

Article

Multi-Modal Bearing Fault Diagnosis Based on Multiscale Feature Enhancement and Adaptive Fusion

Wenyan Zhu ¹, Zhiyu Du ¹, Tongtong Liu ^{1,*} , Jinfeng Huang ², Feng Jin ³ and Chao Zhang ⁴

¹ School of Mechanical Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China

² State Key Laboratory of Tribology in Advanced Equipment, Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China

³ Energy Equipment Intelligent Inspection and Safety Assurance Laboratory, Inner Mongolia Institute of Special Equipment Inspection and Research, Baotou 010020, China

⁴ School of Digital and Intelligent Industry, Inner Mongolia University of Science and Technology, Baotou 014010, China

* Correspondence: liuzhaishi@imust.edu.cn

Abstract

Bearings are critical components of modern large-scale rotating machinery, such as wind turbines, and their operating condition directly affects system safety, reliability, and operation and maintenance costs. Vibration and current signals reflect the mechanical responses and electromechanically coupled responses induced by bearing faults, respectively, and the two modalities differ in their characteristic representations of fault information. When vibration and current features are fused only once at the end of a network, early complementary information cannot be incorporated into subsequent feature learning; moreover, a single-level weighting scheme cannot capture variations in modality contributions across different samples and feature depths. To address these limitations, a Multi-Scale Feature Enhancement and Adaptive Fusion Network (MSFEAF-Net) is developed. The network employs a dual-branch multi-scale residual architecture to extract vibration and current features, enables early cross-modal interaction through bidirectional cross-attention after the first residual stage, and enhances the vibration and current modalities using one-dimensional spatial attention and channel attention, respectively, according to their distinct feature representations. Subsequently, sample-dependent modality weights are generated at three feature levels, and a feature pyramid is used to aggregate feature information across different levels. Finally, parallel multi-scale convolutions and channel-spatial attention are employed to further refine the features for bearing condition identification. Experimental results show that MSFEAF-Net achieves a test accuracy of 99.88% on the PU dataset under conventional operating conditions. Under strong noise at -8 dB, it maintains accuracies of 91.78% and 91.32% on the PU and KAIST datasets, respectively. In the cross-condition task involving simultaneous variations in rotational speed, torque, and radial load, the model achieves an accuracy of 92.45%. These results validate the effectiveness of the proposed method for fault diagnosis under strong noise and complex operating conditions.

Keywords: rolling bearings; fault diagnosis; deep learning; multimodal fusion; attention mechanism



Academic Editor: Hui Ma

Received: 31 August 2026

Revised: 24 September 2026

Accepted: 29 September 2026

Published: 2 October 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

As global installed wind power capacity continues to expand and wind turbines evolve toward higher power ratings and greater structural complexity, their operational

safety has attracted increasing attention [1]. Rolling bearings are critical supporting components in rotating machinery and primarily serve to transmit loads and maintain rotational accuracy [2]. When localized damage occurs in a bearing, the resulting fault-induced vibrations may give rise to dynamic eccentricity and alter the generator stator current response through electromechanical coupling effects, such as unbalanced magnetic pull [3]. If an incipient fault is not detected promptly, the damage may progressively worsen and potentially affect the normal operation of adjacent components, including the gearbox and generator [4]. During actual operation, rotational speed, load, and environmental conditions frequently vary, while signal propagation paths may attenuate incipient fault features. Therefore, effectively exploiting the complementary information provided by different sensors has become an important research direction for improving bearing fault diagnosis under complex operating conditions [5].

Traditional bearing fault diagnosis methods primarily extract time-domain, frequency-domain, and time–frequency-domain features from vibration [6], current [7], or acoustic emission signals [8]. To address the nonstationarity of fault-induced impulses and weak modulation characteristics, empirical mode decomposition (EMD) [9] and wavelet transform (WT) [10] have been employed to decompose signals and extract fault information at different time scales and within different frequency ranges. He et al. [11] proposed a fault diagnosis method integrating variational mode decomposition (VMD), energy entropy, and deep learning. Dong et al. [12] used empirical wavelet transform (EWT) to decompose signals into multiple frequency components, thereby enhancing feature extraction and improving the diagnostic input. Yu et al. [13] proposed a method combining the short-time Fourier transform and synchrosqueezed wavelet transform to integrate time-domain and frequency-domain information, achieving favorable diagnostic accuracy and robustness across multiple datasets. These methods can extract physically meaningful fault features from specific frequency bands or scales, providing an important foundation for bearing condition identification. However, their diagnostic performance generally depends on the selection of decomposition parameters, analysis frequency bands, and handcrafted features. When the rotational speed, load, or noise level changes, preset parameters and feature combinations may not consistently preserve effective fault information. Moreover, processing vibration and current signals separately makes it difficult to fully characterize the complementary relationship between the two types of measurements, thereby limiting the integrated use of multisource monitoring information.

With the development of deep learning, convolutional neural networks [14], deep neural networks [15], residual networks [16], long short-term memory networks [17], and Transformers [18] have been widely applied to the automatic learning of bearing fault features. Existing studies have mainly followed two directions to improve feature representation capability. The first expands the local receptive field through multi-scale convolutions, residual connections, and attention mechanisms. For example, Tang et al. [19] introduced a dual multi-scale quadratic attention CNN that integrates quadratic neurons with multi-scale attention, while Liao et al. [20] designed a multi-scale residual network with an enhanced gated recurrent unit (MSRN-EGRU). The second direction combines convolutional feature extraction with sequential dependency modeling. For instance, Chen et al. [21] combined a multi-scale CNN with an LSTM network for multi-frequency feature extraction; Guo et al. [22] enhanced a CNN–LSTM architecture using an attention mechanism; and Song et al. [23] employed an improved particle swarm optimization algorithm to optimize the hyperparameters of a CNN–BiLSTM model. In addition, generative data augmentation and hybrid networks have been used to alleviate class imbalance and insufficient training data [24]. These methods have improved local feature extraction, long-range dependency modeling, and diagnostic performance under limited-sample con-

ditions. However, increasing network depth or stacking general-purpose modules cannot automatically determine the appropriate stage at which multimodal information should interact during feature extraction, nor can it clarify how diagnostic contributions should be allocated between vibration and current features at different network depths.

