

Article

Multimodal Heterogeneous CNN with Adaptive Modality Fusion for Intelligent Fault Diagnosis of Bearings

Chang Sun, Chenkun Wang, Shiwei Huang and Tianci Zhang * 

School of Mechanical and Electrical Engineering, Central South University, Changsha 410083, China; 8210231909@csu.edu.cn (C.S.); 8203251020@csu.edu.cn (C.W.); 8203240705@csu.edu.cn (S.H.)

* Correspondence: modaoztc@csu.edu.cn

Abstract

In industrial equipment fault diagnosis, vibration and acoustic signals are highly complementary yet exhibit significant differences in frequency distribution and noise sensitivity. Traditional multimodal methods generally rely on homogeneous feature extractors and direct feature concatenation, which may fail to capture modality-specific characteristics and introduce irrelevant information during fusion. To address this, we propose a novel multimodal heterogeneous convolutional neural network framework. Specifically, separate 1D CNN branches are designed for vibration and acoustic signals. Their architectural differences are determined by the characteristics of each sensing modality. The vibration branch focuses on extracting high-level discriminative fault features, including impulse responses and modulated components from vibration signals, while the acoustic branch is designed to preserve fragile high-frequency details of acoustic signals. Furthermore, an adaptive cross-attention fusion module is introduced to dynamically model cross-modal dependencies, assigning Softmax-based weights to enhance dominant features and suppress noise. Experiments based on bearing fault experimental data demonstrate that the proposed heterogeneous architecture significantly outperforms traditional homogeneous models. The dynamic weighting mechanism effectively prevents inferior noisy modalities from degrading overall performance, achieving high diagnostic accuracy. Although validated on rolling bearing fault diagnosis, the proposed heterogeneous multimodal framework is not restricted to bearings and can be readily extended to other intelligent condition monitoring tasks involving heterogeneous sensor fusion, such as gearboxes, motors, and other rotating machinery.

Keywords: bearing fault diagnosis; multimodal feature fusion; adaptive cross-attention; heterogeneous CNN



Academic Editor: Xuefang Xu

Received: 15 June 2026

Revised: 24 July 2026

Accepted: 27 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Rolling bearings are among the most critical components in rotating machinery. Their health condition directly affects the reliability, operational safety, and maintenance cost of machinery [1]. As industrial systems become increasingly high-speed, complex and continuously operated, rolling bearing fault diagnosis has emerged as a central research focus for condition monitoring and predictive maintenance [2]. In recent years, deep learning has emerged as a key technological approach for intelligent fault diagnosis of rolling bearings, owing to its end-to-end feature extraction capability [3]. Comprehensive systematic reviews have confirmed that approaches including convolutional neural networks (CNNs) [4], autoencoders (AEs) [5], generative adversarial networks (GANs) [6] and transfer learning [7]

exhibit promising application potential for rolling bearing fault identification [3]. Overall, deep learning methods have provided a solid technical foundation for rolling bearing fault diagnosis and have facilitated the gradual shift toward automation and intelligence in this field.

In rolling bearing fault diagnosis, vibration signals effectively reflect the impact responses and changes in structural condition caused by localized damage. Therefore, they remain one of the most widely used sources of diagnostic information [8]. With the advancement of deep learning methods, vibration-based intelligent diagnosis methods have advanced substantially, and related research has gradually shifted from relying on manually designed features to utilizing deep learning models to extract fault information from vibration responses. For example, Ding et al. proposed a time-frequency Transformer model based on the self-attention mechanism to extract more discriminative fault features from vibration signals [9]. Wang et al. developed an attention-guided joint learning CNN that combines bearing fault diagnosis with vibration signal denoising to improve the model's diagnostic robustness in noisy environments [10]. Han et al. proposed a rolling bearing fault diagnosis system that combines convolutional neural networks (CNNs) and support vector machines (SVMs). The system takes time-domain vibration data as input, uses CNNs for feature extraction, and employs SVMs to improve classification performance and generalization capabilities under low-sample-size conditions [11]. However, in real-world industrial settings, vibration measurements typically rely on contact sensors, and the quality of the data collected is easily affected by factors such as sensor mounting methods, measurement point locations, and structural transmission paths, thereby limiting the ability of intelligent diagnostic models based on vibration data to characterize complex operating conditions [12]. Furthermore, when the number of labeled samples is limited, deep neural networks relying solely on vibration data are prone to overfitting, which can result in poor model generalization [13].

In addition to vibration signals, acoustic signals, owing to their non-contact acquisition capability, can reflect bearing operating conditions from the perspective of acoustic radiation. Therefore, they are widely regarded as an effective source of diagnostic information and serve as a complement to vibration signals [14]. There have been studies combining vibration signals with acoustic signals for bearing fault diagnosis. For example, You et al. proposed a deep learning method guided by constraints derived from the fusion of acoustic and vibration data, incorporating multi-physics information from acoustic and vibration into the diagnostic model. This approach not only improved the accuracy of bearing fault detection but also enhanced the model's interpretability [15]. Shi et al. proposed a two-stage acoustic-vibration fusion method that enhances the detection of subtle faults in rolling bearings by integrating acoustic and vibration data from multiple measurement points [16]. These studies indicate that incorporating acoustic information into the intelligent fault diagnosis process enables models to obtain more comprehensive fault characterization than vibration input alone, making combined acoustic-vibration diagnosis a promising research direction in intelligent bearing fault diagnosis.

On the other hand, diagnosing bearing faults with limited labeled data is also a key challenge in this field. Because it is costly and difficult to obtain and label fault samples during the actual operation of industrial equipment, deep learning models are prone to overfitting and poor generalization under few-shot conditions [17,18]. To improve fault diagnosis performance under limited labeling conditions, researchers have begun to incorporate self-supervised learning into intelligent bearing fault diagnosis [19]. For example, Cui et al. proposed a vibration signal-based bearing fault diagnosis model using self-attention contrastive learning, which employs unlabeled vibration data for self-supervised feature pre-training and is fine-tuned using a small number of labeled samples,

thereby improving the model's fault recognition capability under conditions of limited labeled data [20]. Furthermore, few-shot learning and meta-learning methods have also been applied to the fault diagnosis of rotating machinery to enhance the models' ability to adapt quickly under conditions of extremely limited data [21]. Li et al. proposed a model-agnostic meta-learning-based method for few-shot bearing fault diagnosis. By learning transferable prior knowledge of faults from multi-task vibration data, the method enables rapid identification of bearing faults under complex operating conditions [22]. However, based on existing research, most of these methods still focus on single-modal vibration signals, and it remains challenging to diagnose rolling bearing faults using multimodal vibration and acoustic data when there are limited fault samples.

Overall, while existing research on intelligent bearing diagnosis based on multimodal data has achieved promising results, it still has two main limitations. First, most existing studies use the same network architecture to extract fault features from vibration and acoustic data. Nevertheless, data from different modalities exhibit significant differences in physical properties, frequency band distribution, and information density. If a homogeneous network is used for feature extraction, it is often difficult to simultaneously address the representation requirements of different modalities, resulting in insufficient feature extraction for some modalities or excessive compression of detailed information [23]. Second, most existing studies employ simple concatenation of features to achieve feature fusion between acoustic and vibration. However, under few-shot conditions, without effective modality quality assessment and dynamic weighting mechanisms, simple fusion methods are prone to introducing interference from low-quality modalities, thereby weakening the discriminative power of the fused features [24,25].

To address the above issues, this paper proposes a multimodal heterogeneous convolutional neural network with adaptive modality fusion, and validates its effectiveness through the task of few-shot intelligent fault diagnosis of rolling bearings. Unlike traditional homogeneous dual-branch architectures, this paper constructs separate one-dimensional convolutional feature extraction branches with distinct structural configurations based on the differences in information density, frequency band distribution, and temporal resolution between vibration and acoustic signals, thereby enhancing the specificity of feature extraction for different modalities. During feature fusion, an adaptive cross-attention mechanism is introduced to model interactions between the two modalities. It dynamically allocates modality weights at the sample level, preserving complementary information while suppressing interference from low-quality features. Although the proposed framework is validated on rolling bearing fault diagnosis in this study, its heterogeneous feature extraction and adaptive cross-modal fusion strategy are designed according to the characteristics of heterogeneous sensing modalities rather than the characteristics of rolling bearings themselves. Therefore, the proposed framework also has the potential to be extended to other intelligent condition monitoring tasks involving multimodal vibration and acoustic information, such as gearbox fault diagnosis, motor condition monitoring, and health assessment of other rotating machinery. This broader applicability further highlights the practical engineering value of the proposed framework.

The main contributions of this paper are summarized as follows:

1. A multimodal heterogeneous convolutional neural network framework is proposed for few-shot bearing fault diagnosis. The framework is specifically designed to extract features tailored to the inherent differences between vibration and acoustic signals, thereby avoiding representation imbalance and loss of fine-grained information that may arise in homogeneous network architectures.
2. An adaptive cross-attention fusion mechanism is developed. Through bidirectional cross-modal interaction and dynamic weight allocation, it enables sample-level adaptive

adjustment of the contribution levels of different modalities, resulting in more effective fusion of vibration and acoustic features.

3. Systematic experimental validation is conducted based on bearing fault experimental data. The experiments validated the effectiveness of the proposed method under few-shot conditions with 1–20 training samples per class, demonstrating that the method exhibits significant advantages over comparative methods in terms of diagnostic accuracy, stability, and engineering application potential.

The remainder of this paper is organized as follows: Section 2 presents the overall framework of the proposed multimodal heterogeneous convolutional neural network. Section 3 conducts experimental analysis of the proposed method based on bearing fault datasets. Section 4 discusses its performance advantages, limitations, and future research directions. Section 5 summarizes the main contributions and conclusions of this paper.

2. Materials and Methods

2.1. Overall Framework

In actual industrial operating environments, although vibration signals and acoustic radiation signals share the same source, they differ significantly in terms of transmission paths, frequency band distribution, and sensitivity to ambient noise [26,27]. Traditional homogeneous network feature extraction often results in the loss of high-frequency details in acoustic signals, while direct feature concatenation can easily lead to low-quality features contaminating high signal-to-noise ratio vibration features [28]. To address the above issues, this paper proposes a novel multimodal heterogeneous intelligent diagnosis framework.

As shown in Figure 1, the proposed method consists of three components: a multimodal heterogeneous feature extraction module, an adaptive cross-attention fusion module, and a fault classification module. Given vibration and acoustic sequences of length L , denoted as x_v and x_a respectively, where $x_v \in \mathbb{R}^{L \times 1}$ and $x_a \in \mathbb{R}^{L \times 1}$, the model first uses a heterogeneous two-branch 1D convolutional network to extract vibration and acoustic modal features. It then establishes cross-modal dependencies through bidirectional cross-attention and performs sample-level dynamic weighting using Softmax weights. Finally, the fused features are fed into a fully connected classifier, which outputs the predicted fault categories. The computational workflow of the entire model can be represented as follows:

$$F_v = f_v(x_v; \theta_v) \quad (1)$$

$$F_a = f_a(x_a; \theta_a) \quad (2)$$

$$F_{fused} = \Psi(F_v, F_a) \quad (3)$$

$$\hat{y} = C(F_{fused}) \quad (4)$$

where F_v and F_a denote the high-dimensional feature representations of the vibration and acoustic modalities, respectively. $f_v(\cdot)$ and $f_a(\cdot)$ represent the feature extraction mappings for the corresponding modalities, respectively. θ_v and θ_a are the trainable parameters for the respective branches. $\Psi(\cdot)$ denotes the adaptive cross-attention fusion function. F_{fused} represents the final fused representation. $C(\cdot)$ denotes the fault classifier, and $\hat{y}$ is the output class probability distribution.

