering acoustic feature extraction, relation alignment, and distribution modeling.

These limitations motivate an acoustic few-shot fault diagnosis method that captures multi-scale spectral features and task-specific relational structures under scarce and variable sound samples.

Building upon these advances in metric-based meta-learning, this study further extends the paradigm to small-sample fault diagnosis with acoustic signals, addressing the challenges of limited labeled data, noisy environments, and complex working conditions. The main contributions of this paper are as follows:

1. A spectral feature backbone with hierarchical receptive fields is employed to extract discriminative representations from Mel spectrograms, capturing both fine-grained cues and broader contextual patterns.
2. A relation-aware alignment mechanism is designed to enhance the interaction between query and support sets, thereby improving semantic consistency in few-shot classification.
3. A distribution-aware Wasserstein metric is introduced to model statistical differences between class distributions, which strengthens class separability under limited data conditions.

2 Related Work

2.1 Industrial Acoustic and Few-Shot Fault Diagnosis

Industrial fault diagnosis has gradually shifted from manual inspection and handcrafted signal analysis to data-driven intelligent recognition. Acoustic signals are attractive because they can be collected non-invasively and reflect abnormal impacts, friction, leakage, and rotating defects. Recent deep models further improve automatic feature learning for acoustic and vibration diagnosis [17]. However, their performance often depends on sufficient labeled samples and stable operating conditions.

Few-shot fault diagnosis aims to recognize fault categories with only a small number of labeled examples. Existing studies address this issue through data augmentation, transfer learning, and meta-learning. For example, augmentation-based methods expand minority classes, while meta-learning methods learn

transferable diagnostic knowledge from episodic tasks. Recent approaches such as MSN, BRPN, and elastic prototypical learning improve feature transfer under limited samples, but most of them are still mainly developed for vibration signals or general time-frequency representations rather than noisy industrial acoustic signals.

2.2 Metric and Relation-Based Few-Shot Learning

Metric-based few-shot learning classifies query samples by comparing them with support samples or class prototypes in an embedding space. A typical prototype-based formulation is

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\theta(x_i), \quad \hat{y} = \arg \min_k d(f_\theta(x_q), c_k), \quad (1)$$

where c_k denotes the prototype of class k , S_k is its support set, and $d(\cdot)$ is a distance metric. Matching Networks use attention over labeled examples [18], while Prototypical Networks classify samples according to distances from class prototypes [19]. These methods are simple and effective, but their comparisons are generally performed in a learned embedding space using attention-based or prototype-based similarity measures. Such formulations may still be limited in capturing fine-grained local discriminative patterns and noise-robust relations for acoustic faults with overlapping spectral patterns and strong interference.

Relation-based methods further learn a task-specific comparison function instead of using a predefined metric. A generic relation score can be written as

$$r_k = g_\phi(f_\theta(x_q), c_k), \quad \hat{y} = \arg \max_k r_k, \quad (2)$$

where $g_\phi(\cdot)$ estimates the similarity between the query embedding and the class prototype. Relation Networks estimate support-query similarity through a learnable relation module [20], and later variants introduce attention or prototype refinement to improve cross-sample interaction. Nevertheless, many relation-based methods still focus on pairwise similarity and do not explicitly combine relation alignment with class distribution modeling, making it difficult to fully characterize intra-class compactness and inter-class separability in small-sample acoustic diagnosis.

2.3 Limitations of Existing Methods

In summary, existing studies have advanced acoustic diagnosis, few-shot learning, metric learning, and relation modeling, but several limitations remain. These limitations become more evident in noisy industrial acoustic signals, where time-varying spectra and overlapping fault harmonics make local feature extraction and fixed support-query metrics less stable. Traditional signal decomposition methods depend on handcrafted components; standard CNN-based methods require sufficient labeled data; metric-based few-shot methods often use fixed distances; and relation-based methods rarely model feature distributions. To address these issues, this work proposes an HCF-RAMA-DAWM framework that links multi-scale acoustic feature extraction, support-query relation alignment, and Wasserstein distribution modeling into a unified feature-relation-distribution diagnostic chain.

3 The Proposed Method

3.1 Overall Process of the Proposed Method

As shown in Fig. 1, raw acoustic signals are converted into Mel-frequency spectrograms and divided into support and query sets. HCF extracts multi-scale features, RAMA and DAWM obtain $\hat{Y}_{b,i}^R$ and $D_{b,i}^2$, and

prediction uses $p_{b,i} = \text{softmax}_i(\lambda \widehat{Y}_{b,i}^R - (1 - \lambda) D_{b,i}^2)$, $\hat{y}_b = \arg \max_i p_{b,i}$. Here, B , C , H , and W denote batch size, channel number, feature height, and feature width, respectively; n is the number of fault classes, and superscripts s and q indicate support and query samples.

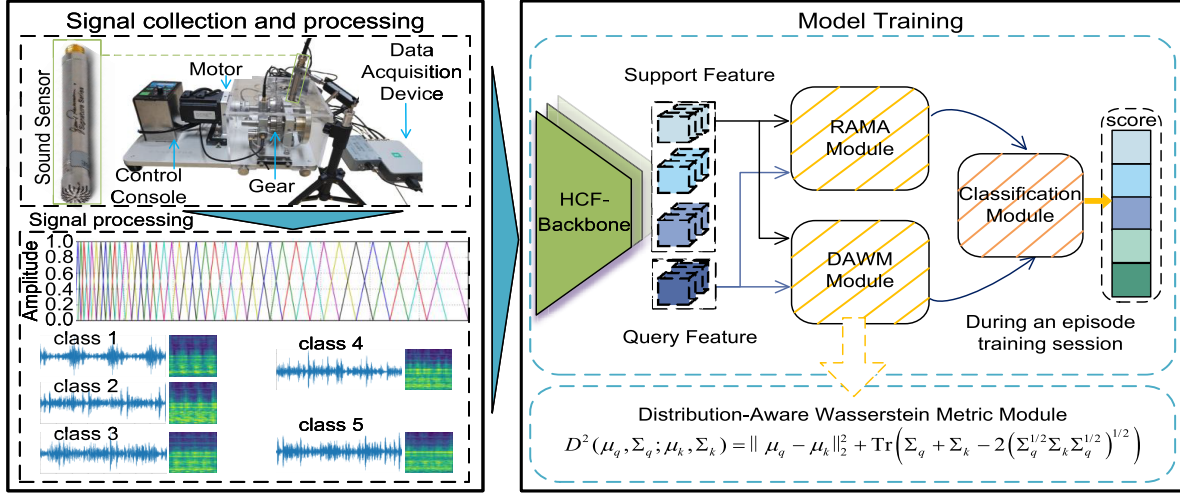


Figure 1: Architecture of the proposed acoustic-based few-shot fault diagnosis framework for industrial scenarios.

3.2 Hierarchical Convolutional Feature Backbone (HCF Backbone)

The Hierarchical Convolutional Feature (HCF) backbone extracts multi-scale Mel-spectrogram representations using a downsampling stem and repeated multi-branch depthwise convolutional blocks, as shown in Fig. 2.