In addition to network architecture, the compactness, class structure, and separability of feature representations are important factors affecting diagnostic performance. Zhao et al. [25] combined dimensionality reduction with graph-based sparse subspace clustering for rolling bearing fault identification, demonstrating that properly organizing low-dimensional features while preserving intra-class structure can improve the discrimination of fault categories. Spirito et al. [26] constructed feature indices characterizing graphical density, orientation, and curvature from the symmetrized dot patterns of vibration signals and combined them with a feedforward neural network for rolling bearing fault detection, further indicating that both effective feature organization and the classification model influence diagnostic performance. These studies suggest that the evaluation of fault diagnosis models should not focus solely on network size and final classification accuracy but should also examine how intermediate features evolve as they propagate through the network and assess their class separability. For multimodal networks, it is also necessary to further investigate how information from different modalities contributes to feature learning and whether the feature distributions before and after fusion are more conducive to class discrimination.

To overcome the incomplete information provided by a single signal, multimodal fault diagnosis methods have attracted increasing attention. Chen et al. [27] employed spatiotemporal attention to adjust feature responses across different temporal positions and channels. Yang et al. [28] used a network combining convolution and attention to achieve deep feature fusion of vibration and current signals. Cui et al. [29] integrated multimodal feature fusion with marginal and conditional distribution alignment, while Liu et al. [30] performed feature fusion at different stages through multi-channel adaptation and multi-domain correlation. Localized bearing damage is typically manifested in vibration signals as transient impulses with temporal locality, and spatial attention can highlight the critical positions corresponding to these fault-induced impulses [31]. In contrast, bearing fault responses in current signals are relatively weak and are susceptible to interference from the power-supply fundamental frequency, harmonics, and noise. Channel attention can enhance fault-related channels while suppressing redundant responses [32]. These studies indicate that, although existing methods can improve the utilization of multimodal information to some extent, three interrelated issues remain. First, some methods perform only a single fusion operation, primarily at the end of the network, preventing early complementary information from contributing to subsequent feature extraction. Second, the contributions of vibration and current signals to the classification task may vary across input samples and feature depths, and fusion weights assigned at a single level cannot adequately characterize such variations. Third, relying solely on the fusion result from the final level may overlook local impulses, modulation responses, and transitional features preserved in intermediate layers.

To provide a clearer comparison of the technical contributions and limitations of existing studies, representative research on bearing fault diagnosis is summarized in terms of its main contributions and remaining research gaps, as presented in Table 1.

Table 1. Comparison of representative studies on bearing fault diagnosis.

Representative Studies	Research Contributions	Research Gaps
He et al. [11]	Enhanced the extraction of nonstationary fault components	Reliance on decomposition parameters and failure to exploit complementary multimodal information
Dong et al. [12] and Yu et al. [13]	Improved multi-scale and time–frequency feature representations	Results affected by preprocessing procedures and parameter selection, with insufficient cross-modal interaction
Tang et al. [19] and Liao et al. [20]	Accounted for both local receptive fields and temporal dependencies	Reliance on a single modality and inability to characterize variations in modality contributions
Chen et al. [21], Guo et al. [22], and Song et al. [23]	Improved sequence modeling and diagnostic performance under limited-sample conditions	The stacking of network modules does not address when multimodal interaction should occur or how features should be fused across different levels
Chen et al. [27] and Yang et al. [28]	Exploited the complementarity between mechanical and current responses	Insufficient consideration of fusion levels and variations in modality weights
Cui et al. [29] and Liu et al. [30]	Enhanced cross-domain or multi-stage feature associations	Early cross-modal interaction, hierarchical dynamic weighting, and multi-level aggregation have not yet been integrated into a unified framework

To address the aforementioned research gaps, this study proposes a Multi-Scale Feature Enhancement and Adaptive Fusion Network (MSFEAF-Net). The primary innovation of the proposed network lies in its dynamic multimodal fusion mechanism, which is both sample-dependent and feature-level-dependent. This mechanism enables the network to adaptively assign the diagnostic contributions of vibration and current signals at different feature depths, while progressively aggregating the fusion results from multiple levels through a feature pyramid. The main contributions of this study are summarized as follows:

- (1) A dual-branch multi-scale feature extraction and early cross-modal interaction architecture is developed. While retaining independent feature-learning pathways for vibration and current signals, bidirectional cross-attention is introduced after the first residual stage to incorporate early complementary information into subsequent feature extraction. Different attention mechanisms are then employed to enhance the features according to the representational characteristics of the two signals.
- (2) A sample-dependent and level-dependent modality fusion method is developed. Based on each input sample, the network adaptively generates modality weights for the vibration and current features at different feature levels. A feature pyramid is subsequently used to aggregate the multi-level fusion results, thereby improving the utilization of diagnostic information across different levels.
- (3) Comprehensive experiments are conducted on the PU and KAIST datasets, including single-modality and multimodal comparisons, evaluations under different signal-to-noise ratios, small-sample experiments, module ablation studies, feature visualization, and cross-operating-condition experiments. These experiments comprehensively evaluate the diagnostic performance and effectiveness of the proposed network under different conditions.