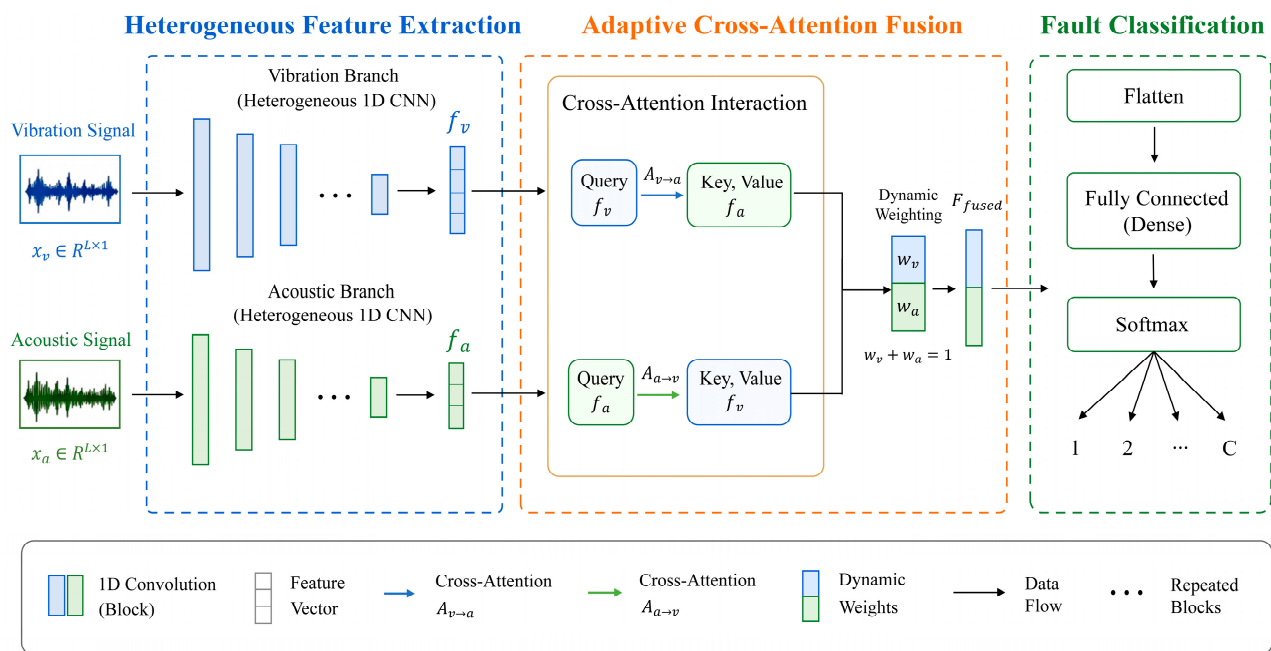


Figure 1. The overall architecture of the proposed multimodal heterogeneous CNN with adaptive modality fusion.

2.2. Multimodal Heterogeneous Feature Extraction

Vibration signals and acoustic signals differ significantly in terms of their physical generation mechanisms, frequency distributions, and sensitivity to noise [29,30]. Vibration signals typically contain stronger structural impulse responses and modulated components, requiring deeper networks to expand the receptive field and extract high-level discriminative features. In contrast, acoustic signals often preserve more fragile high-frequency details. Using excessively deep convolutional and pooling operations can easily lead to a rapid decline in temporal resolution, resulting in the loss of critical information. Therefore, this paper adopts a heterogeneous two-branch structure rather than a homogeneous counterpart, and designs separate network architectures for the vibration branch and the acoustic branch based on the signal characteristics of the input modalities.

Following the standard one-dimensional convolutional neural network architecture [4], the basic operation of the l -th convolutional module can be expressed as

$$h^{(l)} = \text{Pool}\left(\sigma\left(\text{Conv1D}\left(h^{(l-1)}\right)\right)\right) \quad (5)$$

where $h^{(l-1)}$ and $h^{(l)}$ represent the output feature maps of the $(l-1)$ -th and l -th layers, respectively. $\text{Conv1D}(\cdot)$ denotes the one-dimensional convolution operation, $\sigma(\cdot)$ denotes the nonlinear activation function, and $\text{Pool}(\cdot)$ denotes the max-pooling operation.

In this study, LeakyReLU is employed as the nonlinear activation function [31], which is formulated as

$$\sigma(x) = \begin{cases} x, & x \geq 0 \\ \alpha x, & x < 0 \end{cases} \quad (6)$$

where α represents the negative slope coefficient and is set to 0.2 in this work. Compared with the conventional ReLU function, LeakyReLU preserves a nonzero gradient in the negative input region, thereby mitigating the dying ReLU problem and facilitating more stable network training.

Based on the above formulations, the vibration and acoustic branches are both composed of stacked one-dimensional convolutional modules, with their overall feature extrac-

tion mappings defined in Equations (1) and (2), respectively. To accommodate the inherent differences between vibration and acoustic signals, distinct architectural configurations are employed for the two branches. The vibration branch is tailored to capture global fault-related semantic patterns, while the acoustic branch is designed to retain fine-grained high-frequency details during the hierarchical feature abstraction process. This heterogeneous architecture enables the extraction of modality-specific representations and provides more complementary feature representations for subsequent cross-modal fusion.

2.3. Adaptive Cross-Attention Fusion Mechanism

After obtaining the vibration feature representation F_v and the acoustic feature representation F_a , direct feature concatenation may introduce residual environmental noise from the acoustic modality into the fused representation, thereby degrading the quality of multimodal feature fusion. To alleviate this issue, an adaptive cross-attention fusion mechanism is proposed. The proposed mechanism consists of two stages: cross-modal interaction and dynamic feature weighting [32,33].

2.3.1. Cross-Modal Interaction

To model the correlation between the two modalities, a bidirectional cross-attention mechanism is developed based on the scaled dot-product attention proposed in the Transformer architecture [34]. Specifically, the vibration features are used as the query, while the acoustic features are used as the keys and values, thereby establishing a cross-attention operation from vibration to acoustic:

$$Q_v = F_v W_Q^v \quad (7)$$

$$K_a = F_a W_K^a \quad (8)$$

$$V_a = F_a W_V^a \quad (9)$$

where Q_v , K_a , and V_a denote the query matrix of the vibration modality, the key matrix of the acoustic modality, and the value matrix of the acoustic modality, respectively. W_Q^v , W_K^a , and W_V^a are the corresponding trainable projection matrices.

Accordingly, the vibration-guided acoustic representation can be formulated as

$$A_{v \rightarrow a} = \text{Softmax} \left(\frac{Q_v K_a^T}{\sqrt{d_k}} \right) V_a \quad (10)$$

where $A_{v \rightarrow a}$ denotes the cross-modal enhanced representation obtained by querying acoustic features with vibration features. d_k is the dimensionality of the key vectors and is introduced to scale the dot-product attention scores for improved numerical stability during training [34]. This process enables the model to utilize the relatively stable structural response information contained in the vibration modality to impose conditional constraints on the acoustic modality, thereby highlighting fault-related acoustic components and reducing noise interference.

Similarly, by using the acoustic features as the queries and the vibration features as the keys and values, the vibration-enhanced features guided by the acoustic modality can be obtained as

$$A_{a \rightarrow v} = \text{Softmax} \left(\frac{Q_a K_v^T}{\sqrt{d_k}} \right) V_v \quad (11)$$

where $A_{a \rightarrow v}$ denotes the cross-modal enhanced representation obtained by querying the vibration features with the acoustic features. Q_a , K_v , and V_v denote the acoustic query matrix, vibration key matrix, and vibration value matrix, respectively. Corresponding

to Equation (10), this process enables the complementary information contained in the acoustic modality to be reversely transferred to the vibration features, thereby further enhancing the correlation modeling between the two modalities.

Through Equations (10) and (11), the proposed model is able to establish bidirectional cross-modal dependency relationships before feature fusion, thereby enhancing the sharing of fault-related information and suppressing noise components unrelated to faults to a certain extent.

2.3.2. Dynamic Feature Weighting

Since the signal quality of the two modalities varies across different samples, relying solely on cross-modal interaction is insufficient to achieve optimal fusion. To address this, this paper further introduces a Softmax-based dynamic weighting mechanism to allocate modal contributions sample-wise.

First, the bidirectionally enhanced cross-modal features are concatenated to obtain the globally fused feature vector:

$$z = A_{v \rightarrow a} \oplus A_{a \rightarrow v} \quad (12)$$

where z denotes the globally fused representation obtained by concatenating the bidirectionally enhanced cross-modal features, and $\oplus$ represents the feature concatenation operation.

Subsequently, the adaptive weights for the two modalities are generated through a fully connected layer followed by a Softmax function:

$$[w_v, w_a] = \text{Softmax}(W_{\text{attn}}z + b_{\text{attn}}) \quad (13)$$

where W_{attn} and b_{attn} denote the parameters of the weight generation layer, and w_v and w_a represent the importance weights of the vibration and acoustic modalities, respectively. Through this weight generation process, the model can adaptively adjust the relative contributions of the two modalities to the final diagnostic outcome based on the quality of the features of each modality in the current sample.

The weights of the two modalities satisfy the normalization constraint:

$$w_v + w_a = 1 \quad (14)$$

Equation (14) indicates that the weights of the two modalities are normalized through the Softmax function such that their sum equals 1, thereby ensuring relative contribution allocation at the sample level.

Finally, the fused feature representation is formulated as

$$F_{\text{fused}} = w_v A_{v \rightarrow a} \oplus w_a A_{a \rightarrow v} \quad (15)$$

Equation (15) shows that the proposed method does not employ fixed equal-weight fusion. Instead, it dynamically adjusts the contributions of the two modalities according to the actual information quality of each modality in a given sample. When the acoustic modality is severely contaminated by noise, its weight is automatically reduced; when it provides effective complementary information, its weight is correspondingly increased. This mechanism allows the model to preserve the complementarity between modalities while avoiding low-quality modalities dominating the fusion space, thereby improving the robustness of multimodal fusion under few-shot conditions [35].

2.4. Fault Classification and Objective Function

After obtaining the fused feature representation F_{fused} , the model flattens it and feeds it into a fully connected classification layer. The Softmax classifier [34] is then employed to

generate the probability distribution of each fault class. Suppose there are C fault classes, the classification output can be expressed as

$$\hat{y} = \text{Softmax}(W_c F_{\text{fused}} + b_c) \quad (16)$$

where W_c and b_c denote the weight matrix and bias term of the classification layer, and $\hat{y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_C]$ represents the predicted probability of the sample belonging to each fault class.

During training, the cross-entropy loss function, commonly used in multi-class classification tasks, is employed as the optimization objective. Let the training set consist of N samples, C denote the total number of fault classes, $y_{i,c}$ represent the ground-truth label of the i -th sample for class c , and $\hat{y}_{i,c}$ denote the predicted probability of the i -th sample belonging to class c . The loss function is then defined as

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}) \quad (17)$$

By minimizing Equation (17), the model jointly optimizes all parameters in the heterogeneous feature extraction branches, the adaptive cross-attention fusion module, and the classification layer, enabling the predicted results to gradually approximate the ground-truth label distribution.

In summary, the proposed method constructs a complete diagnostic pipeline, including heterogeneous feature extraction, bidirectional cross-modal interaction, dynamic modality weighting, and fault classification. This framework fully exploits the complementary advantages of vibration and acoustic modalities in fault diagnosis and effectively suppresses the interference of noisy modalities on decision boundaries. As a result, it provides a clear methodological foundation for multimodal, vibration- and acoustic-driven intelligent diagnosis of rolling bearing faults.

3. Validity Analysis Based on Rolling Bearing Fault Experimental Data

3.1. Experimental Setup and Data Description

In this study, to validate the effectiveness of the proposed method, the rolling bearing fault dataset released by the University of Ottawa was employed [36]. The dataset was collected from a bearing test rig operating under constant rotational speed and constant load conditions. The data acquisition system mainly consisted of a driving motor, the tested rolling bearing, an accelerometer, a microphone, a load sensor, a Hall-effect speed sensor, a signal conditioner, and a data acquisition device, which were used to synchronously acquire vibration and acoustic signals during bearing operation.

For these two types of signals, tests were conducted on the inner race, outer race, rolling elements, and cage of 20 bearings under constant nominal rotational speed and load conditions. A total of 60 datasets were collected, covering three health conditions: healthy, developing fault, and fault. Based on the original dataset, seven diagnostic conditions were further selected and organized in this study, including normal condition (NC-0), developing inner-race fault (IF-1), inner-race fault (IF-2), outer-race fault (OF-1), ball fault (BF-1), developing cage fault (CF-1), and cage fault (CF-2), which were used to construct a seven-class fault diagnosis task.

As illustrated in Figure 2, the representative fault types correspond to defects located on different bearing components, including the inner race, outer race, and rolling element. The defect locations are indicated by red circles to provide a clear visualization of the physical characteristics of the investigated bearing conditions.