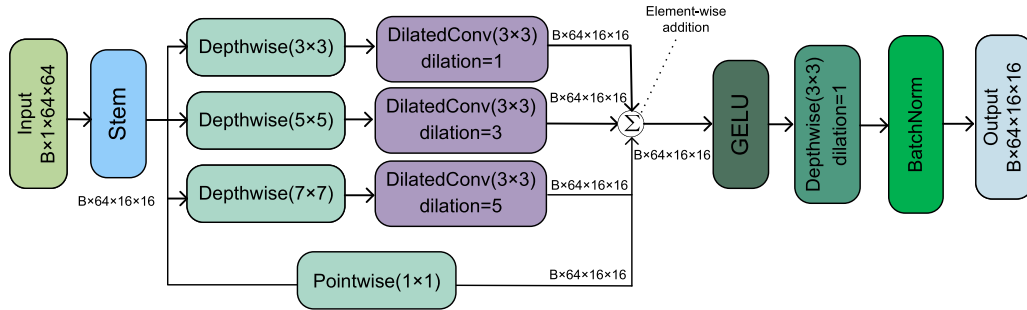


Figure 2: Hierarchical convolutional feature backbone (HCF-backbone) for multi-level spectral feature extraction. Tensor dimensions and element-wise addition are marked in the revised HCF block.

The input spectrogram $X \in \mathbb{R}^{B \times 1 \times H \times W}$ first passes through two 3×3 convolutional stem layers with stride 2 and padding 1, batch normalization, and ReLU activation, expanding channels from $1 \rightarrow 32 \rightarrow 64$ and reducing the resolution to $H/4 \times W/4$. The stem feature X_{stem} is then processed by a 3×3 depthwise convolution with layer normalization and GELU activation.

The output is then fed into a multi-branch module consisting of three parallel branches with different receptive fields, formulated as:

$$X_{\text{branch},i} = \text{DConv}_{D_i}(\text{DWConv}_{K_i}(X_{\text{stem}})) \quad (3)$$

where $K_i = 3, 5, 7$ and $D_i = 1, 3, 5$. Each branch output is normalized before fusion, enabling the branches to capture multi-scale spectral features with different receptive fields.

The three branch outputs are fused with the pointwise-convolved stem output X_{stem} , then passed through GELU activation, depthwise convolution, and batch normalization to produce the final feature map.

Two HCF blocks are stacked in this study.

3.3 Relation-Aware Metric Alignment Module (RAMA)

The Relation-Aware Metric Alignment Module (RAMA) refines support and query features with self-attention and then aligns query features with class prototypes through cross-attention, as shown in Fig. 3.

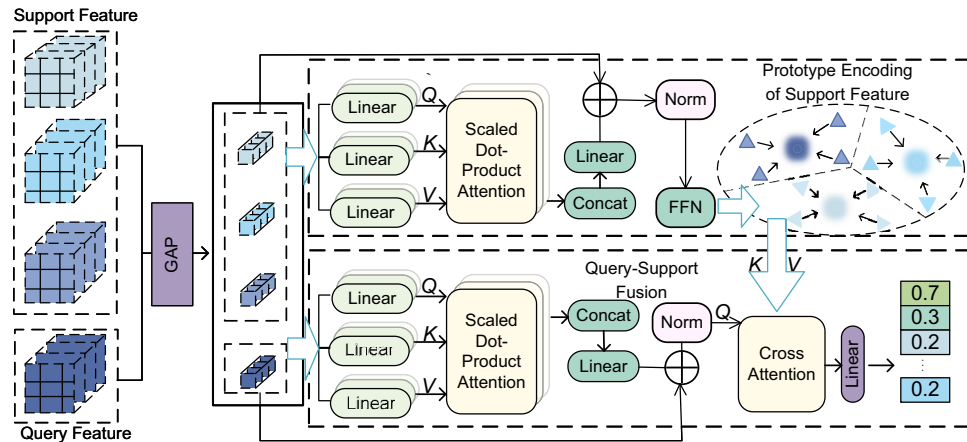


Figure 3: Relation-aware metric alignment module.

After HCF, each support or query sample produces $X_F \in \mathbb{R}^{B \times C \times H \times W}$. A global descriptor is obtained by average pooling:

$$X_{\text{avg},c} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_{HCF,i,j,c}, \quad c = 1, 2, \dots, C \quad (4)$$

The pooled feature is then refined by 4-head self-attention and a feedforward network, as summarized in Table 2. For class i , the refined support representations $\tilde{X}_{i,b}^s$ are averaged to form the prototype:

$$P_i = \frac{1}{B} \sum_{b=1}^B \tilde{X}_{i,b}^s \quad (5)$$

Refined query features are then aligned with prototypes by cross-attention and mapped to relation scores:

$$\hat{Y}_{b,i}^R = \text{CrossAttn}(X_b^q, p_i) W + b, \quad i = 1, \dots, n \quad (6)$$

where $\text{CrossAttn}(\cdot)$ denotes attention-based query-prototype alignment and $\hat{Y}_b^R$ is the relation score vector of query X_b^q over all n classes.

The key layer settings and tensor dimensions of HCF and RAMA are summarized in Table 2.

Table 2: Architecture summary of HCF and RAMA.

Module	Main Settings	Input → Output
HCF stem	Two 3×3 convolutional stages; channels $1 \rightarrow 32 \rightarrow 64$; stride 2; padding 1	$B \times 1 \times 64 \times 64 \rightarrow B \times 64 \times 16 \times 16$
HCF block	Three depthwise-dilated branches with $K_i = 3, 5, 7$ and $D_i = 1, 3, 5$; element-wise addition; repeated twice	$B \times 64 \times 16 \times 16 \rightarrow B \times 64 \times 16 \times 16$
RAMA projection	GAP followed by 4-head self-attention; $Q/K/V$ and output projections are $64 \rightarrow 64$; head dimension 16	$B \times 64 \times 16 \times 16 \rightarrow B \times 64$
RAMA FFN and score	FFN $64 \rightarrow 64 \rightarrow 64$; cross-attention projection $64 \rightarrow 1$ per class	Prototype 1×64 ; query score $B \times 1$ per class

3.4 Distribution-Aware Wasserstein Metric Module (DAWM)

DAWM complements RAMA by measuring query-to-prototype relations with first- and second-order feature statistics. Compared with Euclidean or Mahalanobis distances, it captures both feature-location shifts and covariance differences.