The rest of this paper is structured as follows. Section 2 explains the necessary theoretical background. Section 3 clarifies the detailed network modules of MSFEAF-Net.

Section 4 provides the experimental tests and discusses the results. Lastly, Section 5 brings the conclusion and gives future study paths.

2. Theoretical Foundations

2.1. One-Dimensional Convolutional Neural Network

As a classic deep learning architecture, convolutional neural networks (CNNs) rely on local receptive fields and shared weights. They work exceptionally well for image recognition and time-series analysis [33], its network architecture is illustrated in Figure 1. In rotating machinery health monitoring, both vibration and current signals constitute typical one-dimensional sequences; consequently, the adoption of a 1D-CNN can effectively extract their local temporal features.

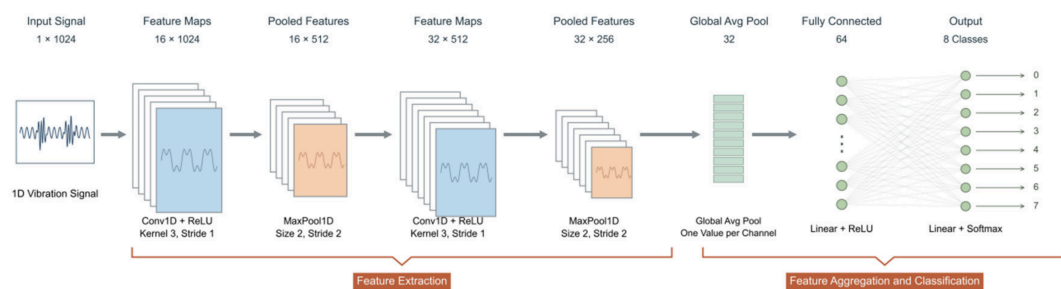


Figure 1. Architecture of the one-dimensional convolutional neural network.

For any sequential input vector $x(t)$, Equation (1) defines the standard 1D convolution processing:

$$y(t) = \sum_{i=0}^{k-1} x(t+i) \cdot w(i) + b \quad (1)$$

where $w(i)$ represents the filter parameter matrix, while k and b signify the window length and bias value, respectively.

After the convolution operation, the ReLU activation function is applied to nonlinearly map the convolution output, which can be expressed as

$$z(t) = \sigma(y(t)) = \max(0, y(t)) \quad (2)$$

where $\sigma(\cdot)$ denotes the ReLU activation function, and $z(t)$ represents the activated output feature.

By sliding convolutional kernels over the input signal to perform local weighted summation, a one-dimensional convolutional neural network can extract characteristic patterns within different temporal windows.

2.2. Residual Network

As the network depth increases, conventional CNNs are susceptible to vanishing gradients or performance degradation. To alleviate this issue, the residual network (ResNet) introduces skip connections to implement identity mappings, thereby enabling the network to learn residual functions rather than direct mappings [34]; its network architecture is illustrated in Figure 2. Its fundamental form can be mathematically expressed as

$$y = F(x) + x \quad (3)$$

where $F(x)$ denotes the residual mapping function, and x represents the input features.

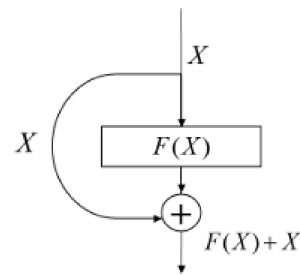


Figure 2. Architecture of the residual network.

The residual mapping function can also be expressed as the difference between the target mapping and the input features:

$$F(x) = H(x) - x \quad (4)$$

where $H(x)$ denotes the target mapping to be learned by the network.

This architecture effectively alleviates the training difficulties of deep networks while preserving raw signal information, which is of paramount significance for identifying weak fault features.

2.3. Convolutional Block Attention Module

The convolutional block attention module (CBAM) is a lightweight adaptive attention mechanism that can achieve adaptive feature enhancement and irrelevant feature suppression from both channel and spatial dimensions [35]. This module is composed of two sub-modules connected in series: the channel attention and spatial attention modules, and its overall network architecture is illustrated in Figure 3.

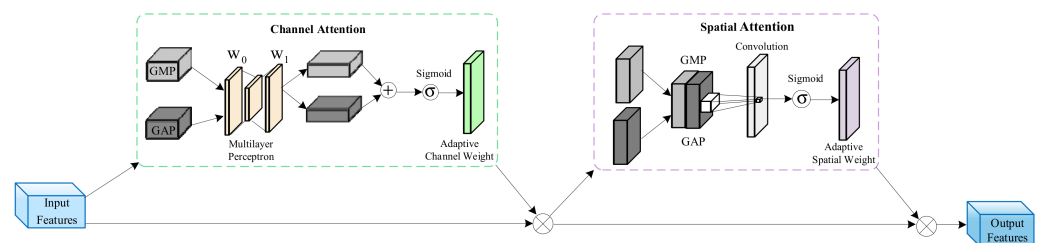


Figure 3. Architecture of the convolutional block attention module (CBAM).

The channel attention module computes the importance weights of individual channels by modeling inter-channel dependencies. Specifically, global average pooling and max pooling are first applied to the input feature map to derive two distinct channel descriptors. Subsequently, a channel attention map is generated by routing these descriptors through a shared multi-layer perceptron (MLP) and a sigmoid activation function. Concurrently, the spatial attention module focuses on the salient regions within the feature maps. To this end, average pooling and max pooling operations are executed along the channel dimension on the channel-refined features. The resulting two feature maps are then concatenated and processed through a 7×1 convolution and a sigmoid function to generate the spatial attention map. By sequentially cascading these two stages of attention, the CBAM achieves a comprehensive and refined calibration of features. Its complete operational workflow can be mathematically formulated by Equations (5) and (6).

$$F' = F \otimes M_c(F) \quad (5)$$

$$F'' = M_s(F') \otimes F' \quad (6)$$

where $\otimes$ denotes element-wise multiplication, $M_c(F)$ represents the intermediate feature representation scaled by channel attention, and $M_s(F')$ signifies the final feature representation refined by spatial attention.