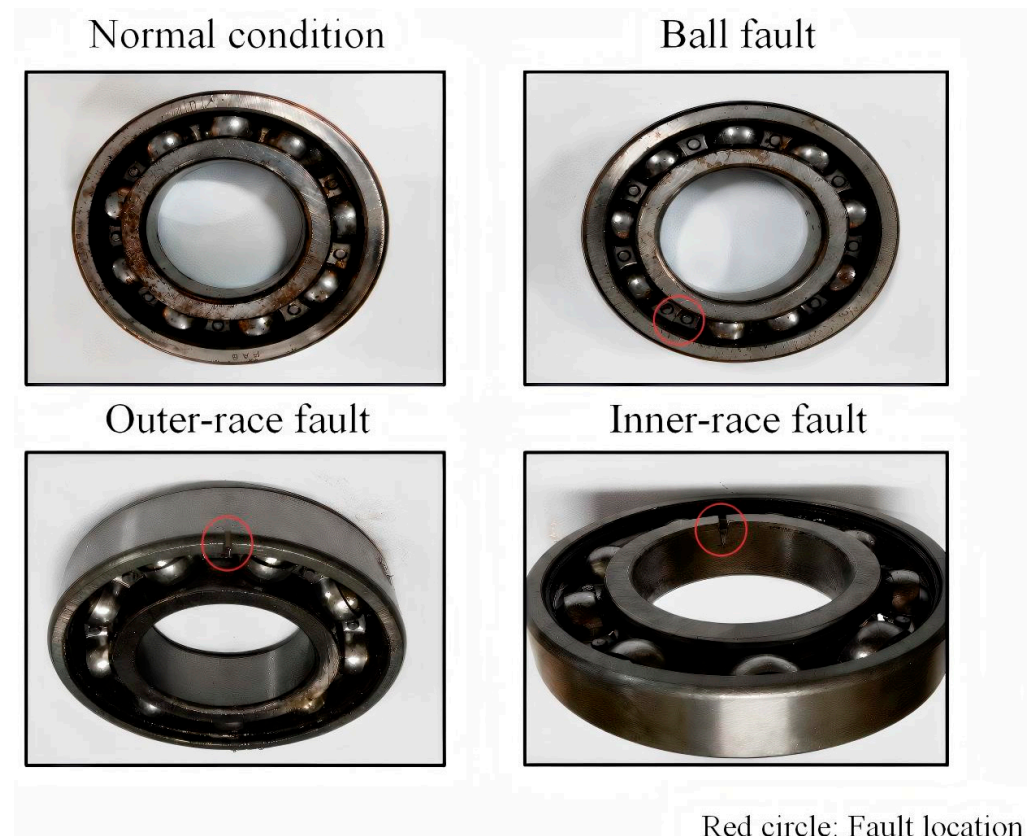


Figure 2. Representative photographs of different bearing conditions.

To ensure synchronized acquisition of multimodal information, the experimental platform was equipped with dedicated sensors for vibration, acoustic emission, rotational speed, load, and temperature measurements. All sensing devices were connected to a unified data acquisition system through a signal conditioner, enabling simultaneous recording of multiple sensor outputs. Detailed information regarding the instrumentation is summarized in Table 1.

Table 1. Instrumentation used for data acquisition.

Instrument	Manufacturer	Model	Function
Accelerometer	PCB Piezotronics, Depew, NY, USA	623C01	Vibration acquisition
Microphone	PCB Piezotronics, Depew, NY, USA	130F20	Acoustic acquisition
Load Cell	OMEGA Engineering, Norwalk, CT, USA	LCM302	Load measurement
Hall-effect Sensor	OMEGA Engineering, Norwalk, CT, USA	OMDC-MPU-A	Rotational speed measurement
Thermocouple	OMEGA Engineering, Norwalk, CT, USA	KTSS-HH	Temperature monitoring
Signal Conditioner	PCB Piezotronics, Depew, NY, USA	482C	Signal conditioning
Data Acquisition System	National Instruments, Austin, TX, USA	USB-6212	Synchronous data acquisition

The vibration and acoustic signals were synchronously acquired using the same data acquisition system at an identical sampling frequency of 42 kHz. Each recording lasted 10 s, resulting in 420,000 sampling points for each signal channel. Since both sensing

modalities shared the same acquisition hardware and sampling frequency, no additional temporal synchronization or resampling was required before subsequent signal processing and multimodal fusion.

To further illustrate the differences in vibration and acoustic responses under different bearing conditions, representative signal segments corresponding to the seven diagnostic conditions are presented in Figure 3.

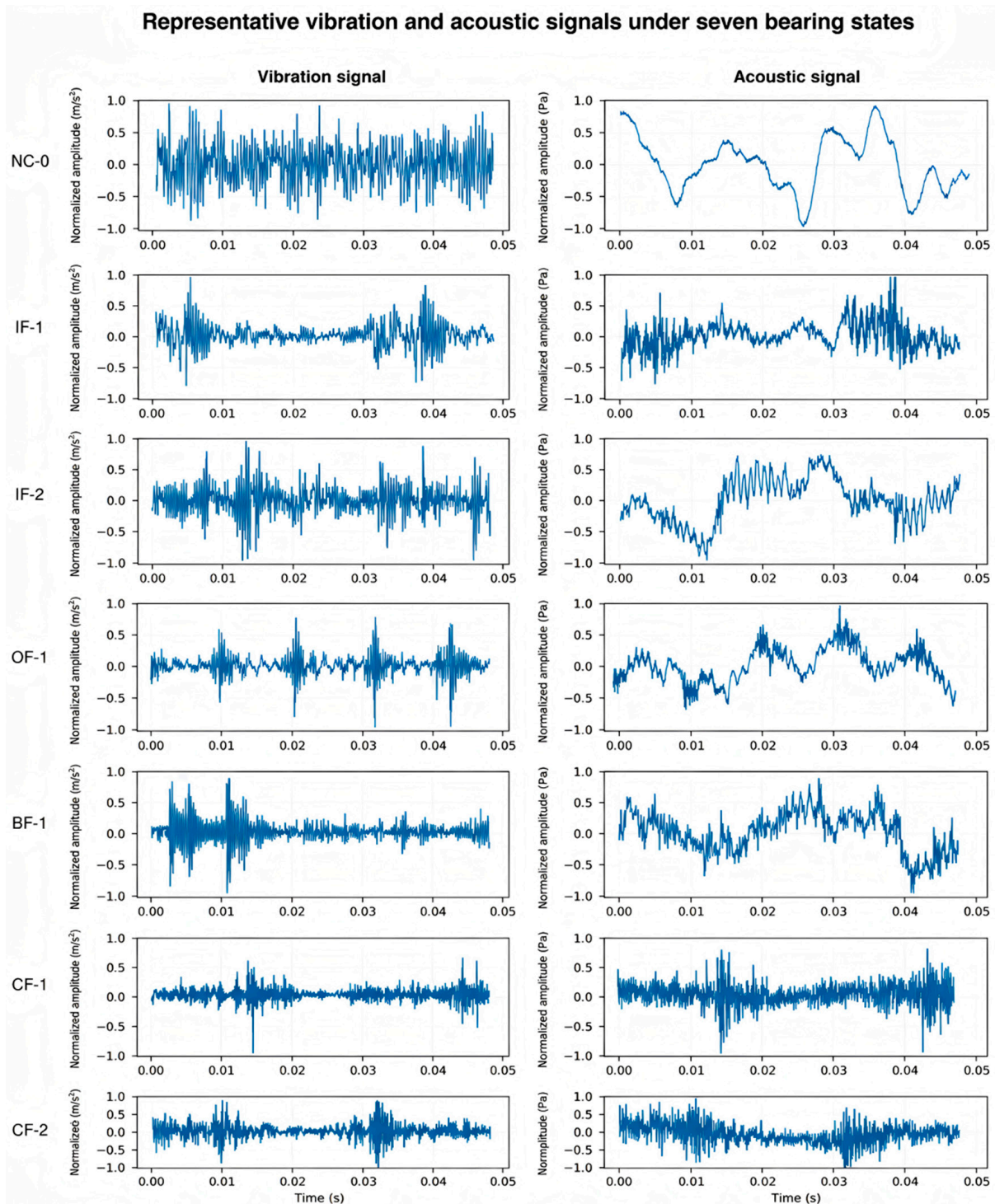


Figure 3. Time-domain waveforms of vibration and acoustic signals under seven bearing health conditions.

The Adamax optimizer with a learning rate of 0.001 was employed in this study, and the model was trained for 100 epochs. The data processing and model training were con-

ducted on a computer running the 64-bit Windows 11 operating system, equipped with a 13th Gen Intel® Core™ i9-13900HX CPU @ 2.20 GHz and an NVIDIA GeForce RTX 4060 Laptop GPU. Python 3.8 was used as the programming language, and TensorFlow 2.4 was adopted as the machine learning framework.

In this study, a heterogeneous dual-branch convolutional network was employed to process vibration signals and acoustic signals separately, with different network configurations adopted for the two branches. The corresponding design principles have been described in the previous section. Here, the network parameter settings consistently used in the subsequent experiments are presented. The detailed configuration is listed in Table 2.

Table 2. Specific structural parameters of the multimodal heterogeneous CNN.

Network Branch	Convolutional Layers	Output Channels	Kernel Size /Stride	Activation Function /Pooling
Vibration Branch	Conv1–Conv2	32	First Layer: 16/1 Other Layers: 4/1	LeakyReLU ($\alpha = 0.2$) MaxPooling
	Conv3–Conv4	64		
	Conv5–Conv6	128		
	Conv7–Conv8	256		
	Conv9–Conv10	512		
Acoustic Branch	Conv1–Conv2	32	First Layer: 16/1 Other Layers: 4/1	LeakyReLU ($\alpha = 0.2$) MaxPooling
	Conv3–Conv4	64		
	Conv5–Conv6	128		
	Conv7–Conv8	256		
	Conv9	512		
Feature Fusion & Classification	Bidirectional Cross-Attention + Dynamic Weighting	/		Weighted Fusion + Dense Classification Output

To ensure the reproducibility of the experimental results, a consistent data preprocessing and partitioning strategy was adopted under different training sample sizes. The originally synchronized vibration and acoustic signals were first segmented into non-overlapping windows with a fixed length of 2048 sampling points. Vibration segments and acoustic segments corresponding to the same temporal position were then paired to form bimodal samples, thereby preserving the temporal correspondence between the two modalities. Each bimodal sample consists of one vibration segment and one acoustic segment corresponding to the same temporal position, with each segment containing 2048 consecutive sampling points. This window-based sample construction strategy has also been widely adopted in previous few-shot fault diagnosis studies [17].

Subsequently, 1–20 bimodal samples from each fault category were randomly selected as the training set, while the remaining bimodal samples were used as the test set. Both the training and test sets were constructed from bimodal samples generated by segmenting the signals collected from the same bearing. The training and testing sets were mutually exclusive, and each bimodal sample was assigned exclusively to either the training set or the testing set without any overlap. To reduce the randomness introduced by data partitioning, the experiments were repeated 10 times using different random seeds, and the average values of all evaluation metrics were reported.

Except for the number of training samples per class, the input length, optimizer, learning rate, number of training epochs, and experimental environment were kept consistent throughout all experiments.

3.2. Comparison Methods

This study selected three relevant baseline methods and two advanced multimodal diagnostic algorithms published in recent years as comparison benchmarks to verify the superiority of the proposed method. Specific details are as follows:

1. Relevant Baseline Diagnostic Methods

DCNN-V: A unimodal 1D convolutional neural network utilizing only vibration signal input. This network employs a 10-layer structure identical to the vibration branch in the proposed method, testing the diagnostic capability of vibration data when operating independently.

DCNN-A: A unimodal 1D convolutional neural network utilizing only acoustic signal input. This network employs a 9-layer structure identical to the acoustic branch in the proposed method, testing the diagnostic capability of acoustic data when operating independently.

DCNN-Concat: A baseline multimodal feature concatenation network. This method adopts the same dual-input architecture as the proposed method. However, the attention mechanism is removed after feature extraction. The extracted features from the two modalities are directly concatenated end-to-end into a single feature vector. This vector is subsequently fed into a fully connected layer for fault classification.

2. Advanced Multimodal Diagnostic Algorithms

Two representative advanced fusion algorithms published in recent years in the field of machinery fault diagnosis were selected to demonstrate the superiority of the proposed method over existing state-of-the-art techniques:

Multimodal Residual Convolutional Fusion Network (MRCFN) [37]: This model introduces a residual network architecture and employs skip connections to construct a deep multisensor feature extractor. Combined with a multimodal fusion strategy, it enhances bearing fault diagnosis performance under noisy and few-shot conditions.