For each class $k \in \{1, \dots, n\}$, all spatial features from its K support samples are flattened to obtain vectors $\{x_i\}_{i=1}^{N_k}$, where N_k is the total number of spatial features aggregated for that class. Based on these, the statistical prototype (μ_k, Σ_k) is constructed, with the mean and covariance defined as:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i, \quad \Sigma_k = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (x_i - \mu_k)(x_i - \mu_k)^\top \quad (7)$$

The query feature q is processed in the same way, yielding (μ_q, Σ_q) . The Wasserstein-2 distance between the query distribution and the class- k prototype is defined as:

$$D^2(\mu_q, \Sigma_q; \mu_k, \Sigma_k) = \|\mu_q - \mu_k\|_2^2 + \text{Tr} \left(\Sigma_q + \Sigma_k - 2 \left(\Sigma_q^{1/2} \Sigma_k \Sigma_q^{1/2} \right)^{1/2} \right) \quad (8)$$

where $\|\cdot\|_2$ is the Euclidean norm, $\text{Tr}(\cdot)$ denotes the matrix trace, and $\Sigma^{1/2}$ represents the matrix square root of covariance Σ .

For numerical stability under few-shot conditions, the covariance matrices are regularized by adding a small diagonal term, and the covariance term is evaluated using symmetric matrix square-root operations when possible; otherwise, a trace-based surrogate is used. For each query, distances to all n class prototypes form a distribution-aware distance profile, and the predicted class is obtained by the minimum Wasserstein distance.

4 Experimental Validation

4.1 Experimental Settings

Using Adam settings in Table 3, fixed training and evaluation/test pools were formed before episodic sampling; each episode draws $K = 5$ support and $Q = 1$ query sample/class without replacement (8-way for MIMII/CUMTB, 5-way for PT Gear Fault) and optimizes relation loss plus query cross-entropy for 50 episodes/epoch and 50 epochs ($T = 2500$); evaluation/testing uses 50 episodes.

Table 3: Representative training details for reproducibility.

Dataset	Representative Setting	LR/Epochs	Training Time	Convergence Behavior
MIMII	6 dB, 30 samples/class, 5-shot	0.001/50	85.58 min (102.70 s/epoch)	Loss decreased rapidly and validation accuracy stabilized in later epochs; best epoch 35.
PT Gear Fault	50 samples/class, 5-shot	0.0005/50	40.65 min (48.78 s/epoch)	Validation accuracy rose quickly and remained stable after convergence; best epoch 18.
CUMTB	Cond_1, 1 rpm, 30 samples/class, 8-way 5-shot	0.001/50	39.58 min (47.50 s/epoch)	Loss decreased from 7.0354 to 0.0014, and validation accuracy reached its best value at epoch 29 and remained generally stable later.

RAMA uses 4-head self-attention, and the RAMA-DAWM balancing parameter λ is 0.6 for MIMII and 0.7 for PT Gear Fault. Experiments ran on an Intel i5-13400 CPU, 16 GB RAM, and NVIDIA GeForce RTX 4060 GPU. Unless specified, samples/class denotes training examples per class.

All datasets used 64×64 log-Mel spectrograms; pools came from MIMII file lists, PT within-recording earlier/later non-overlapping segments, and CUMTB class-wise samples. Checkpoints were selected by validation accuracy; source code and split files will be available upon reasonable request.

4.2 Case Study 1: Validation Based on the MIMII Dataset

4.2.1 Description of the MIMII Dataset

The MIMII dataset [21] contains eight categories from four industrial machine types: valves, pumps, fans, and slider rails, recorded under normal and anomalous conditions. It includes 14,719 normal and 3300 anomalous segments collected as 16-bit, 16 kHz recordings in reverberant factory environments, covering faults such as contamination, leakage, rotational imbalance, and rail damage, with background noise mixed at 6, 0, and -6 dB SNR. Each audio was converted into log-Mel spectrograms using a 1024-sample STFT window, a 512-sample shift, 64 Mel filter banks, and five consecutive frames per feature vector.

4.2.2 Results of Various Methods under Small Sample Conditions

Fig. 4 shows representative confusion matrices and t-SNE visualizations of ResNet18, MatchNet, CosineNet, and the proposed method under 6 dB SNR with 30 samples per class. In this representative visualization, MatchNet achieves 90.38% accuracy but suffers cross-category confusion, mainly between normal and faulty states of the same machine (fan and valve), while CosineNet (90.00%) partially alleviates this. ResNet18 and the proposed method reach 92.5% and 94.5%, respectively, with the proposed model showing the clearest class separation, effectively enhancing intra-class compactness and inter-class discriminability.

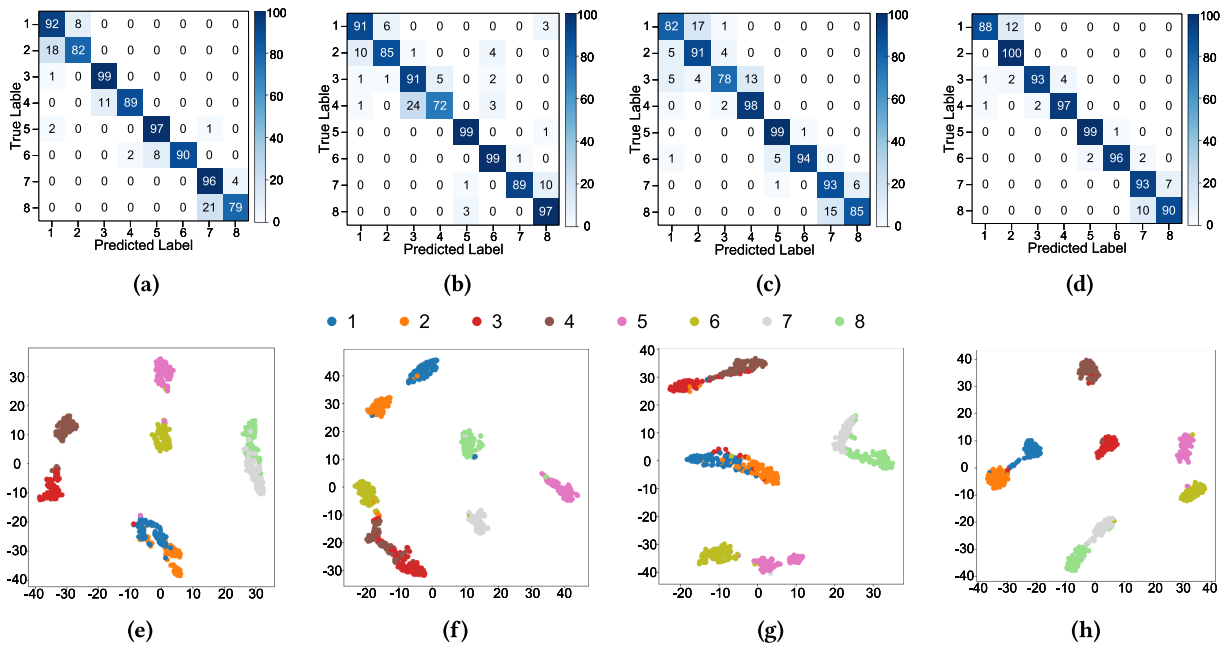


Figure 4: Confusion matrices and t-SNE visualizations of four representative models on the MIMII dataset (6 dB SNR, 30 samples per class). (a) ResNet18 (Confusion Matrix); (b) MatchNet (Confusion Matrix); (c) CosineNet (Confusion Matrix); (d) Proposed (Confusion Matrix); (e) ResNet18 (t-SNE); (f) MatchNet (t-SNE); (g) CosineNet (t-SNE); (h) Proposed (t-SNE).