Through the aforementioned workflow, the CBAM effectively enhances the discriminative power of features. Furthermore, its plug-and-play characteristic facilitates seamless integration into existing networks, thereby significantly bolstering model performance in tasks such as fault diagnosis.

2.4. Transformer Model

The Transformer framework operates primarily on self-attention mechanisms. Through global feature modeling, it handles long-range sequence dependencies with high efficiency [36]; its basic architecture is illustrated in Figure 4. Its fundamental paradigm revolves around computing correlation scores among different positions within a sequence to achieve the adaptive, weighted fusion of features. Specifically, given an input feature sequence, it is first mapped into Query (Q), Key (K), and Value (V) matrices via linear projections. The corresponding self-attention weight calculation is mathematically formulated as shown in Equation (7):

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

where d_k denotes the dimension of the key vectors, and $\sqrt{d_k}$ serves as the scaling factor.

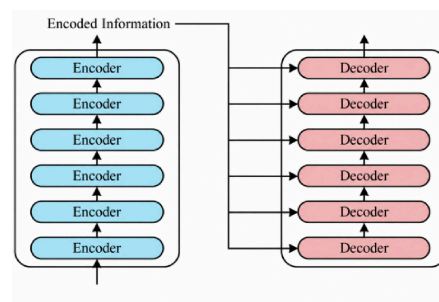


Figure 4. Architecture of the Transformer network.

Through the aforementioned computation, the feature representation at each individual position can comprehensively integrate information from all positions within the sequence, thereby achieving global dependency modeling. Compared with conventional convolutional or recurrent structures, the self-attention mechanism can capture long-range dependent features with superior efficiency.

To further elevate the representation capacity of the model, the Transformer incorporates a multi-head attention mechanism. By configurationally deploying multiple attention heads in parallel, this mechanism projects the input into distinct representation subspaces to perform feature learning, thereby enhancing the model's capability to characterize complex patterns. The overall mathematical expression of the multi-head attention is formulated as shown in Equation (8):

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W_O \quad (8)$$

where h denotes the number of attention heads, and W_O represents the output projection matrix.

Through this multi-head architecture, the model is capable of extracting complementary information from diverse representation subspaces, thereby enriching the feature representation with enhanced diversity and comprehensiveness.

Equations (1)–(8) describe the fundamental operations of convolution, residual mapping, CBAM, and the Transformer adopted in this study, respectively, with their theoretical foundations provided in Refs. [33–36]. The characteristics and functions of these four basic network modules are summarized in Table 2.

Table 2. Characteristics and functions of the basic network modules.

Module	Core Architecture	Key Characteristics	Source
1D-CNN	Local convolution and weight sharings	Extracts local temporal patterns with relatively low computational cost	Ref. [33]
ResNet	Residual mapping and skip connections	Alleviates degradation in deep networks while preserving the original information	Ref. [34]
CBAM	Channel attention and spatial attention	Recalibrates features along the channel and temporal-position dimensions	Ref. [35]
Transformer	Self-attention and multi-head attention	Models long-range dependencies and cross-sequence relationships	Ref. [36]

3. Design of the MSFEAF-Net Model

3.1. Overall Framework of the MSFEAF-Net Model

To address the limitations that a single fusion operation at the end of the network cannot effectively exploit early complementary information and that single-level weighting cannot capture variations in modal contributions across different samples and feature depths, this study proposes a multi-scale feature enhancement and adaptive fusion network (MSFEAF-Net), whose overall architecture is illustrated in Figure 5. MSFEAF-Net is a complete diagnostic model composed of multiple functional modules connected sequentially. These modules are not independent models used for comparison; instead, they perform distinct tasks, including feature extraction, cross-modal enhancement, within-level modality fusion, cross-level feature aggregation, and feature selection before classification. First, the vibration and current signals are separately fed into a dual-branch multi-scale residual network, where one-dimensional convolutions at different scales extract local temporal features, while residual structures ensure the stability of deep feature learning. The feature enhancement module is placed only after the first multi-scale residual block (MSRB1). It employs self-attention and bidirectional cross-attention to capture intra-modal dependencies and cross-modal complementary relationships, respectively, and then applies one-dimensional spatial attention and channel attention to enhance the vibration and current features. The two enhanced feature streams are subsequently passed through the remaining MSRBs for further feature extraction. At the 128-, 256-, and 512-channel levels, the dynamic feature fusion module generates fusion weights for the two modalities based on each input sample through global average pooling, feature concatenation, a two-layer fully connected mapping, and softmax normalization, thereby achieving sample-dependent and level-dependent feature weighting. Subsequently, the multi-scale feature pyramid aggregates the fusion results from the three levels along a deep-to-shallow pathway. Finally, the adaptive feature selection module refines the fused features through parallel multi-scale convolutions and joint channel–spatial attention, after which the classifier outputs the bearing condition. By sequentially integrating early cross-modal interaction, hierarchical

modality weighting, and cross-level feature aggregation, the proposed network enables complementary information from different stages to contribute jointly to fault identification.