Multimodal Self-supervised Learning Network (MSLN) [38]: This method constructs a multimodal self-supervised learning framework using vibration and acoustic data. A homogeneous dual-encoder structure and a contrastive learning mechanism are employed for pretraining to enhance feature representation capability. The learned representations are subsequently utilized for few-shot rolling bearing fault identification.

To ensure a fair comparison among different methods, all models were evaluated under the same experimental settings, including identical data partitioning strategies, input signal lengths, training epochs, optimizer parameters, and evaluation criteria. For the baseline models, including DCNN-V, DCNN-A, and DCNN-Concat, the network training framework was implemented in a manner consistent with that of the proposed method. For the two advanced multimodal approaches, namely Multimodal Residual Convolutional Fusion Network (MRCFN) and Multimodal Self-supervised Learning Network (MSLN), the implementations were reproduced according to the publicly available network architectures described in the original papers. Only the input dimensions and output layer sizes were modified to match the seven-class classification task, while the core network architectures and fusion mechanisms remained unchanged. Training hyperparameters were adjusted when necessary to ensure compatibility with the Ottawa bearing dataset, without altering the original model structures. This ensured that all methods could be consistently applied to the few-shot seven-class fault diagnosis task.

3.3. Evaluation Criteria

A multidimensional evaluation metric system was employed in this study to comprehensively and objectively assess the performance of the proposed heterogeneous multimodal CNN model in the rolling bearing fault diagnosis task. In addition to the most intuitive metric, accuracy, precision and F1-score were also adopted to evaluate the model's recognition capability across different fault classes. Moreover, the receiver operating characteristic (ROC) curve was introduced to assess the overall performance of the model under varying classification thresholds. Confusion matrices and t-SNE feature visualization tech-

niques were further employed to provide an intuitive analysis of the model's classification details and feature learning effectiveness.

1. Accuracy

Accuracy is the most commonly used metric to evaluate the overall performance of a model, representing the proportion of correctly classified samples to the total number of test samples. It is calculated as follows [39]:

$$Accuracy = \frac{N_{correct}}{N_{total}} \times 100\% \quad (18)$$

where $N_{correct}$ denotes the number of correctly predicted samples, and N_{total} denotes the total number of test samples.

2. Precision and Recall

In industrial fault diagnosis, it is not sufficient to evaluate only the classification accuracy; the sensitivity of the model to specific fault types must also be considered. Precision reflects the proportion of samples predicted as a certain fault that actually belong to that fault class. The higher the precision, the lower the model's false positive rate. Recall reflects the proportion of all actual samples of a certain fault that are correctly identified by the model. The higher the recall, the lower the model's false negative rate.

3. F1-score

The F1-score is the harmonic mean of precision and recall and is used to comprehensively evaluate model performance. In industrial applications, the occurrence frequency of certain severe faults is often much lower than that of normal operating conditions, resulting in an imbalanced distribution of data samples. In such cases, if only accuracy or precision is considered, the model may achieve a high score by favoring the majority class, namely normal samples. In contrast, the F1-score simultaneously considers both false positives and false negatives, thereby providing a more objective evaluation. It is calculated as follows [39]:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (19)$$

4. ROC Curve and AUC

The receiver operating characteristic (ROC) curve is a visualization tool for comprehensively evaluating the performance of classification models. For the seven-class fault diagnosis task considered in this study, the macro-averaging strategy was adopted to calculate the false positive rates and true positive rates of all fault classes, thereby generating a global ROC curve. The area under the ROC curve (AUC) represents the area enclosed by the ROC curve and is typically within the range of 0.5 to 1. A higher AUC value indicates better discrimination capability of the model for different fault characteristics.

5. Confusion Matrix

To further analyze the performance of the model on each specific fault category, confusion matrices were employed for visualization. In a confusion matrix, each column represents the ground-truth labels of the samples, while each row represents the labels predicted by the model.

6. t-SNE Feature Visualization

To verify the effectiveness of feature extraction by the heterogeneous CNN, the t-distributed stochastic neighbor embedding (t-SNE) algorithm was employed to reduce the dimensionality of high-dimensional features. By projecting the extracted high-dimensional features onto a two-dimensional plane and observing the distribution of data points with different colors (i.e., the clustering patterns of different fault conditions), the discriminative capability of the learned fault features can be intuitively evaluated.

3.4. Analysis of Diagnostic Results Under Few-Shot Conditions

In practical industrial environments, equipment typically operates under normal conditions for extended periods, making fault samples extremely difficult to obtain. In addition, manual annotation is often costly and labor-intensive. Therefore, the feature learning capability and generalization performance of intelligent diagnostic models under limited training samples are important criteria for evaluating their practical engineering value.

To verify the robustness of the proposed heterogeneous multimodal CNN model with adaptive modality fusion under few-shot conditions, variable-sample-size experiments were designed in this study. The number of training samples for each fault category was set to 1, 5, 10, 15, and 20, respectively, while the remaining samples were used as the test set.

1. Classification Performance Analysis under Different Training Sample Sizes

The average performance of the proposed model in terms of Accuracy, Precision, and F1-score under different training sample sizes is presented in Table 3.

Table 3. Performance of the proposed model under different training sample sizes.

Number of Training Samples	Accuracy (%)	Precision (%)	F1-Score (%)
1	56.20 ± 7.13	62.09 ± 8.75	52.84 ± 8.35
5	86.67 ± 4.11	88.70 ± 3.99	86.35 ± 4.33
10	95.82 ± 1.25	96.26 ± 1.10	95.69 ± 1.30
15	96.49 ± 2.31	96.94 ± 1.82	96.48 ± 2.38
20	97.83 ± 2.52	97.96 ± 2.33	97.81 ± 2.60

As can be observed from Table 3, when the number of training samples is reduced to an extreme level (i.e., only one sample per class), the model is unable to effectively learn fault decision boundaries due to insufficient data representation. Consequently, all evaluation metrics remain at relatively low levels and exhibit considerable fluctuations. When the number of training samples increases to five per class, the diagnostic performance improves significantly, benefiting from the complementary feature-learning capability of the multimodal network. As the number of training samples further increases to 10 or more per class, the feature learning capability of the model is effectively exploited, and both Accuracy and F1-score rapidly increase to above 95% and become relatively stable.

2. ROC Curve Analysis under Different Training Sample Sizes

To further evaluate the classification sensitivity and specificity of the proposed model under few-shot conditions, Figure 4 presents the macro-average ROC curves obtained under five different training sample sizes.

As shown in Figure 4, the diagnostic performance of the model exhibits a clear step-wise improvement as the amount of training data increases. When only one training sample is available for each class, the model is unable to establish reliable classification boundaries, resulting in a relatively low ROC curve and an AUC value of 0.7347. When the number of training samples increases to five per class, the ROC curve shifts substantially toward the upper-left corner, and the AUC value increases significantly to 0.9801, benefiting from the effective multimodal feature fusion mechanism. When the number of training samples reaches 10 per class, the AUC value further increases to 0.9998, indicating excellent classification performance. Thereafter, under the settings of 15 and 20 training samples per class, the model maintains stable performance and approaches the ideal classification condition.

3. Confusion Matrix Analysis

To provide a more detailed class-wise evaluation and facilitate comparison among different methods, Figure 5 presents the confusion matrices of the proposed method together with three representative comparison methods under the 10-shot setting. The diagonal

elements denote correctly classified samples, whereas the off-diagonal elements indicate the corresponding misclassification probabilities. The confusion matrices provide an intuitive visualization of the classification performance for each fault category and reveal the detailed inter-class confusion patterns.

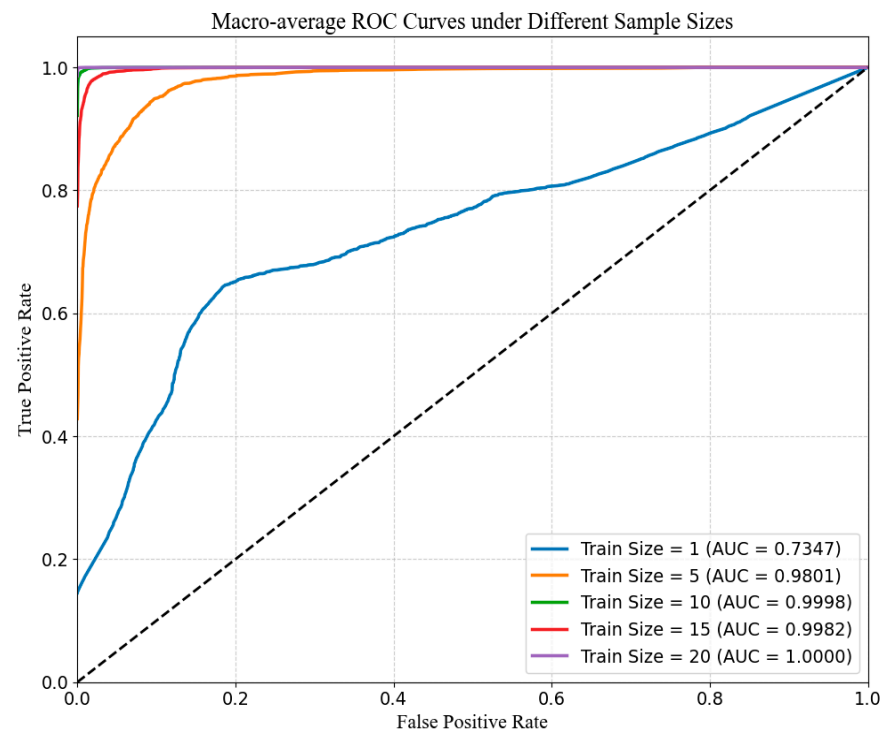


Figure 4. Macro-average ROC curves for multi-class classification under different training sample sizes (note: the dashed diagonal line represents the performance of a random classifier).

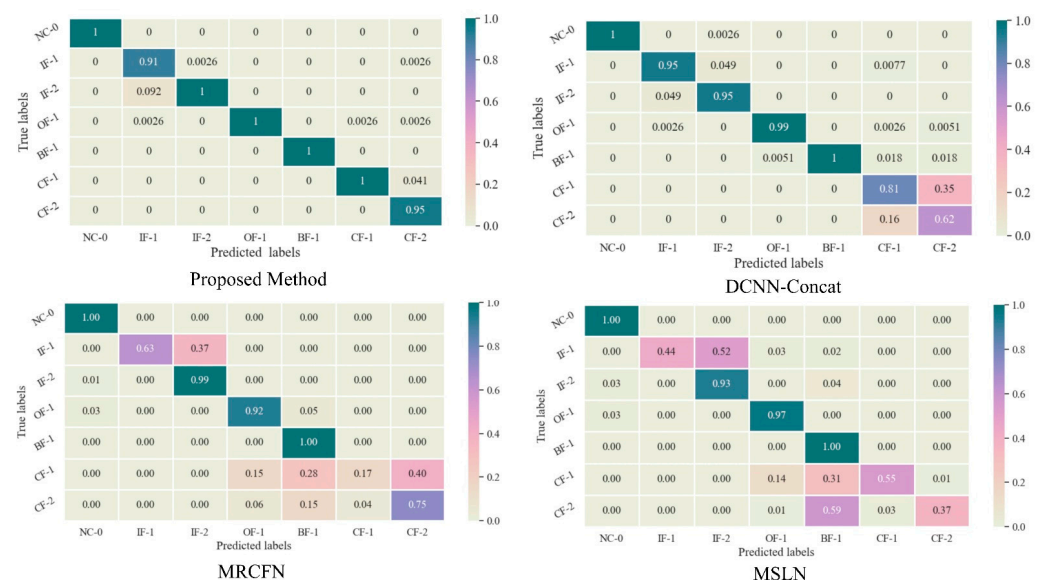


Figure 5. Confusion matrices of the proposed method and representative comparison methods under the 10-shot setting.

As shown in Figure 5, compared with the representative comparison methods, the proposed method exhibits a noticeably more concentrated diagonal distribution and substantially fewer off-diagonal misclassifications, indicating superior class discrimination under limited training samples. The proposed method achieves perfect recognition for

NC-0, OF-1, BF-1, and IF-2, while maintaining recognition accuracies above 90% for the remaining fault categories.