4.2.3 The Impact of Sample Size on Model Recognition Performance

Table 4 reports the mean classification accuracy with standard deviation of ResNet18, ProtoNet, RelaNet, MatchNet, CosineNet, and the proposed method on MIMII under 6, 0, and -6 dB SNR with different samples per class. The proposed method achieves consistently strong performance under noisy and limited-sample conditions.

Fig. 5 further shows the performance trends across SNR levels and samples per class. The proposed method outperforms all baselines in every setting; for example, at 6 dB with only 10 training samples per class, it exceeds the best baseline by nearly 4%. The degradation at -6 dB is mainly caused by noise masking weak fault cues and increasing spectral overlap.

4.2.4 Comparison with State-of-the-Art Methods

To provide direct comparisons with recent foundation-model-based and Transformer-based few-shot baselines, the released FSAMLM, UniFault, BearLLM, and ET-Mahalanobis codebases were adapted to the same MIMII acoustic setting. FSAMLM was evaluated using its Tip-Adapter-F branch; since a released LoRA checkpoint was not available, the raw CLIP-based Tip-Adapter-F result was used. UniFault was evaluated by fine-tuning its Tiny pretrained model. BearLLM was evaluated by adapting its DCT reference-signal encoder and classification network to the eight-class acoustic setting. ET-Mahalanobis was evaluated by adapting its ensemble Transformer feature extractor and Mahalanobis metric-learning classifier to the eight-class acoustic setting. All reproduced baseline results are reported as the mean accuracy with standard deviation over five random seeds. The comparison with the proposed method is summarized in Table 5.

Table 4: Performance comparison of different models under varying SNR levels and samples per class.

Models	Samples/Class (6 dB)				Samples/Class (0 dB)				Samples/Class (-6 dB)				
	10	20	30	10	30	50	70	100	10	30	50	70	100
ResNet18	74.85 ± 1.07	81.93 ± 0.91	89.61 ± 0.62	57.98 ± 1.54	80.02 ± 0.99	88.10 ± 0.67	92.13 ± 0.60	95.91 ± 0.44	56.06 ± 1.72	67.98 ± 1.25	79.95 ± 0.92	81.64 ± 0.93	88.02 ± 0.68
ProtoNet	73.34 ± 1.09	85.01 ± 0.87	91.54 ± 0.55	58.91 ± 1.56	80.62 ± 0.93	90.13 ± 0.66	94.51 ± 0.52	95.42 ± 0.49	56.84 ± 1.61	71.94 ± 1.21	78.34 ± 0.94	83.41 ± 0.91	88.25 ± 0.64
RelaNet	74.31 ± 1.11	83.64 ± 0.86	92.81 ± 0.56	57.23 ± 1.57	80.15 ± 0.97	89.88 ± 0.62	93.97 ± 0.55	95.24 ± 0.46	55.86 ± 1.67	70.41 ± 1.20	78.46 ± 0.99	82.21 ± 0.89	86.47 ± 0.71
MatchNet	74.32 ± 1.13	80.87 ± 0.90	90.41 ± 0.63	58.25 ± 1.53	80.41 ± 0.98	88.96 ± 0.64	93.31 ± 0.58	94.84 ± 0.47	55.03 ± 1.66	69.64 ± 1.30	79.32 ± 0.92	81.45 ± 0.97	85.64 ± 0.69
CosineNet	72.81 ± 1.14	81.53 ± 0.94	88.74 ± 0.66	57.73 ± 1.61	78.93 ± 1.00	87.57 ± 0.74	91.41 ± 0.61	94.23 ± 0.45	54.13 ± 1.71	67.62 ± 1.39	78.03 ± 1.05	81.42 ± 0.95	84.98 ± 0.78
Proposed	78.56 ± 1.08	87.91 ± 0.72	94.53 ± 0.48	59.21 ± 1.42	83.71 ± 0.91	92.40 ± 0.57	96.21 ± 0.38	97.40 ± 0.31	58.21 ± 1.55	74.41 ± 1.21	83.34 ± 0.95	86.14 ± 0.79	90.38 ± 0.64

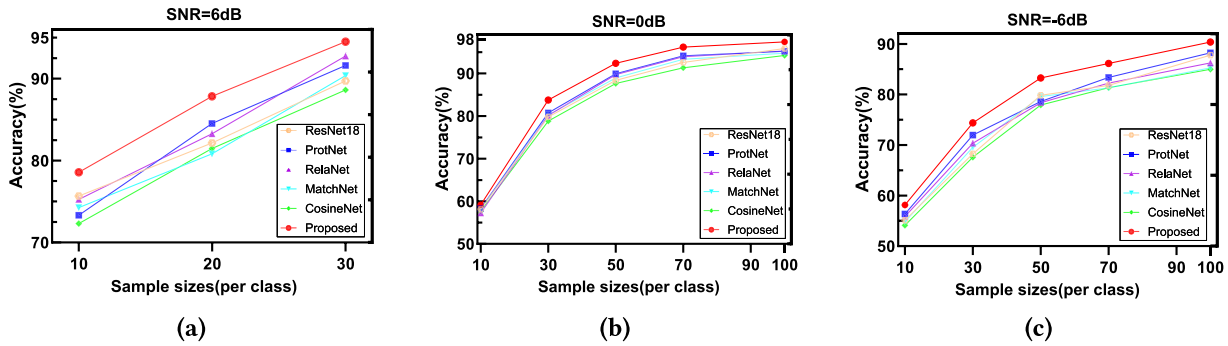


Figure 5: Accuracy comparison of different models under varying numbers of samples per class and SNR conditions on the MIMII dataset. (a) SNR = 6 dB; (b) SNR = 0 dB; (c) SNR = -6 dB.

Table 5: Comparison with reproduced recent baselines on the MIMII dataset.