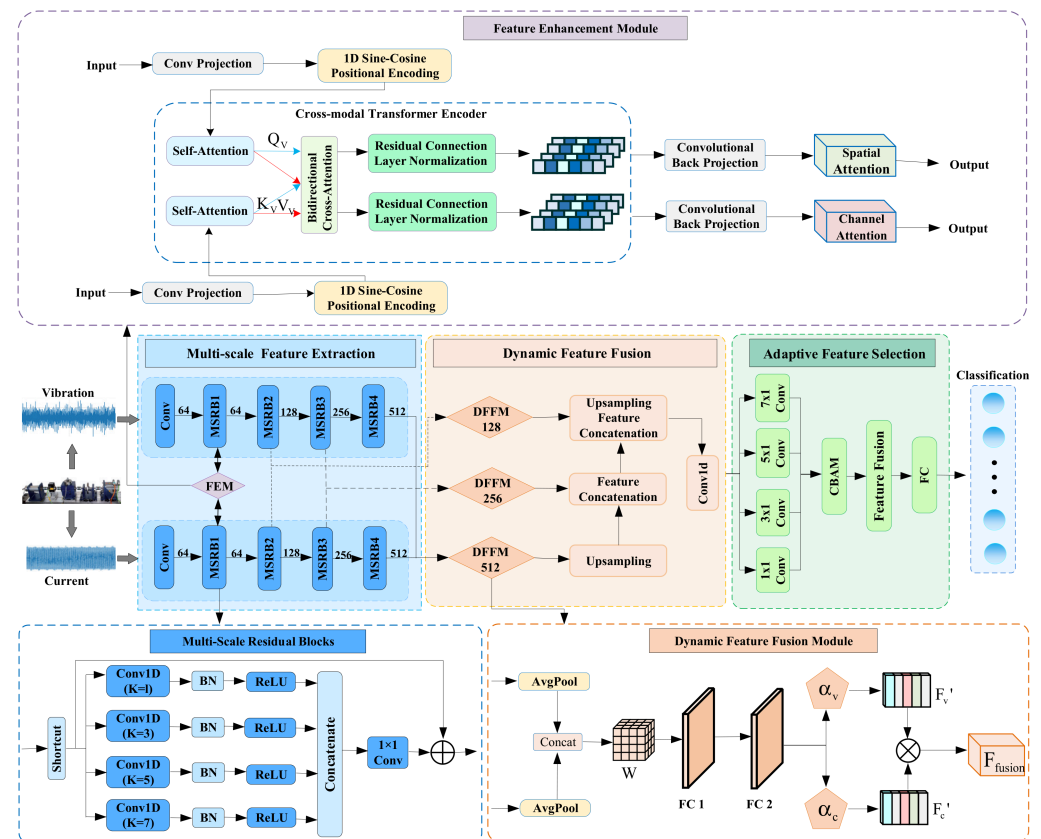


Figure 5. Overall architectural framework of the MSFEAF-Net model.

3.2. Dual-Branch Multi-Scale Residual Feature Extraction Module

This study constructs a dual-branch multi-scale residual feature extraction network with identical branch architectures but independently learned parameters to effectively extract features from bearing vibration and current signals. The network takes one-dimensional temporal vibration and current signals as inputs and first maps each single-channel signal to a 64-channel feature representation through an initial convolutional block. Each branch then consists of four stacked multi-scale residual blocks (MSRBs), with output channel dimensions of 64, 128, 256, and 512, respectively. The detailed network architecture is shown in the lower-left corner of Figure 2. The identical branch architectures ensure that the two modalities have corresponding feature levels, whereas the use of non-shared parameters enables each branch to independently learn feature representations from its respective input signal. Specifically, the input feature matrix of the l -th MSRB can be mathematically formulated as

$$F_{in}^l \in \mathbb{R}^{C_{in} \times L_{in}} \quad (9)$$

where C_{in} and L_{in} denote the number of input channels and the sequence length, respectively.

To comprehensively capture the high-frequency transient impulses excited by diverse faults and the low-frequency electromagnetic fundamental wave characteristics, the module deploys four different scales of convolutional kernels ($K \in \{1, 3, 5, 7\}$) in parallel. For the k -th parallel scale branch, the feature transformation can be mathematically expressed as

$$F_k^l = \sigma \left(f_{\text{BN}} \left(f_{\text{Conv}}^{k \times 1} \left(F_{\text{in}}^l \right) \right) \right), k \in \{1, 3, 5, 7\} \quad (10)$$

where $f_{\text{Conv}}^{k \times 1}(\cdot)$ denotes the 1D convolution operation with a kernel size of k and a stride of 1; $f_{\text{BN}}(\cdot)$ represents the Batch Normalization operator; and $\sigma(\cdot)$ signifies the ReLU non-linear activation function.

Specifically, adaptive padding is applied to all branches to keep the output dimensions equal to the input length. Subsequently, the four feature matrices embedding distinct receptive fields are concatenated along the channel dimension, and this concatenation process can be mathematically expressed as

$$F_{\text{concat}}^l = \text{Concat} \left(F_1^l, F_3^l, F_5^l, F_7^l \right) \in \mathbb{R}^{4C_{\text{mid}} \times L_{\text{in}}} \quad (11)$$

Because the number of feature channels expands fourfold after concatenation, a 1×1 1D convolutional layer is introduced to compress the channel dimensionality, thereby facilitating cross-channel information interaction:

$$F_{\text{fusion}}^l = f_{\text{Conv}}^{1 \times 1} \left(F_{\text{concat}}^l \right) \in \mathbb{R}^{C_{\text{out}} \times L_{\text{in}}} \quad (12)$$

Finally, to safeguard gradient stability during deep forward propagation, the fused multi-scale features are element-wise aggregated with the identity mapping from the shortcut branch, yielding the ultimate output representation of the MSRB:

$$F_{\text{out}}^l = F_{\text{fusion}}^l + H_{\text{shortcut}} \left(F_{\text{in}}^l \right) \quad (13)$$

where H_{shortcut} denotes the dimensionality alignment operator. When $C_{\text{in}} = C_{\text{out}}$, this operator serves as an identity mapping; when the number of channels changes (i.e., $C_{\text{in}} \neq C_{\text{out}}$), the operator employs a 1D convolution to perform a linear projection, thereby achieving channel alignment.