Although a small number of misclassification cases remain, they are primarily observed between fault categories with similar physical characteristics. Specifically, approximately 9% of the incipient inner-race fault samples are misclassified as advanced inner-race faults, while about 4% of the advanced cage fault samples are classified as incipient cage faults. Since these fault pairs occur at the same component and mainly differ in damage severity, their feature distributions remain highly similar under few-shot conditions, making them more difficult to distinguish.

In contrast, the comparison methods exhibit considerably more severe confusion between similar fault categories, particularly for the two cage fault conditions. For example, DCNN-Concat and MRCFN show substantial mutual confusion between CF-1 and CF-2, whereas MSLN frequently misclassifies CF-2 as BF-1. These results demonstrate that the proposed heterogeneous feature extraction and adaptive fusion strategy produce more discriminative fault representations under few-shot conditions.

4. t-SNE Visualization Analysis of Multimodal Features

To further investigate the feature representation capability of the proposed method, the high-dimensional features extracted from the testing samples were projected into a two-dimensional space using the t-SNE algorithm. The model was trained under the 10-shot setting, while the feature visualization was generated using the testing samples. Figure 6 illustrates the distributions of vibration features (circles), acoustic features (inverted triangles), and fused features (stars).

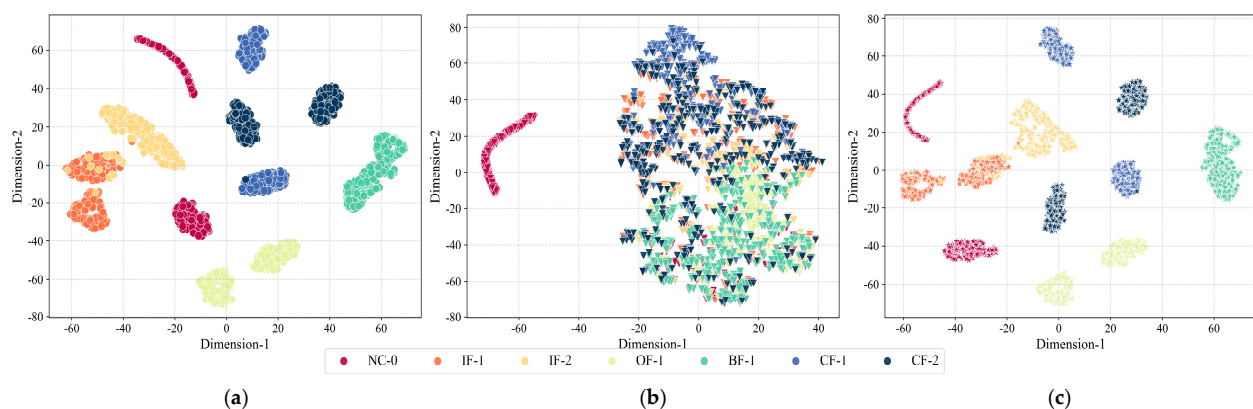


Figure 6. t-SNE visualization of feature distributions of testing samples for (a) vibration features (circles), (b) acoustic features (inverted triangles), and (c) fused multimodal features (stars).

As shown in Figure 6a, the vibration features (circles), serving as the dominant modality, exhibit good discriminative capability and generally form independent clusters. However, a certain degree of overlap still exists at the boundaries between some similar fault categories. This indicates that vibration signals contain richer fault-related information than acoustic signals and provide a more reliable basis for fault discrimination.

As shown in Figure 6b, the acoustic features (inverted triangles) exhibit a highly overlapping and scattered distribution in the feature space, which is consistent with the physical characteristic that acoustic signals are susceptible to environmental noise contamination. This indicates that acoustic signals alone are insufficient to achieve reliable fault classification.

Figure 6c illustrates the distribution of the final fused features (stars). Compared with the single-modality feature distributions shown in Figure 6a,b, the fused features exhibit noticeably more compact intra-class distributions and clearer inter-class separations. This improvement indicates that the proposed multimodal fusion strategy effectively exploits

the complementary fault information contained in vibration and acoustic signals while effectively integrating the complementary information provided by the acoustic modality. These results demonstrate that the proposed multimodal fusion strategy effectively enhances feature discrimination by improving intra-class compactness and inter-class separability.

3.5. Performance Comparison Between the Proposed Method and Comparison Methods

3.5.1. Comparison with Existing Methods

Based on the analysis in Section 3.4, ten training samples per class are sufficient to form relatively stable classification boundaries while maintaining strict few-shot constraints. Therefore, ten samples per class were selected as the unified few-shot benchmark condition. Under the same data partitioning, input length, number of training epochs, optimizer settings, and evaluation metrics, the proposed method was compared horizontally with the various baseline models listed in Section 3.2. Each method was independently repeated 10 times, and the results were averaged to reduce the influence of random factors.

Table 4 presents the overall diagnostic performance of different methods under the condition of ten training samples per class. It can be observed that the proposed method achieves the best results on the three core metrics, with Accuracy, Precision, and F1-score reaching 95.82%, 96.26%, and 95.69%, respectively. This indicates that the proposed heterogeneous feature extraction and adaptive modality fusion strategy can effectively construct a highly discriminative multimodal fault feature space even under extremely limited sample support.

Table 4. Performance comparison of different methods with 10 training samples per class.

Model	Accuracy (%)	Precision (%)	F1-Score (%)
DCNN-V	89.48 ± 3.52	89.83 ± 3.42	89.22 ± 3.65
DCNN-A	62.63 ± 5.58	64.03 ± 5.96	62.50 ± 5.64
DCNN-Concat	89.65 ± 2.55	90.40 ± 2.77	89.40 ± 2.62
MRCFN	83.60 ± 3.09	86.65 ± 2.70	82.69 ± 3.64
MSLN	77.22 ± 8.20	81.05 ± 7.70	75.96 ± 8.86
Proposed Method	95.82 ± 1.25	96.26 ± 1.10	95.69 ± 1.30

Compared with the single-modal vibration model DCNN-V, the Accuracy of the proposed method improves from 89.48% to 95.82%, and the F1-score increases from 89.22% to 95.69%. This demonstrates that, on the basis of the vibration modality already providing strong fault representation, the appropriate introduction of acoustic auxiliary information with effective fusion can further enhance the model's ability to recognize complex fault conditions. In contrast, the single-modal acoustic model DCNN-A can only partially reflect equipment anomalies, achieving an Accuracy of 62.63%, which is significantly lower than the other methods. This suggests that acoustic signals are more susceptible to environmental noise under few-shot conditions and are insufficient alone to support stable fault diagnosis.

Compared with the basic multimodal concatenation model DCNN-Concat, the proposed method improves Accuracy by 6.17% and F1-score by 6.29%. This indicates that simple feature concatenation, although able to utilize information from both modalities, cannot effectively model the dependencies between modalities and tends to directly introduce noise from the acoustic modality into the fusion space, thus limiting the final diagnostic performance. The heterogeneous extraction and adaptive fusion mechanisms proposed in this study, while preserving complementary information, effectively suppress negative transfer effects.

Furthermore, compared with the two advanced multimodal diagnostic methods MR-CFN and MSLN, the proposed method demonstrates a clear performance advantage under a unified experimental framework. MR-CFN achieves an average Accuracy of 83.60% and F1-score of 82.69%, while MSLN achieves an average Accuracy of 77.22% and F1-score of 75.96%, both lower than the proposed method (Accuracy 95.82%, F1-score 95.69%). This indicates that, compared with these two state-of-the-art diagnostic models, the proposed method is more effective in fusing and extracting features from vibration and acoustic data.

3.5.2. Robustness Evaluation Under Sensor Noise

In practical industrial applications, vibration and acoustic sensors are inevitably affected by environmental interference, resulting in the degradation of signal quality. To evaluate the robustness of the proposed multimodal diagnosis framework under such conditions, Gaussian sensor noise was independently introduced into either the vibration branch or the acoustic branch during the testing stage, while the other sensing modality remained clean. This experimental setting simulates the practical scenario in which only one sensor is contaminated while the remaining sensing channel continues to provide reliable information. Three signal-to-noise ratio (SNR) levels (20 dB, 10 dB, and 0 dB) were considered to represent different degrees of sensor contamination. All experiments were independently repeated ten times under identical experimental settings, and all reported numerical results correspond to the average performance over these repeated runs.

As reported in Table 5, diagnostic performance declines gradually with decreasing signal-to-noise ratio, regardless of whether noise is injected into the acoustic or vibration branch. Severe sensor interference inevitably deteriorates the quality of the acquired signals, resulting in the distortion of fault-related characteristics. Consequently, the discriminative capability of the extracted features is weakened, leading to a gradual decrease in diagnostic performance. Nevertheless, the proposed method maintains high diagnostic performance across all tested noise levels, demonstrating strong robustness against sensor contamination.

Table 5. Robustness evaluation of the proposed method under different sensor noise conditions.

Noise Type	SNR (dB)	Accuracy (%)	Precision (%)	F1-Score (%)
Clean	/	95.82 ± 1.25	96.26 ± 1.10	95.72 ± 1.26
Acoustic	20	94.72 ± 1.35	95.13 ± 1.28	94.34 ± 1.52
Acoustic	10	93.22 ± 1.94	93.75 ± 2.01	93.11 ± 2.03
Acoustic	0	91.03 ± 2.67	90.87 ± 2.55	90.98 ± 2.58
Vibration	20	94.41 ± 1.68	94.66 ± 1.66	94.13 ± 1.73
Vibration	10	93.17 ± 2.23	94.37 ± 2.76	93.33 ± 2.72
Vibration	0	90.88 ± 2.51	90.79 ± 2.43	91.02 ± 2.81

To further investigate the adaptive behavior of the proposed fusion mechanism, the average modal weights generated by the adaptive cross-attention module were analyzed under different acoustic noise levels. To isolate the effect of acoustic interference, Gaussian noise was added only to the acoustic modality, while the vibration modality remained unchanged. As illustrated in Figure 7, the weight assigned to the acoustic modality gradually decreases as the acoustic noise level increases, whereas the vibration modality receives a correspondingly higher weight. This trend indicates that the proposed fusion mechanism automatically relies more on the modality with better signal quality when one modality becomes unreliable. Consequently, the influence of the corrupted acoustic signal on the final diagnostic decision is reduced, which explains the stable performance observed under different noise conditions in Table 5.

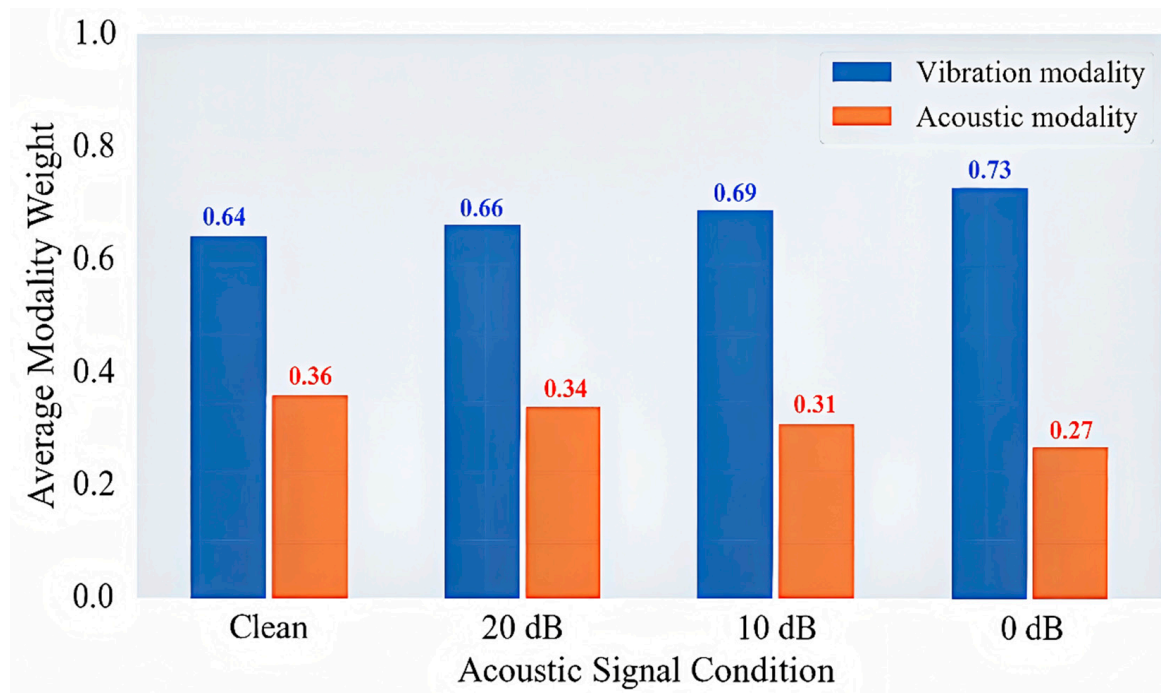


Figure 7. Average adaptive modality weights under different acoustic noise levels.