Models	Samples/Class (6 dB)			Samples/Class (0 dB)			Samples/Class (-6 dB)		
	10	20	30	10	30	50	10	30	50
ET-Mahalanobis [22]	77.28 ± 1.05	85.32 ± 0.83	92.16 ± 0.59	57.44 ± 1.58	82.02 ± 0.94	90.98 ± 0.64	56.98 ± 1.63	71.77 ± 1.18	82.21 ± 0.91
UniFault [23]	77.33 ± 0.98	85.83 ± 0.72	94.17 ± 0.52	57.58 ± 1.02	81.98 ± 0.88	90.51 ± 0.94	56.42 ± 1.29	73.57 ± 1.09	82.37 ± 0.73
FSAMLM [24]	78.25 ± 1.01	86.61 ± 0.66	91.65 ± 0.59	58.43 ± 1.24	82.31 ± 0.74	92.09 ± 0.89	57.31 ± 1.49	71.68 ± 1.18	83.02 ± 1.05
BearLLM [25]	78.42 ± 0.82	88.08 ± 0.68	94.36 ± 0.41	59.17 ± 1.25	82.50 ± 0.75	91.99 ± 0.54	57.03 ± 1.24	72.59 ± 1.13	82.95 ± 1.02
Proposed	78.56 ± 1.08	87.91 ± 0.72	94.53 ± 0.48	59.21 ± 1.42	83.71 ± 0.91	92.40 ± 0.57	58.21 ± 1.55	74.41 ± 1.21	83.34 ± 0.95

The reproduced FSAMLM, UniFault, BearLLM, and ET-Mahalanobis baselines are competitive in several settings, but the proposed method achieves the highest mean accuracy in eight of the nine reproduced comparisons, with higher mean accuracies at 6 dB with 30 samples per class, 0 dB with 50 samples per class, and -6 dB with 30 and 50 samples per class. These results support the effectiveness of the proposed acoustic relation-distribution modeling under noisy and limited-sample conditions.

To further evaluate robustness beyond fixed SNR settings, three MIMII scenarios were tested using the 6 dB, 30-samples-per-class, 5-shot setting: same-domain sampling, real machine background noise by mixing id_04 queries with normal-machine recordings from id_00, id_02, and id_06 at random 0–6 dB SNR, and machine-ID domain shift using id_04 as support and id_00/id_02/id_06 as query domains. As shown in Table 6, the proposed method remains robust to real machine background noise, while the larger degradation under machine-ID shift indicates that cross-device distribution mismatch is more challenging than additive noise.

Table 6: Additional robustness evaluation under real noise and domain shift.

Scenario	Setting	Accuracy (%)
Same-domain reference	id_04 support/query under the original sampling protocol	94.25 ± 0.48
Real machine background noise	id_04 queries mixed with normal recordings from id_00/id_02/id_06 at random 0–6 dB SNR	87.58 ± 0.73
Machine-ID domain shift	id_04 support domain and id_00/id_02/id_06 query domains	76.27 ± 0.94

4.3 Case Study 2: Validation Based on the PT Gear Fault Dataset

4.3.1 Test Bench Development Based on Acoustic Signals

A gear-fault acoustic dataset was collected on the PT400-MINI test rig (Fig. 6a), which simulates spur gear faults at 0–2500 rpm. Acoustic signals were recorded using a GRAS 46AE microphone and NI9234 DAQ at 51.2 kHz. The dataset contains five gear conditions, including normal condition (NO), missing tooth (MT), broken tooth (BT), worn tooth (WT), and cracked tooth (CT) (Fig. 6b), with five recordings per gear condition under varying conditions. Signals were segmented into 2048-point windows, producing 250 segments per recording; earlier/later non-overlapping segments formed training/evaluation-test pools, and same-condition recordings share a label. All segments were converted into log-Mel spectrograms using the same preprocessing pipeline as MIMII.

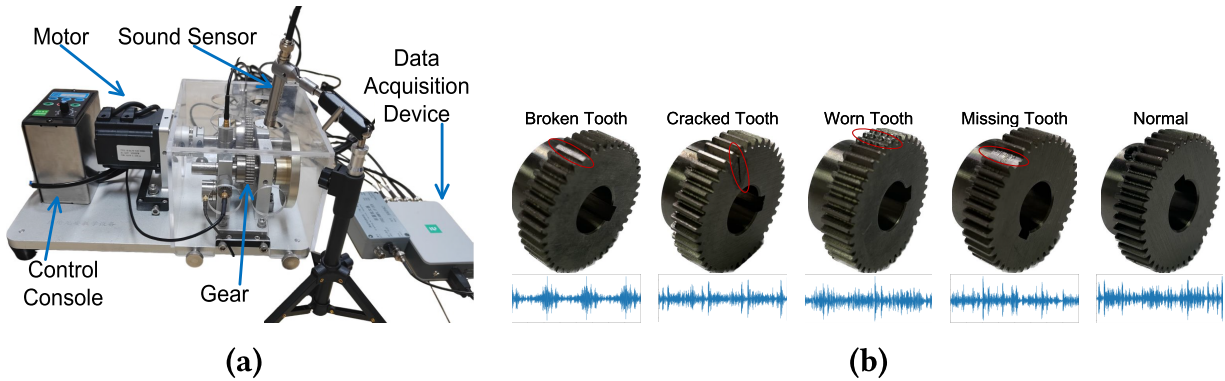


Figure 6: PT400-MINI test rig and gear fault categories. (a) PT400 gear fault test bench; (b) Five gear types and their corresponding sound signals.

4.3.2 Comparison of Diagnostic Results Using Multiple Methods

Fig. 7 compares six methods with 30 samples per class. In the confusion matrices, ResNet18 (95.0%) and CosineNet (96.0%) often confuse normal conditions with fault classes, MatchNet (96.6%) mainly confuses cracked-tooth samples with other faults, and ProtoNet (95.8%) and RelaNet (96.4%) still show errors involving the normal and cracked-tooth classes. The proposed method achieves the highest accuracy (99.2%), with clearer class boundaries and minimal confusion.

The t-SNE panels show that support and query embeddings form well-separated fault-class clusters, indicating that the proposed method learns a structured and discriminative feature space with both fine-grained fault cues and contextual patterns under variable operating conditions.

4.3.3 Fault Diagnosis Results under Various Numbers of Samples per Class

The experimental results on the PT Gear Fault dataset under varying operating conditions are presented in Table 7. A variety of few-shot learning methods, including ResNet18, ProtoNet, RelaNet, MatchNet, CosineNet, and the proposed approach, were evaluated using limited training samples per class (10, 20, 30, 40, and 50 samples per class). This setup allows assessment of each model's performance under small-sample conditions and its adaptability to varying operating environments.

From the results, it is observed that conventional CNN-based models such as ResNet18 can achieve relatively high accuracy when sufficient support samples are available, but their performance tends to be lower under extremely small-sample settings. ProtoNet, as a metric-based method, shows moderate improvement by leveraging prototype representations, while RelaNet and MatchNet further enhance classification

by exploiting inter-sample relational information. CosineNet also benefits from similarity-based learning, showing a slightly higher accuracy than previous metric-learning baselines in most cases.