From a macroscopic perspective, as the network depth increases, the four MSRBs within the dual branches adopt a channel-incremental stacking strategy, where the number of feature channels expands from 64 to 128, 256, and 512 sequentially. To maximize the utilization of information complementarity across different network hierarchies, this paper designs a unique layer-wise shunting and extraction mechanism. Instead of simple spatial concatenation, the outputs of MSRB2, MSRB3, and MSRB4 serve as shunting nodes for multi-scale features, vertically routing characteristics from diverse levels into the multi-scale dynamic feature fusion architecture:

$$\begin{cases} F_{\text{DFFM},128} = F_{\text{out}}^2 \in \mathbb{R}^{128 \times L_2} \\ F_{\text{DFFM},256} = F_{\text{out}}^3 \in \mathbb{R}^{256 \times L_3} \\ F_{\text{DFFM},512} = F_{\text{out}}^4 \in \mathbb{R}^{512 \times L_4} \end{cases} \quad (14)$$

where $F_{\text{DFFM},128}$, $F_{\text{DFFM},256}$ and $F_{\text{DFFM},512}$ represent the low-level waveform features, mid-level localized impulse characteristics, and high-level global fault information dispatched to the dynamic feature fusion modules DFFM128, DFFM256, and DFFM512, respectively.

This hierarchically extracted, multi-scale feature output mechanism not only preserves the subtle transient fault abruptions captured by the shallow network, but also integrates the highly discriminative insights from the deep network. This mechanism significantly enhances the model's anti-interference robustness under complex operating conditions and heavy background noise.

3.3. Feature Enhancement Module

A feature enhancement module (FEM) is designed, and its specific network architecture is illustrated in the upper dashed box of Figure 2. Adopting a twin-stream symmetrical architecture, this module first utilizes 1D convolutional projection layers to uniformly map the dual-branch features into 256-dimensional feature sequences, thereby eliminating dimensional heterogeneity. Concurrently, 1D sine-cosine positional encodings are introduced to dynamically embed temporal information. Subsequently, the dual-branch encoded sequences are routed into a cross-modal Transformer encoder. Within this encoder, self-attention mechanisms are first employed to establish intra-modal long-term sequential dependencies, generating their respective Query, Key, and Value matrices. Following this, the feature matrices from both branches enter the core interaction layer to undergo deep non-linear interaction, wherein bidirectional cross-attention is executed to fully mine fault-related cross-modal dynamic correlations. Finally, residual connections and layer normalization (LN) are applied to effectively suppress deep network degradation. The layer-wise evolutionary formulas are mathematically expressed as follows:

$$\hat{Z}_v = f_{LN} \left(Z_v + \text{Softmax} \left(\frac{Q_v K_c^T}{\sqrt{d_k}} \right) V_c \right) \quad (15)$$

$$\hat{Z}_c = f_{LN} \left(Z_c + \text{Softmax} \left(\frac{Q_c K_v^T}{\sqrt{d_k}} \right) V_v \right) \quad (16)$$

where $f_{LN}(\cdot)$ denotes the layer normalization operator, and $\sqrt{d_k}$ represents the scaling factor.

After intermediate feature alignment, the output feature sequences are restored to their original channel dimensions through a one-dimensional convolutional back-projection layer (Convolutional Back Projection). Subsequently, considering the differences in the formation mechanisms and feature manifestations of fault information in vibration and current signals, a modality-specific attention enhancement mechanism is introduced at the end of the module. Specifically, one-dimensional spatial attention is applied to the vibration branch, whereas channel attention is applied to the current branch.

After propagating through the mechanical structure, transient impulses generated by localized bearing damage typically appear as local peaks or damped oscillations at specific temporal positions in the vibration sequence, and their discriminative information exhibits strong temporal locality. Therefore, the vibration branch employs one-dimensional spatial attention to enhance critical impulse locations while suppressing responses from irrelevant temporal regions. In contrast, bearing faults affect the stator current through electromechanical coupling effects, including dynamic eccentricity, unbalanced magnetic pull, and torque fluctuations. The resulting fault-related modulation components have relatively low amplitudes and can be easily masked by the power-supply fundamental frequency, harmonics, and noise. After multi-channel convolutional mapping, this fault information can be represented by the combined responses of different latent feature channels. Therefore, the current branch employs channel attention to recalibrate the relative contributions of the latent channels. Based on the temporal locality of vibration fault information and the differentiated responses of current fault information across latent channels, this modality-specific configuration enhances effective information along the temporal-position and feature-channel dimensions, respectively. The specific formulations are mathematically expressed as

$$M_v(F) = H_v \otimes \sigma \left(f_{Conv}^{7 \times 1} \left([f_{Avg}(H_v); f_{Max}(H_v)] \right) \right) \quad (17)$$

$$M_c(F) = H_c \otimes \sigma \left(f_{MLP} \left(f_{Avg}(H_c) \right) + f_{MLP} \left(f_{Max}(H_c) \right) \right) \quad (18)$$

where $\otimes$ denotes element-wise multiplication, $\sigma(\cdot)$ signifies the sigmoid activation function, $f_{\text{Avg}}(\cdot)$ and $f_{\text{Max}}(\cdot)$ represent the average pooling and max pooling operators along the channel or spatial dimension, respectively, and $f_{\text{MLP}}(\cdot)$ corresponds to the shared multi-layer perceptron mapping.