Vibration and acoustic signals provide different yet complementary fault information, so the degradation of one sensing modality does not necessarily lead to the loss of all discriminative features. During feature fusion, the adaptive cross-attention mechanism automatically assigns different weights to the two modalities according to the extracted feature responses. Consequently, the contribution of the corrupted modality is automatically reduced, while the cleaner modality contributes more to the final diagnostic decision.

A comparison between the two noise conditions further shows that vibration sensor corruption has a greater influence on the diagnostic results than acoustic sensor corruption. This difference is consistent with the characteristics of rolling bearing diagnosis, where vibration signals usually carry more direct fault impulses and therefore contribute more significantly to the final decision. Even so, the decrease in diagnostic performance remains moderate, indicating that the acoustic modality still provides effective supplementary information when the vibration signal quality deteriorates.

Overall, the results show that the proposed framework remains stable under different levels of sensor noise. The robustness evaluation confirms that the multimodal heterogeneous fusion strategy can effectively mitigate the influence of sensor interference and is therefore suitable for practical bearing fault diagnosis under realistic operating conditions.

3.6. Analysis of the Training Process

Although the aforementioned evaluation metrics demonstrate the final diagnostic performance of the model, they are insufficient to reveal the internal weight adjustment process of the network. Therefore, under the few-shot condition of ten training samples per class, this section investigates the evolution of the evaluation metrics and internal weights throughout 100 training epochs, and compares the results with those of the baseline feature concatenation method.

1. Evolution of Training Loss and Testing Accuracy

Figure 8 illustrates the dual-axis curves of training loss and testing accuracy throughout the entire training process of the proposed method.

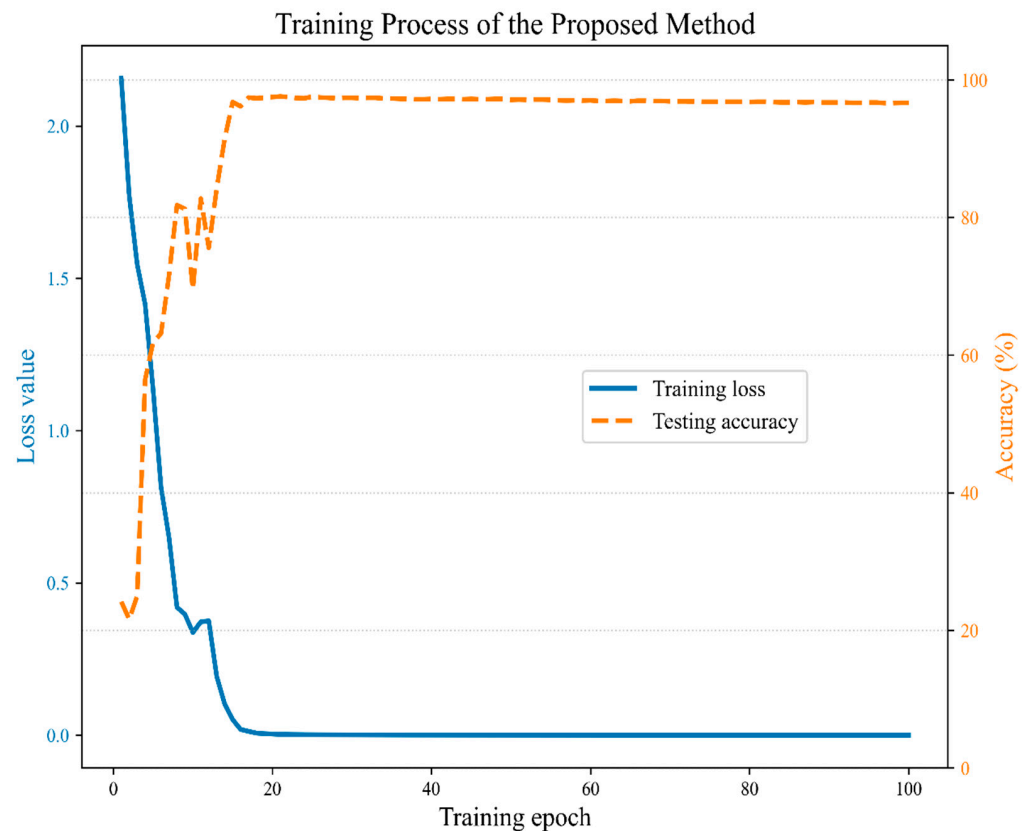


Figure 8. Evolution of training loss and testing accuracy for the proposed method.

As shown in Figure 8, during the initial stage of training, the model's loss rapidly decreases, while the corresponding testing accuracy exhibits several pronounced oscillations. This phenomenon is reasonable because the proposed method incorporates an adaptive fusion mechanism, and in the early stages, the network not only needs to learn the classification rules but also continuously explores and adjusts the fusion weights between the two modalities. After approximately 20 training epochs, the testing accuracy enters a high-level region around 96% and remains relatively stable. This indicates that the model converges quickly under the few-shot condition and maintains highly stable performance in the later stages.

2. Comparative Analysis of the Three Evaluation Metrics

To demonstrate the superiority of the fusion mechanism proposed in this paper from multiple perspectives, Figure 9 compares the training processes of our method and the DCNN-Concat baseline across three metrics: accuracy, precision, and F1-score.

Observing these three subplots, a common pattern can be identified: the baseline method DCNN-Concat, which uses simple direct concatenation without additional optimization of feature weights, shows a relatively fast increase in the early training stages. However, as training progresses, its Accuracy and F1-score quickly plateau around 90%, indicating that its feature space is easily affected by noise.

In contrast, the proposed method, although experiencing brief fluctuations in the early stage, exhibits a clear improvement in performance during later epochs. After approximately 10 training epochs, the red curve quickly surpasses the blue baseline and maintains a significant advantage across all three metrics. This observation directly demonstrates that the adaptive feature weighting mechanism designed in this work effectively filters out irrelevant information, thereby endowing the model with a higher performance ceiling and stronger resistance to interference.

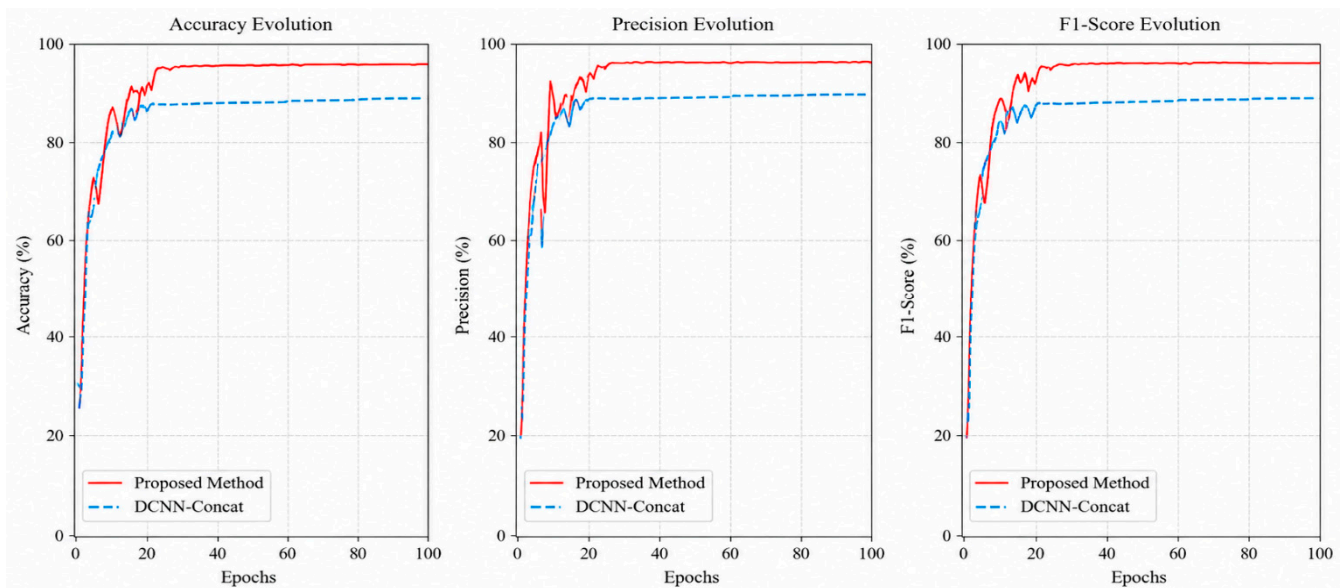


Figure 9. Evolution of Accuracy, Precision, and F1-score during Training.

3. Analysis of Adaptive Modality Fusion Weights

To investigate which modality—acoustic or vibration—is more important in joint diagnosis, this study extracts the numerical changes in the adaptive weight layer in the model during the training process, as shown in Figure 10.

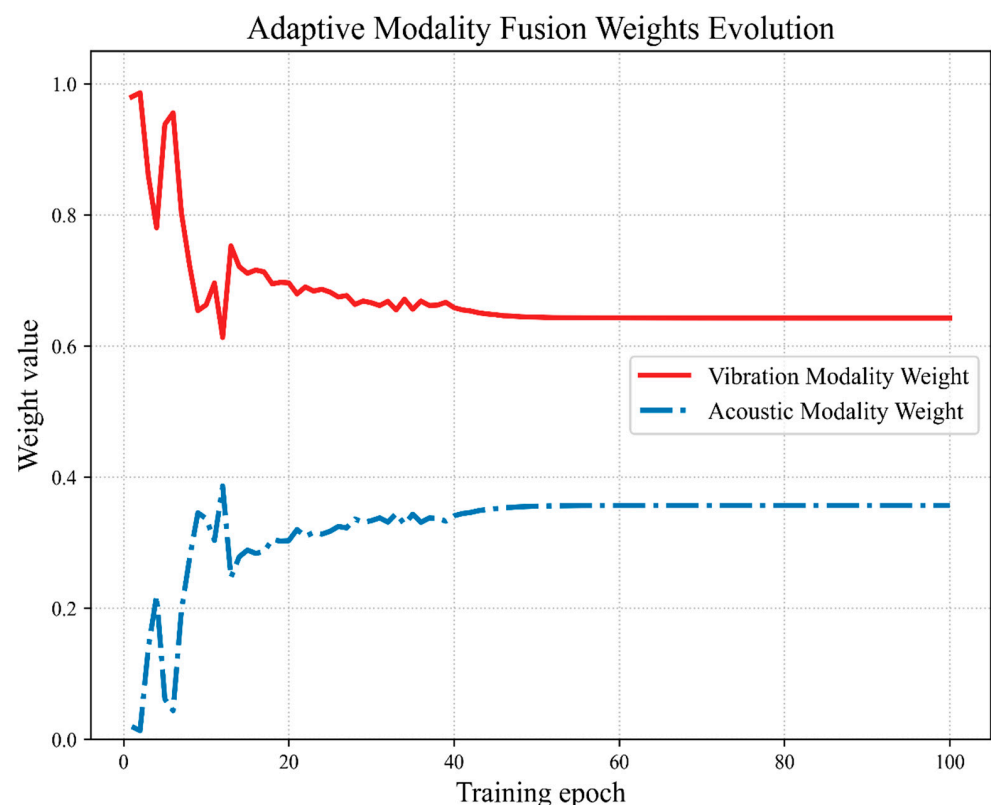


Figure 10. Evolution of adaptive modal weights over training iterations.

As shown in Figure 10, the model gradually forms a relatively stable allocation of modality contributions during the training process. The training trajectory exhibits a clear optimization trend. In the early training stage, the weight of the vibration modality increases rapidly, while the weight of the acoustic modality correspondingly decreases.

Subsequently, both curves maintain a highly stable trend throughout the entire training period. Eventually, the adaptive weights of two modalities gradually converge to a stable allocation. The weight of the vibration modality stabilizes at approximately 0.64, while the acoustic modality weight remains around 0.36.