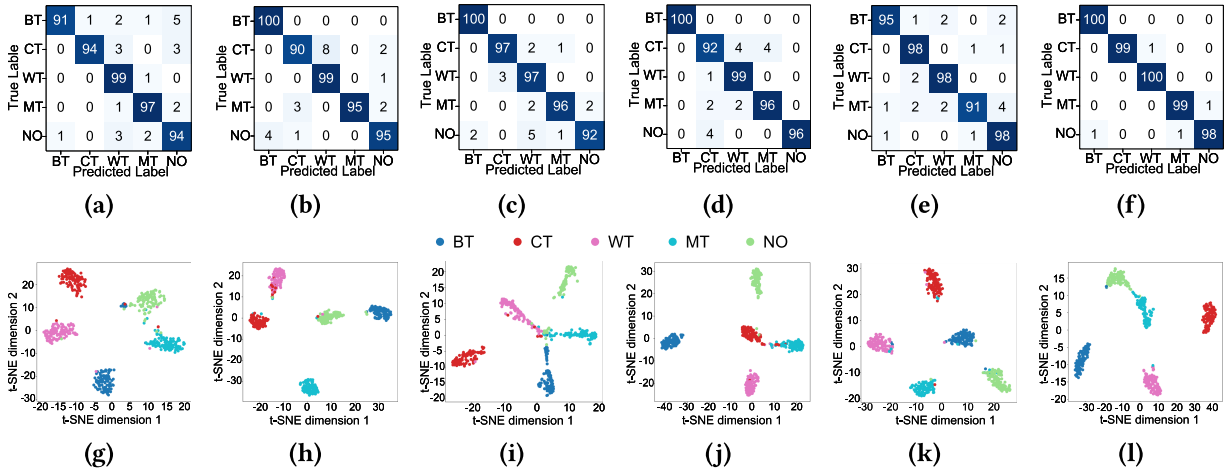


Figure 7: Comparison of the six methods on the PT Gear Fault dataset with 30 training samples per class. (a) ResNet18; (b) ProtoNet; (c) RelaNNet; (d) MatchNet; (e) CosineNet; (f) Proposed method; (g) ResNet18; (h) ProtoNet; (i) RelaNNet; (j) MatchNet; (k) CosineNet; (l) Proposed method.

Table 7: Evaluation of various methods on the PT Gear Fault dataset across different numbers of training samples per class.

Models	Samples per Class				
	10	20	30	40	50
ResNet18	87.55 ± 0.76	92.46 ± 0.58	94.35 ± 0.49	96.94 ± 0.35	98.12 ± 0.27
ProtoNet	88.67 ± 0.72	93.14 ± 0.55	95.87 ± 0.42	97.41 ± 0.34	98.76 ± 0.25
RelaNNet	89.75 ± 0.68	93.31 ± 0.54	96.51 ± 0.39	97.84 ± 0.31	98.67 ± 0.26
MatchNet	88.48 ± 0.73	94.07 ± 0.50	96.69 ± 0.37	97.53 ± 0.32	98.46 ± 0.28
CosineNet	89.46 ± 0.76	93.59 ± 0.47	97.21 ± 0.39	97.76 ± 0.30	99.07 ± 0.31
Proposed	90.14 ± 0.56	96.49 ± 0.42	98.71 ± 0.33	99.31 ± 0.27	99.47 ± 0.19

The proposed method consistently achieves the highest accuracy across all training scenarios, ranging from $90.14 \pm 0.56\%$ with 10 samples per class to $99.47 \pm 0.19\%$ with 50 samples per class. Notably, even in the smallest-sample scenario, it outperforms all compared models, indicating strong robustness and effectiveness in extracting discriminative acoustic features. This superior performance can be attributed to the model's ability to capture both fine-grained cues and broader contextual patterns between support and query samples, which enhances its generalization across varying operational conditions and fault types. These results demonstrate the proposed approach's practical applicability for industrial acoustic fault diagnosis tasks with small-sample and cross-condition challenges.

4.3.4 Computational Complexity Analysis

To evaluate deployment feasibility, all compared methods were profiled with 64×64 Mel-spectrograms on an Intel i5-13400 CPU, 16 GB RAM, and an NVIDIA GeForce RTX 4060 GPU using Python 3.10, PyTorch 2.5.1, and CUDA 12.1. Few-shot methods used the corresponding 5-shot single-query episode setting, while

ResNet18 and CosineNet used one-query inference under the same input size. Latency was averaged over 100 synchronized runs after 20 warm-up passes, and memory denotes peak GPU allocation. As shown in Table 8, the proposed method keeps a small parameter count and moderate memory consumption, with an acceptable accuracy-complexity tradeoff for small-sample industrial acoustic fault diagnosis.

Table 8: Computational complexity comparison with FLOPs, parameter count, inference latency, and memory consumption.

Method	Parameter Count (M)	FLOPs (G)	Inference Latency (ms)	Memory Consumption (MB)
ResNet18	11.174	0.283	2.87 ± 0.73	61.12
ProtoNet	0.120	3.953	17.53 ± 2.92	14.83
RelaNet	5.170	8.818	20.73 ± 2.77	42.35
MatchNet	0.120	3.953	18.02 ± 2.75	16.27
CosineNet	11.174	0.283	2.61 ± 0.75	61.12
Proposed	0.387	5.120	33.82 ± 4.30	18.37

4.4 Case Study 3: CUMTB Dataset

To examine generalization beyond the two main datasets, the proposed method was evaluated on the public CUMTB wind-turbine pitch-bearing acoustic dataset [26], which contains pitch-bearing measurements under different operating conditions. Cond_1 was used to form an additional 8-class validation set from the original 1 rpm acoustic signals, including one healthy class and seven single-fault classes. All signals were converted to log-Mel spectrograms using the same preprocessing pipeline as MIMII and PT Gear Fault, and 8-way 5-shot tests were conducted with 10, 20, 30, and 50 samples per class over five repeated runs.

As shown in Table 9, the mean accuracy increases from $81.19 \pm 0.61\%$ with 10 samples per class to $97.42 \pm 0.43\%$ with 50 samples per class. Although the very-small-sample result is affected by the different equipment type and pitch-bearing fault mechanism, the low standard deviations and consistent improvement indicate stable adaptation on CUMTB, further supporting the generalization ability of the proposed framework.

Table 9: Five-run validation results on CUMTB.

Samples/Class	Run 1	Run 2	Run 3	Run 4	Run 5	Mean $\pm$ SD
10	80.34	81.04	81.19	81.34	82.04	81.19 ± 0.61
20	85.43	86.04	86.42	86.80	87.41	86.42 ± 0.75
30	91.27	92.05	92.37	92.69	93.47	92.37 ± 0.81
50	96.82	97.32	97.42	97.52	98.02	97.42 ± 0.43

4.5 Ablation Experiments

4.5.1 The Impact of Multiple Modules Constituting the Proposed Model

Ablation experiments on MIMII and PT Gear Fault evaluate the contribution of HCF, RAMA, and DAWM (Table 10). Trial 1 uses only ResNet18, Trial 2 adds HCF for multi-scale spectral feature extraction, Trial 3 further introduces RAMA for support-query relation alignment, and Trial 4 combines HCF with DAWM for distribution-aware discrimination.