After modality-specific attention enhancement, the enhanced vibration and current features are fed into the subsequent multi-scale residual blocks, allowing early cross-modal information to contribute to feature extraction at later levels and thereby providing highly discriminative feature representations for the subsequent multi-scale dynamic feature fusion module (DFFM).

3.4. Dynamic Feature Fusion Architecture

The contributions of different modalities to the diagnostic task are affected by the fault category, operating condition, and signal quality, and may also vary with network depth. Performing only a single feature concatenation before the classifier or applying identical fusion coefficients to all samples cannot adequately capture these differences. Therefore, a Dynamic Feature Fusion Module (DFFM) is deployed at each of the three feature levels with 128, 256, and 512 channels to generate corresponding modality weights based on the feature representations of the current input sample. A multi-scale feature pyramid is subsequently employed to aggregate the fusion results from the three levels. This architecture handles same-level modality fusion and cross-level feature aggregation separately, as illustrated in Figure 2.

First, global average pooling is applied to the enhanced current and vibration features, compressing each channel along the temporal dimension into a single statistical value. The resulting features are then concatenated so that the global descriptors of both modalities jointly serve as the input for weight generation. Consequently, the weight assigned to one modality is determined not only by its own features but also by the current response of the other modality. This process can be expressed as follows:

$$G = \text{Concat}(G_v, G_c) \quad (19)$$

The concatenated features are mapped to two modality-specific scores through a two-layer fully connected network and then normalized using the softmax function:

$$\alpha = W_2 \delta(W_1 G + b_1) + b_2 \quad (20)$$

$$\alpha = [\alpha_v, \alpha_c] \quad (\alpha_v + \alpha_c = 1) \quad (21)$$

where W_1 , W_2 , b_1 and b_2 are learnable parameters used for weight generation, and δ denotes the nonlinear activation function; α_v and α_c represent the vibration and current weights assigned to the current sample at this feature level, respectively.

Finally, the generated modality weights are used to scale the two feature streams, and the fused feature at the current level is obtained through element-wise weighted summation. The resulting level-specific fused feature can be expressed as

$$F_{\text{fusion}} = \alpha_v \cdot F'_v + \alpha_c \cdot F'_c \quad (22)$$

On this basis, to fully exploit the complementary information available at different feature depths, a multi-scale feature pyramid module (MSFPM) is introduced to progressively integrate heterogeneous fused features along a deep-to-shallow pathway. As shown in the orange region of Figure 2, the MSFPM first applies linear interpolation upsampling to the 512-channel deep features and concatenates them with the 256-channel impulse features at the corresponding node. Following the two-stage upsampling process, the resulting

composite features are further aggregated with the 128-channel shallow time-domain waveform features. Finally, the high-dimensional multi-scale representation is processed by a one-dimensional convolutional layer (Conv1d) for channel balancing and feature smoothing, and the resulting pyramid-smoothed feature (F_{pyramid}) can be expressed as

$$F_{\text{pyramid}} = \sigma \left(f_{\text{BN}} \left(f_{\text{Conv1D}}^{1 \times 1} \left(F_{\text{cat}}^{128} \right) \right) \right) \in \mathbb{R}^{C_{\text{out}} \times L_{128}} \quad (23)$$

This pyramid fusion mechanism preserves the class-level semantics of the deep features while supplementing them with local impulses, modulation responses, and transitional information retained in the shallow and intermediate layers. The resulting multi-scale aggregated features are then fed into the adaptive feature selection module for further refinement before classification. The principal improvement of this study lies in deploying input-dependent modality weighting at multiple feature levels and combining it with a feature pyramid to aggregate the fusion results from each level, thereby reducing the loss of intermediate information caused by performing only a single fusion operation at the end of the network.

3.5. Adaptive Feature Selection Module

After dynamic modality fusion and feature pyramid aggregation, the output features contain information from both different modalities and different network depths. However, cross-level concatenation and convolutional integration may retain redundant responses and features that are only weakly relevant to the classification task. To further refine the fusion results, this study constructs an Adaptive Feature Selection Module (AFSM). As illustrated in the green-shaded region on the right side of Figure 2, this module is deployed downstream of the feature pyramid smoothing layer, and its front-end core consists of parallel 1D convolutional layers with four distinct scales ($1 \times 1, 3 \times 1, 5 \times 1, 7 \times 1$). Upon injecting the smoothed and calibrated pyramid features into this layer, spatial reconstruction and multi-scale adaptive filtering are performed through these 1D convolutional layers embedding diverse receptive fields, thereby achieving the preliminary spatial screening of fault features. This spatial evolutionary process can be mathematically formulated as

$$F_{\text{sc}} = \sum_{k \in \{1, 3, 5, 7\}} \sigma \left(f_{\text{BN}} \left(f_{\text{Conv}}^{k \times 1} \left(F_{\text{pyramid}} \right) \right) \right) \in \mathbb{R}^{C \times L} \quad (24)$$

where $f_{\text{Conv}}^{k \times 1}(\cdot)$ denotes the 1D convolution operator of the k -th scale, while $f_{\text{BN}}(\cdot)$ and $f_{\text{BN}}(\cdot)$ represent the batch normalization layer and the ReLU non-linear activation function, respectively.