This clearly delineated distribution of weights indicates an objective pattern in mechanical fault diagnosis: the acoustic modality, with a weight as high as 0.36, is not entirely ignored by the network but serves as important supplementary information in the joint decision-making process. This evolution process directly demonstrates that the proposed adaptive fusion mechanism can dynamically assign weights to the strong vibration signals and weak acoustic signals according to their physical quality in the data, thereby improving the effectiveness of multimodal information fusion.

3.7. Analysis of Computational Efficiency

In practical industrial online monitoring scenarios, fault diagnosis models are required not only to achieve high diagnostic accuracy but also to complete training and inference efficiently under limited computational resources. Therefore, in addition to classification performance, training time and inference time per sample are also important indicators for evaluating the engineering applicability of a diagnostic model. Accordingly, this section comprehensively compares the proposed method and the benchmark methods in terms of average training time, inference time per sample, and diagnostic accuracy. The detailed comparison results are presented in Table 6.

Table 6. Comparison of training time, inference time, and diagnostic accuracy of different methods (10 samples per class).

Model	Training Time (s)	Inference Time (ms/Sample)	Accuracy (%)
DCNN-V	28.58	1.5658	89.48
DCNN-A	18.99	1.1387	62.63
DCNN-Concat	27.23	1.4190	89.65
MRCFN	124.15	5.6438	83.60
MSLN	116.08	2.9473	77.22
Proposed Method	30.87	1.5231	95.82

From the perspective of training time, the proposed method requires an average of 30.87 s to complete a single full training session. This value is slightly higher than that of DCNN-V, DCNN-A, and DCNN-Concat, but is clearly lower than MSLN and MRCFN. This indicates that, although the proposed method incorporates a cross-attention mechanism and dynamic modality weight allocation, the overall network structure still maintains good computational efficiency, capable of meeting the demands for rapid iteration and retraining in engineering scenarios.

Regarding online inference efficiency, the proposed method processes an individual fault sample with an average inference time of 1.5231 ms. Although this value is slightly higher than that of DCNN-V, DCNN-A, and DCNN-Concat, it is significantly lower than MRCFN and MSLN. This demonstrates that the adaptive cross-attention fusion mechanism introduced in this work effectively improves diagnostic performance without introducing unacceptable extra latency, still ensuring a good real-time response. For high-frequency online monitoring tasks, this millisecond-level inference speed is sufficient to meet practical application requirements.

Overall, the proposed method achieves a good balance among diagnostic accuracy, training efficiency, and inference speed. Compared with complex multimodal models, the proposed approach achieves higher diagnostic performance while maintaining relatively

low computational costs; compared with baseline methods, although training and inference times slightly increase, this is traded for significantly improved accuracy. Therefore, the proposed method is more suitable for deployment in industrial online monitoring scenarios with limited computational resources but high real-time performance requirements.

3.8. Analysis of Ablation Experiments

To thoroughly validate the necessity and effectiveness of the key design elements in the proposed heterogeneous CNN model with adaptive modality fusion, this section conducts seven groups of ablation experiments. To ensure absolute fairness in comparison, all variant models are independently executed 10 times under the strict few-shot condition of 10 samples per class to obtain averaged metrics.

Experiments 1 and 2 construct unimodal networks with only vibration signals and only acoustic signals as inputs, respectively, serving as performance baselines for single-source signals. Experiment 3 removes all intelligent fusion mechanisms proposed in this work, retaining only the 10-layer vibration branch and the 9-layer acoustic branch, and directly concatenates the extracted bimodal features. Experiments 4 and 5 build upon the basic concatenation, individually introducing the cross-attention mechanism and independent feature weighting module to verify the individual contributions of each modality module. Experiments 6 and 7 investigate the effect of network asymmetry by constructing a deep 10 + 10 layer symmetric dual-branch network and a shallow 9 + 9 layer symmetric dual-branch network, respectively, and performing their fusion.

The proposed method integrates heterogeneous feature extraction, bidirectional cross-attention, and adaptive dynamic weighting. The comparative results of the three core evaluation metrics (Accuracy, Precision, F1-score) of each ablation model on the test set are summarized in Table 7.

Table 7. Comprehensive diagnostic performance of each ablation model under 10 samples per class.

Modal/Method	Accuracy (%)	Precision (%)	F1-Score (%)	Inference Time (ms)
Vibration-only	89.48 ± 3.52	89.83 ± 3.42	89.22 ± 3.65	1.5658
Acoustic-only	62.63 ± 5.58	64.03 ± 5.96	62.50 ± 5.64	1.1387
Heterogeneous Basic Concatenation	89.65 ± 2.55	90.40 ± 2.77	89.40 ± 2.62	1.4190
Heterogeneous Concatenation + Cross-Attention	91.87 ± 2.73	92.35 ± 2.60	91.74 ± 2.80	1.8149
Heterogeneous Concatenation + Independent Feature Weighting	90.38 ± 2.92	91.25 ± 2.80	90.11 ± 3.01	1.3824
Homogeneous Deep Concatenation	90.96 ± 2.78	91.66 ± 2.57	90.79 ± 2.80	1.5054
Homogeneous Shallow Concatenation	88.79 ± 3.93	89.30 ± 4.28	88.40 ± 4.14	1.4196
Adaptive Modality Fusion Heterogeneous	95.82 ± 1.25	96.26 ± 1.10	95.69 ± 1.30	1.5231

As shown in Table 7, the proposed ablation study progressively introduces the heterogeneous feature extraction branches, cross-attention module, and dynamic weighting strategy to evaluate the contribution of each component. Compared with the baseline model, introducing the heterogeneous branches significantly improves the diagnostic performance, demonstrating that extracting complementary features from vibration and acoustic signals effectively enhances the representation capability of the network. Based on this architecture, incorporating the cross-attention module further improves the diagnostic accuracy, indicating that cross-modal interaction enables the model to capture more discriminative features by strengthening the information exchange between the two modalities. Furthermore, after introducing the dynamic weighting strategy, the diagnostic performance

is further enhanced, suggesting that adaptively assigning modality weights allows the network to emphasize the more informative modality under different fault conditions and thus improves feature fusion. These results demonstrate that each proposed component contributes positively to the overall performance, and their combination achieves the best diagnostic results.

To more intuitively demonstrate the stepwise improvement in overall classification accuracy resulting from different network architecture designs, Figure 11 presents bar plots of the average accuracy for each ablation variant, with error bars showing the standard deviation over 10 runs.

Ablation Study: Performance Comparison of Different Architectures

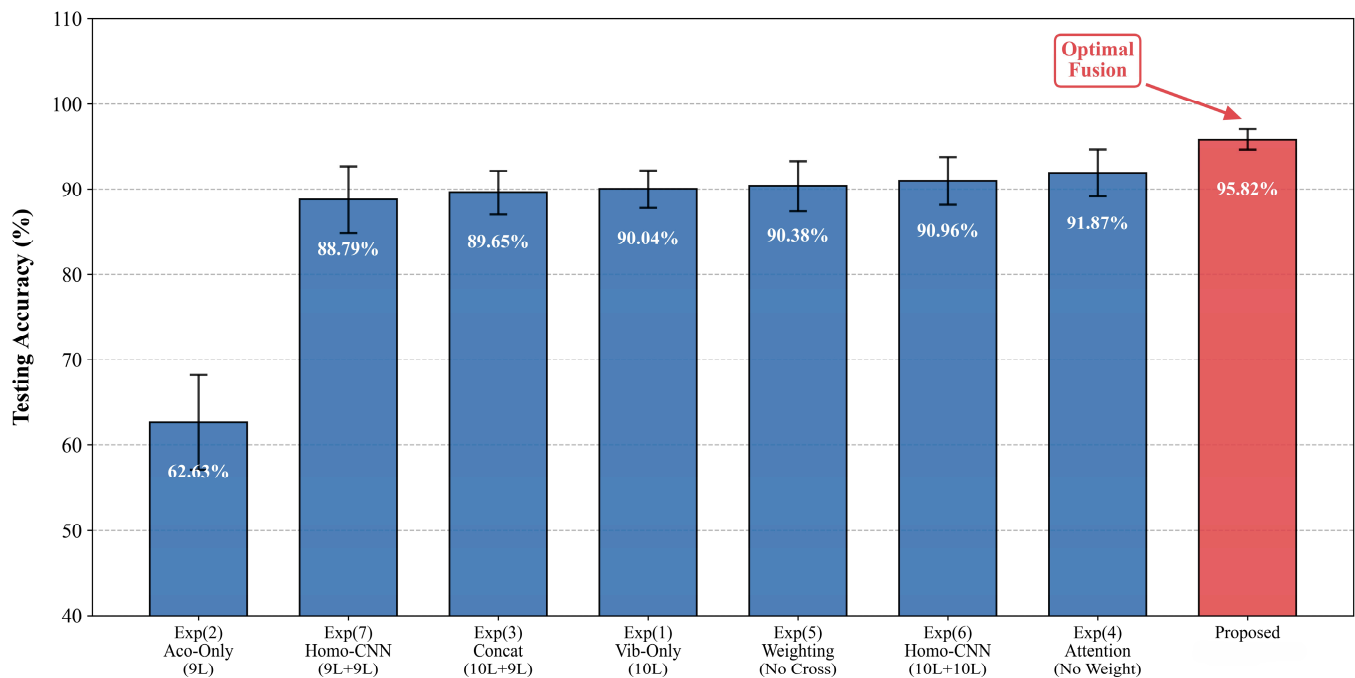


Figure 11. Comprehensive diagnostic performance of different ablation networks under 10 samples per class (note: bars are sorted in ascending order of testing accuracy).

It can be observed that, in the few-shot acoustic–vibration fusion diagnostic task constructed in this study, the proposed method, based on cross-modal perceptual adaptive fusion, can more effectively leverage complementary information from the two modalities, significantly enhancing both the classification performance and stability of the model.

4. Discussion

The proposed multimodal heterogeneous convolutional neural network demonstrates superior diagnostic performance under few-shot conditions and complex noisy environments.

First, the model exhibits strong feature extraction capability in few-shot scenarios. Under the stringent condition of only 10 training samples per fault category, the proposed method still achieved a fault diagnosis accuracy of 95.82%. From a signal analysis perspective, this performance can be primarily attributed to the heterogeneous network architecture. By reducing excessive deep pooling operations in the acoustic branch, the network effectively preserves high-frequency fault-impact information that would otherwise be lost during feature compression. Consequently, the model is able to learn discriminative fault representations and achieve rapid convergence even with extremely limited training data.

Second, the introduced adaptive cross-attention mechanism significantly enhances the noise robustness of the model, which carries significant practical implications. As indicated by the modality-weight evolution results, the network automatically adjusts modality contributions during training. In the later training stage, the vibration modality weight stabilizes at approximately 0.64, whereas the acoustic modality weight converges to approximately 0.36. This distribution is consistent with practical industrial observations. Owing to its higher signal-to-noise ratio, vibration data serve as the primary source of diagnostic information, whereas acoustic data are more susceptible to environmental noise and mainly provide complementary evidence. Through this dynamic weighting strategy, the proposed model reduces the adverse impact of noisy modalities and alleviates the performance degradation commonly observed in conventional feature concatenation methods.

Finally, the proposed method achieves high computational efficiency suitable for real-time online monitoring. The average inference time is only 1.52 ms per sample, which not only enables efficient deployment but also makes the model highly suitable for industrial edge-computing devices with limited computational resources.

In real industrial environments, both acoustic and vibration measurements are susceptible to environmental interference. Background acoustic noise can obscure faint fault-induced acoustic emissions, making incipient faults more difficult to detect and increasing the probability of false-negative predictions. Such interference degrades signal quality, impairs the discriminative capability of the extracted features, reduces the sensitivity of the proposed framework to weak fault signatures, and consequently decreases the overall diagnostic accuracy. As demonstrated by the robustness evaluation in Section 3.5.2, the diagnostic performance gradually decreases with decreasing signal-to-noise ratio. Nevertheless, the proposed multimodal framework alleviates this performance degradation by exploiting the complementary information provided by vibration and acoustic signals. When one sensing modality is contaminated, the adaptive fusion strategy dynamically assigns greater importance to the relatively more reliable modality. Consequently, the proposed framework remains robust under different sensor noise conditions, demonstrating its robustness and practical applicability in real industrial fault diagnosis.

Although the proposed method demonstrates excellent overall diagnostic performance, several limitations remain regarding fine-grained fault classification and validation scenarios. As revealed by the confusion matrices, although the proposed method substantially reduces inter-class confusion compared with the representative methods, a small number of misclassifications still occur between fault categories with similar damage characteristics, especially those sharing the same fault location but different damage severities. This observation indicates that, under few-shot conditions, the current heterogeneous feature extraction and adaptive fusion strategy still has limited sensitivity to subtle differences in fault severity. Further improving the discrimination of similar fault categories remains an important direction for future research.