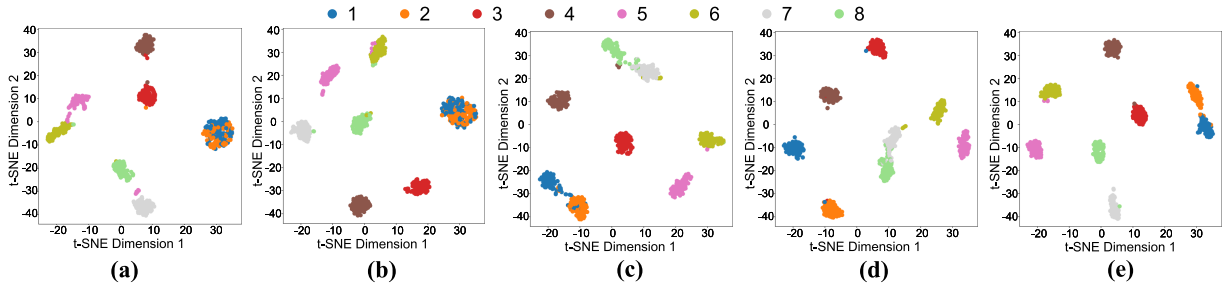
Table 10: Ablation study of different modules on the MIMII and PT Gear Fault datasets with training samples reported per class.

Trials	HCF	RAMA	DAWM	MIMII (Samples/Class)			PT Gear (Samples/Class)		
				10	50	100	10	40	100
Trial 1				75.72 ± 1.12	92.33 ± 0.63	96.79 ± 0.39	87.58 ± 0.78	97.85 ± 0.31	99.28 ± 0.18
Trial 2	✓			76.68 ± 1.05	94.25 ± 0.52	97.61 ± 0.33	88.29 ± 0.72	98.27 ± 0.28	99.49 ± 0.16
Trial 3	✓	✓		76.99 ± 1.09	95.35 ± 0.47	98.61 ± 0.26	88.63 ± 0.69	98.64 ± 0.24	99.63 ± 0.15
Trial 4	✓		✓	77.41 ± 0.98	95.65 ± 0.44	99.03 ± 0.22	89.32 ± 0.66	98.78 ± 0.25	99.60 ± 0.17
Trial 5	✓	✓	✓	78.64 ± 0.91	98.73 ± 0.29	99.32 ± 0.19	90.15 ± 0.61	99.31 ± 0.20	99.86 ± 0.09

Note: ✓ indicates that the corresponding module is included.

Trial 5 integrates all modules and achieves the best overall performance, showing that hierarchical features, contextual relations, and distribution-aware cues jointly improve spectrogram representation under limited samples.

As shown in Fig. 8, Trial 1 produces loose clusters with visible overlap, whereas HCF makes the embeddings more compact and RAMA or DAWM further improves separation. The full model yields the clearest clusters, confirming the complementary contributions of the three modules.

**Figure 8:** t-SNE visualization of ablation variants on the MIMII dataset under 6 dB SNR with 50 samples per class and 5-shot setting. (a) Trial 1; (b) Trial 2; (c) Trial 3; (d) Trial 4; (e) Trial 5.

Using the run-wise accuracies from the same five MIMII runs under 6 dB and 50 samples per class with matched random seeds and data splits, two-sided paired t -tests were conducted for the module comparisons in Table 11. For each comparison, the paired accuracy differences were defined as $d_i = A_i^{\text{with}} - A_i^{\text{without}}$. The null hypothesis was $H_0 : \bar{d} = 0$, the alternative hypothesis was $H_1 : \bar{d} \neq 0$, and $p < 0.05$ was regarded as statistically significant.

Table 11: Module contribution significance by two-sided paired t -test.

Contribution	Comparison	n	$\bar{d}$ (%)	s_d (%)	t	df	p -Value
HCF	Trial 2 vs. Trial 1	5	+1.92	0.44	9.85	4	5.95×10^{-4}
RAMA	Trial 3 vs. Trial 2	5	+1.10	0.19	12.80	4	2.15×10^{-4}
DAWM	Trial 4 vs. Trial 2	5	+1.40	0.23	13.64	4	1.67×10^{-4}
RAMA with DAWM	Trial 5 vs. Trial 4	5	+3.08	0.24	28.39	4	9.16×10^{-6}
DAWM with RAMA	Trial 5 vs. Trial 3	5	+3.38	0.29	25.78	4	1.34×10^{-5}

The positive $\bar{d}$ values show accuracy gains for all tested module additions, and all p -values are below 0.05, rejecting H_0 and confirming that the improvements are statistically significant rather than random sampling fluctuations.

4.5.2 The Impact of Fusion Weight Coefficient λ on the Model

Since λ controls the fusion between RAMA and DAWM, its sensitivity was evaluated by sweeping λ from 0.4 to 0.8 on the two main datasets, as shown in Fig. 9. The best settings were $\lambda = 0.6$ for MIMII with 50 samples per class and $\lambda = 0.7$ for PT Gear Fault with 10 samples per class. The different optima indicate that the two datasets place different demands on relation alignment and distribution modeling: MIMII contains stronger background noise, while PT Gear Fault involves more operating-condition variation. Overall, intermediate λ values best preserve the complementarity between support-query alignment and distribution-aware modeling.

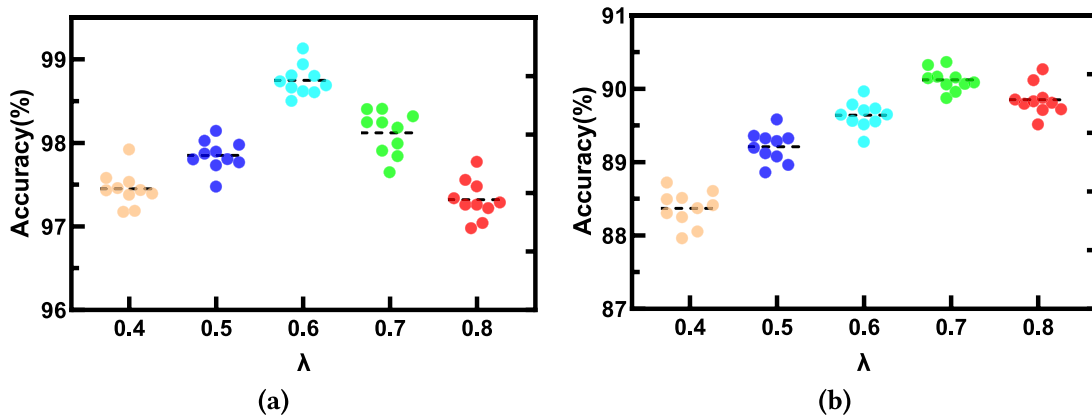


Figure 9: Sensitivity analysis of the fusion weight coefficient λ . The dashed horizontal markers indicate the mean accuracy at each λ value. (a) MIMII dataset with 50 samples per class; (b) PT Gear Fault dataset with 10 samples per class.

5 Conclusions

This work proposes HCF-RAMA-DAWM, a few-shot industrial acoustic fault diagnosis framework that combines multi-scale Mel-spectrogram feature extraction, support-query relation alignment, and Wasserstein distribution modeling. On MIMII and PT Gear Fault, it achieves 94.53% accuracy with 30 samples per class and 99.47% with 50 samples per class, respectively; CUMTB validation and real machine noise and machine-ID shift tests further support its generalization and robustness. Ablation, significance, λ -sensitivity, and complexity analyses confirm the complementary module contributions and an acceptable accuracy-complexity tradeoff. Overall, the framework is practical with limited labeled fault data, while future work will explore zero-shot diagnosis and broader validation under varying conditions.