In addition, the experimental validation in this study is mainly conducted using the bearing dataset collected by the University of Ottawa under constant rotational speed and constant load conditions. Although these controlled operating conditions facilitate objective performance evaluation and fair comparison among different diagnostic methods, they cannot fully represent practical industrial environments, where rotating machinery frequently operates under variable-speed and variable-load conditions. Such operating-condition variations may result in nonstationary vibration and acoustic signals, leading to frequency shifts and feature distribution discrepancies between training and testing data, which may reduce the generalization capability of the proposed model.

To address these limitations, future research will focus on two directions. On the one hand, multi-scale convolutional kernels will be incorporated into the current hetero-

geneous feature-extraction network to improve the discrimination of faults with similar damage sizes. On the other hand, operating-condition-invariant feature learning will be integrated into the proposed framework, together with domain adaptation techniques, to reduce feature distribution discrepancies caused by variations in rotational speed and load. Furthermore, multimodal datasets collected under diverse operating conditions will be introduced to further validate and improve the robustness and generalization capability of the proposed framework in real industrial environments.

5. Conclusions

This paper proposes a multimodal intelligent fault diagnosis method for rolling bearings based on vibration and acoustic signals. By employing a heterogeneous convolutional neural network, effective feature extraction and optimal fusion of multimodal data are achieved. The experimental results indicate that constructing heterogeneous one-dimensional convolutional feature extraction networks for vibration and acoustic signals is more consistent with the physical characteristics of multimodal signals than conventional homogeneous architectures. The proposed structure effectively alleviates the loss of high-frequency acoustic details caused by deep pooling operations and significantly enhances feature extraction capability.

Furthermore, an adaptive cross-attention mechanism is introduced to dynamically allocate modality weights according to data quality. While incorporating complementary acoustic information, this mechanism effectively suppresses the influence of environmental noise. Experimental results show that the proposed model achieves a diagnostic accuracy of 95.82% under a strict few-shot setting with only 10 training samples per class.

Overall, the proposed method demonstrates strong fault feature extraction capability and effective multimodal fusion performance under few-shot conditions. In addition, it achieves a favorable balance between diagnostic accuracy and computational efficiency, indicating promising potential for practical engineering applications. Future research will focus on improving the model's sensitivity to fine-grained fault discrimination and further validating its performance under variable operating conditions.

Author Contributions: Methodology, writing—original draft preparation, C.S.; investigation, writing—original draft preparation, C.W.; writing—original draft preparation, S.H.; writing—review and editing, supervision, funding acquisition, T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No.52505141) and the Natural Science Foundation of Hunan Province (Grant No.2025JJ60284).

Data Availability Statement: The rolling bearing fault dataset released by the University of Ottawa is sourced from publicly available datasets, and the methodology for accessing and obtaining these data resources has been thoroughly described in the manuscript. For more detailed data descriptions, the corresponding author of this paper can be contacted. Depending on the type of request, some of the important data will be provided appropriately.

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

References

1. Zhang, K.; Wang, J.; Shi, H.; Zhang, X.; Tang, Y. A fault diagnosis method based on improved convolutional neural network for bearings under variable working conditions. *Measurement* **2021**, *182*, 109749. [[CrossRef](#)]
2. Schwendemann, S.; Amjad, Z.; Sikora, A. A survey of machine-learning techniques for condition monitoring and predictive maintenance of bearings in grinding machines. *Comput. Ind.* **2021**, *125*, 103380. [[CrossRef](#)]

3. Hakim, M.; Omran, A.A.B.; Ahmed, A.N.; Al-Waily, M.; Abdellatif, A. A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations. *Ain Shams Eng. J.* **2023**, *14*, 101945. [\[CrossRef\]](#)
4. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 6999–7019. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Li, P.; Pei, Y.; Li, J. A comprehensive survey on design and application of autoencoder in deep learning. *Appl. Soft Comput.* **2023**, *138*, 110176. [\[CrossRef\]](#)
6. Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 3313–3332. [\[CrossRef\]](#)
7. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [\[CrossRef\]](#)
8. Liu, J.; Zhang, C.; Jiang, X. Imbalanced fault diagnosis of rolling bearing using improved MsR-GAN and feature enhancement-driven CapsNet. *Mech. Syst. Signal Process.* **2022**, *168*, 108664. [\[CrossRef\]](#)
9. Ding, Y.; Jia, M.; Miao, Q.; Cao, Y. A novel time–frequency Transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. *Mech. Syst. Signal Process.* **2022**, *168*, 108616. [\[CrossRef\]](#)
10. 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 Trans.* **2022**, *128*, 470–484. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Han, T.; Zhang, L.; Yin, Z.; Tan, A.C.C. Rolling bearing fault diagnosis with combined convolutional neural networks and support vector machine. *Measurement* **2021**, *177*, 109022. [\[CrossRef\]](#)
12. Li, C.; Zhou, J.; Wu, X.; Liu, T. Phase-based video vibration measurement and fault feature extraction method for compound faults of rolling bearings. *Adv. Eng. Inform.* **2024**, *62*, 102897. [\[CrossRef\]](#)
13. Zhang, X.; He, C.; Lu, Y.; Chen, B.; Zhu, L.; Zhang, L. Fault diagnosis for small samples based on attention mechanism. *Measurement* **2022**, *187*, 110242. [\[CrossRef\]](#)
14. Wang, X.; Mao, D.; Li, X. Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. *Measurement* **2021**, *173*, 108518. [\[CrossRef\]](#)
15. You, K.; Wang, P.; Huang, P.; Gu, Y. A sound-vibration physical-information fusion constraint-guided deep learning method for rolling bearing fault diagnosis. *Reliab. Eng. Syst. Saf.* **2025**, *253*, 110556. [\[CrossRef\]](#)
16. Shi, H.; Li, Y.; Bai, X.; Zhang, K.; Sun, X. A two-stage sound-vibration signal fusion method for weak fault detection in rolling bearing systems. *Mech. Syst. Signal Process.* **2022**, *172*, 109012. [\[CrossRef\]](#)
17. Zhang, A.; Li, S.; Cui, Y.; Yang, W.; Dong, R.; Hu, J. Limited data rolling bearing fault diagnosis with few-shot learning. *IEEE Access* **2019**, *7*, 110895–110904. [\[CrossRef\]](#)
18. Pan, T.; Chen, J.; Zhang, T.; Liu, S.; He, S.; Lv, H. Generative adversarial network in mechanical fault diagnosis under small sample: A systematic review on applications and future perspectives. *ISA Trans.* **2022**, *128*, 1–10. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Zhang, T.; Chen, J.; He, S.; Zhou, Z. Prior Knowledge-Augmented Self-Supervised Feature Learning for Few-Shot Intelligent Fault Diagnosis of Machines. *IEEE Trans. Ind. Electron.* **2022**, *69*, 10573–10584. [\[CrossRef\]](#)
20. Cui, L.; Tian, X.; Wei, Q.; Liu, Y. A self-attention based contrastive learning method for bearing fault diagnosis. *Expert Syst. Appl.* **2024**, *238*, 121645. [\[CrossRef\]](#)
21. Zhang, T.; Chen, J.; Ye, Z.; Liu, W.; Tang, J. Prior Knowledge-Informed Multi-Task Dynamic Learning for Few-Shot Machinery Fault Diagnosis. *Expert Syst. Appl.* **2025**, *271*, 126439. [\[CrossRef\]](#)
22. Li, C.; Li, S.; Zhang, A.; He, Q.; Liao, Z.; Hu, J. Meta-learning for few-shot bearing fault diagnosis under complex working conditions. *Neurocomputing* **2021**, *439*, 197–211. [\[CrossRef\]](#)
23. Han, D.; Zhang, Y.; Yu, Y.; Tian, J.; Shi, P. Multi-source heterogeneous information fusion fault diagnosis method based on deep neural networks under limited datasets. *Appl. Soft Comput.* **2024**, *154*, 111371. [\[CrossRef\]](#)
24. Chen, J.; Lin, C.; Yao, B.; Yang, L.; Ge, H. Intelligent fault diagnosis of rolling bearings with low-quality data: A feature significance and diversity learning method. *Reliab. Eng. Syst. Saf.* **2023**, *237*, 109343. [\[CrossRef\]](#)
25. Ye, M.; Yan, X.; Jiang, D.; Xiang, L.; Chen, N. MIFDELN: A multi-sensor information fusion deep ensemble learning network for diagnosing bearing faults in noisy scenarios. *Knowl.-Based Syst.* **2024**, *284*, 111294. [\[CrossRef\]](#)
26. Shuai, H.; Li, J.; Wu, L.; Zhang, W. Research on acoustic fault diagnosis of bearings based on spatial filtering and time-frequency domain filtering. *Measurement* **2023**, *221*, 113533. [\[CrossRef\]](#)
27. Pacheco-Chérrez, J.; Fortoul-Díaz, J.A.; Cortés-Santacruz, F.; Alosa-Valerdi, L.M.; Ibarra-Zarate, D.I. Bearing fault detection with vibration and acoustic signals: Comparison among different machine learning classification methods. *Eng. Fail. Anal.* **2022**, *139*, 106515. [\[CrossRef\]](#)
28. Li, X.; Xiao, S.; Li, Q.; Zhu, L.; Wang, T.; Chu, F. The bearing multi-sensor fault diagnosis method based on a multi-branch parallel perception network and feature fusion strategy. *Reliab. Eng. Syst. Saf.* **2025**, *261*, 111122. [\[CrossRef\]](#)
29. Lian, Y.; Wang, J.; Li, Z.; Liu, W.; Huang, L.; Jiang, X. Residual attention guided vision transformer with acoustic-vibration signal feature fusion for cross-domain fault diagnosis. *Adv. Eng. Inform.* **2025**, *64*, 103003. [\[CrossRef\]](#)

30. Xu, Y.; Feng, K.; Yan, X.; Sheng, X.; Sun, B.; Liu, Z.; Yan, R. Cross-Modal Fusion Convolutional Neural Networks With Online Soft-Label Training Strategy for Mechanical Fault Diagnosis. *IEEE Trans. Ind. Inform.* **2024**, *20*, 73–84. [[CrossRef](#)]
31. Dubey, S.R.; Singh, S.K.; Chaudhuri, B.B. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* **2022**, *503*, 92–108. [[CrossRef](#)]
32. Wang, D.; Li, Y.; Jia, L.; Song, Y.; Wen, T. Attention-Based Bilinear Feature Fusion Method for Bearing Fault Diagnosis. *IEEE/ASME Trans. Mechatron.* **2023**, *28*, 1695–1705. [[CrossRef](#)]
33. Zhang, W.; Xu, Q.; Hu, Y.; Xu, C.; Luo, L. Attention-Based Two-Stage Multi-Sensor Feature Fusion Method for Bearing Fault Diagnosis. *IEEE Trans. Ind. Appl.* **2024**, *60*, 8709–8721. [[CrossRef](#)]
34. Zeng, J.; Wang, D.; Chen, P. A Survey on Transformers for Point Cloud Processing: An Updated Overview. *IEEE Access* **2022**, *10*, 86510–86527. [[CrossRef](#)]
35. Song, Q.; Sun, S.; Wang, B.; Song, Q.; Wang, T.; Jiang, H. Noise-robust fault diagnosis network based on multiscale feature enhancement and dynamic cross-modal interaction. *Measurement* **2026**, *257*, 118564. [[CrossRef](#)]
36. Sehri, M.; Dumond, P.; Bouchard, M. University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets. *Data Brief* **2023**, *49*, 109327. [[CrossRef](#)] [[PubMed](#)]
37. Ye, M.; Yan, X.; Hua, X.; Jiang, D.; Xiang, L.; Chen, N. MRCFN: A multi-sensor residual convolutional fusion network for intelligent fault diagnosis of bearings in noisy and small sample scenarios. *Expert Syst. Appl.* **2025**, *259*, 125214. [[CrossRef](#)]
38. Chen, X.; Xue, Y.; Huang, M.; Yang, R. Multi-Modal Self-Supervised Learning for Cross-Domain One-Shot Bearing Fault Diagnosis. *IFAC-Pap.* **2024**, *58*, 746–751. [[CrossRef](#)]
39. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.