Acknowledgement: Not applicable.

Funding Statement: The authors gratefully acknowledge the support of the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202400753).

Author Contributions: Conceptualization, Jialin Li and Shirong Li; methodology, Jialin Li; software, Jialin Li; validation, Yihong Liu, Yang Yang and Shirong Li; formal analysis, Jialin Li and Yihong Liu; investigation, Jialin Li and Yang Yang; resources, Yihong Liu and Shirong Li; data curation, Jialin Li and Yang Yang; writing—original draft

preparation, Jialin Li; writing—review and editing, Yihong Liu, Yang Yang and Shirong Li; visualization, Jialin Li; supervision, Shirong Li; project administration, Shirong Li; funding acquisition, Shirong Li. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The raw data supporting the conclusions of this article will be made available by the authors on request, subject to approval by the institute.

Ethics Approval: Not applicable.

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

References

1. Wang Z, Luo Q, Chen H, Zhao J, Yao L, Zhang J, et al. A high-accuracy intelligent fault diagnosis method for aero-engine bearings with limited samples. *Comput Ind.* 2024;159(3):104099. doi:10.1016/j.compind.2024.104099.
2. Yin H, Xu H, Zhang H, Wang J, Fang X. MC-ABDS: a system for low SNR fault diagnosis in industrial production with intense overlapping and interference. *Appl Acoust.* 2025;227(5):110217. doi:10.1016/j.apacoust.2024.110217.
3. Bornea YP, Vitor AL, Goedtel A, Castoldi MF, Souza WA, Barbara GV. A novel method for detecting bearing faults in induction motors using acoustic sensors and feature engineering. *Appl Acoust.* 2025;234:110627. doi:10.1016/j.apacoust.2025.110627.
4. Deng L, Zhao C, Yan X, Zhang Y, Qiu R. A novel approach for bearing fault diagnosis in complex environments using PSO-CWT and SA-FPN. *Measurement.* 2025;249:117027. doi:10.1016/j.measurement.2025.117027.
5. Wang H, Wang X, Yang Y, Gryllias K, Liu Z. A few-shot machinery fault diagnosis framework based on self-supervised signal representation learning. *IEEE Trans Instrum Meas.* 2024;73:1–14. doi:10.1109/tim.2024.3352689.
6. He Y, He D, Lao Z, Jin Z, Miao J, Lai Z, et al. Few-shot fault diagnosis of turnout switch machine based on flexible semi-supervised meta-learning network. *Knowl Based Syst.* 2024;294:111746. doi:10.1016/j.knosys.2024.111746.
7. Li X. Few-shot wind turbine blade damage early warning system based on sound signal fusion. *Multimed Syst.* 2023;29(5):2913–22. doi:10.1007/s00530-021-00882-7.
8. Xu Y, Fan RZ, He YL, Zhu QX, Zhang Y, Zhang MQ. DeepSMOTE with Laplacian matrix decomposition for imbalance instance fault diagnosis. *Chemom Intell Lab Syst.* 2025;259(3):105338. doi:10.1016/j.chemolab.2025.105338.
9. Fu Z, Liu Z, Ping S, Li W, Liu J. TRA-ACGAN: a motor bearing fault diagnosis model based on an auxiliary classifier generative adversarial network and transformer network. *ISA Trans.* 2024;149:381–93. doi:10.1016/j.isatra.2024.03.033.
10. Lu C, Ma X, Yan K. Chiller fault diagnosis based on improved variational autoencoder and co-training framework: a case study of insufficient samples. *J Build Eng.* 2024;88:109137. doi:10.1016/j.job.2024.109137.
11. Lao Z, He D, Sun H, He Y, Lai Z, Shan S, et al. Few-shot fault diagnosis of switch machine based on data fusion and balanced regularized prototypical network. *Eng Appl Artif Intell.* 2024;135:108847. doi:10.1016/j.engappai.2024.108847.
12. Zhang B, Li L, Liang G, Tan C, Dong F. Elastic slow feature prototypical network for few-shot fault diagnosis of industrial processes. *IEEE Sens J.* 2024;24(21):35264–75. doi:10.1109/jsen.2024.3461731.
13. Zhu Y, Li S, Lang X, Liu L. Multipath Siamese neural network for intelligent fault diagnosis under small samples. *IEEE Trans Instrum Meas.* 2025;74:1–15. doi:10.1109/tim.2025.3597681.
14. Wang D, Wang T, Wang X. Few-shot fault diagnosis for industrial robot transmission systems via a prototypical time-frequency mixer. *IEEE Access.* 2025;13:178045–59. doi:10.1109/access.2025.3620386.
15. Wang H, Lv D, Lin T, Han C, Song L. Task-adaptive unbiased regularization meta-learning for few-shot cross-domain fault diagnosis. *Eng Appl Artif Intell.* 2025;144:110200. doi:10.1016/j.engappai.2025.110200.
16. Xu M, Pan H, Wang S, Sun S. Transformer-embedded task-adaptive-regularized prototypical network for few-shot fault diagnosis. *Electronics.* 2025;14(19):3838. doi:10.3390/electronics14193838.
17. 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. doi:10.1016/j.energy.2024.130326.

18. Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain. p. 3637–45.
19. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 4080–90.
20. Sung F, Yang Y, Zhang L, Xiang T, Torr PHS, Hospedales TM. Learning to compare: relation network for few-shot learning. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT. p. 1199–208.
21. Purohit H, Tanabe R, Ichige K, Endo T, Nikaido Y, Suefusa K, et al. MIMII dataset: sound dataset for malfunctioning industrial machine investigation and inspection. arXiv:1909.09347. 2019.
22. Vu MH, Nguyen VQ, Tran TT, Pham VT, Lo MT. Few-shot bearing fault diagnosis via ensembling transformer-based model with Mahalanobis distance metric learning from Multiscale features. IEEE Trans Instrum Meas. 2024;73:1–18. doi:10.1109/tim.2024.3381270.
23. Eldele E, Ragab M, Qing X, Edward C, Chen Z, Wu M, et al. UniFault: a fault diagnosis foundation model from bearing data. arXiv:2504.01373. 2025.
24. Zhang X, Liu S, Jiang L, Li Y. FSAMLM: a few-shot adaptation multimodal large model for cross-domain fault diagnosis. Appl Soft Comput. 2025;185:113985.
25. Peng H, Liu J, Du J, Gao J, Wang W. BearLLM: a prior knowledge-enhanced bearing health management framework with unified vibration signal representation. Proc AAAI Conf Artif Intell. 2025;39(19):19866–74.
26. Guo Y, He J, Pu J, Wang M, Jiao B, Zhang D, et al. Vibration and acoustic data of pitch bearing in wind turbines under time-varying load for fault diagnosis. Data Brief. 2025;62(1):11187.