Subsequently, the aggregated features are fed into the convolutional block attention module (CBAM), in which the channel attention submodule adaptively emphasizes fault-related feature channels, while the spatial attention submodule precisely captures key temporal segments containing transient impulses. After being screened by the CBAM, the features are fused and transformed into a 1D global discriminative vector. This vector is then fed into a fully connected (FC) layer to undergo deep non-linear mapping, maximizing the spatial separability among different fault categories, and is ultimately processed by the classifier to yield precise and intelligent diagnostic results. In summary, the AFSM effectively eliminates information redundancy and preserves high-purity fault characteristics, ensuring high diagnostic accuracy and robust stability of the model under environments characterized by heavy noise and complex operating conditions.

4. Experimental Verification and Discussion

4.1. Experimental Setup

4.1.1. Dataset Description

This paper uses the Paderborn University (PU) dataset [37] and the Korea Advanced Institute of Science and Technology (KAIST) dataset [38] to experimentally validate the MSFEAF-Net.

PU Dataset: The dataset utilized in this study originates from the KAT Data Center at Paderborn University in Germany. It primarily encompasses real damages induced by drilling, manual electric engraving, electrical discharge machining (EDM), and accelerated life tests, which can effectively simulate actual bearing fault operating environments. The experimental platform for the PU dataset is illustrated in Figure 6, where the collected signals comprise both vibration and current sequences. The test rig captured normal and fault data under bearing operating conditions for a duration of 4 s at a sampling frequency of 64k Hz. Table 3 lists the specific settings for individual test conditions. Artificial defects often lack realistic signatures. Therefore, we use actual fatigue wear from the hardware rig. These real-world samples provide a better platform to verify our model. The comprehensive fault types and relevant parameters are listed in Table 4, which includes seven accelerated life test damaged bearing samples and one healthy bearing sample, with each sample possessing corresponding vibration and current signal segments.

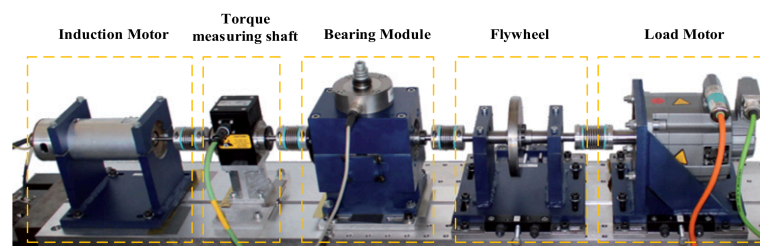


Figure 6. Experimental test rig for the Paderborn University (PU) bearing dataset.

Table 3. Operating Condition Parameters of the PU Dataset.

Conditions	Rotational Speed/rpm	Load Torque/Nm	Radial Force/N	Name
1	900	0.7	1000	N09_M07_F10
2	1500	0.7	1000	N15_M07_F10
3	1500	0.1	1000	N15_M01_F10
4	1500	0.1	400	N15_M01_F04

Table 4. Description of Fault Categories and Sample Configurations in the PU Dataset.

Label	Code	Fault Type	Fault Feature	Level
0	K001	Normal	-	0
1	KA04	OR	Single	1
2	KA16	OR	Repetitive	2
3	KB23	OR and IR	Multiple	2
4	KI14	IR	Multiple	1
5	KI16	IR	Single	3
6	KI18	IR	Single	2
7	KI21	IR	Single	1

KAIST Dataset: This dataset is provided by the Human Lab at the Center for Noise and Vibration Control Plus, Korea Advanced Institute of Science and Technology (KAIST).

It primarily encompasses rotating machinery operating environments simulated through artificially seeded cracks of varying sizes, shaft misalignment, and rotor unbalance. The layout of the experimental platform for the KAIST dataset is illustrated in Figure 7, where the acquired multi-sensor signals comprise vibration, acoustic, temperature, and motor current sequences. Operating at a constant rotational speed of 3010 RPM, the test rig captured normal and fault data of vibration and current at a sampling frequency of 25.6k Hz. The detailed parameters for each operating condition are tabulated in Table 5. The data acquisition durations for the healthy and faulty states are 120 s and 60 s, respectively. Compared with datasets containing only a single operating condition, this dataset introduces different load gradients of 0 N·m, 2 N·m, and 4 N·m, which effectively reflect the real-world load fluctuation challenges in industrial fields. Consequently, it is highly suitable for evaluating the performance of multimodal fusion fault diagnosis algorithms. The comprehensive fault types are listed in Table 6, which covers multiple operational states including healthy state, inner ring fault, outer ring fault, shaft misalignment, and rotor unbalance. Crucially, each sample provides precisely synchronized multi-sensor signal segments.

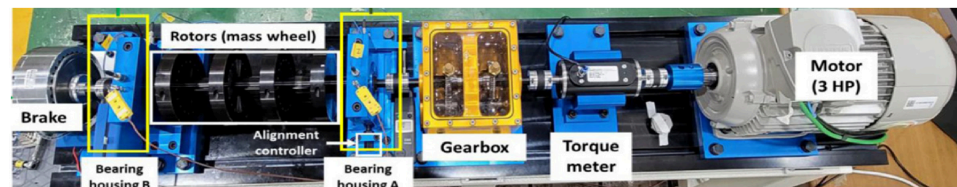


Figure 7. Configuration and sensor layout of the KAIST experimental test rig.

Table 5. Operating Condition Parameters for the KAIST Dataset.

Conditions	Rotational Speed/rpm	Load Torque/Nm
1	3010	0
2	3010	2
3	3010	4

Table 6. Description of Fault Categories and Sample Configurations in the KAIST Dataset.

Label	Code	Fault Type	Crack Size (mm)
0	N	Normal	0
1	OR-03	OR	0.3
2	OR-10	OR	1
3	OR-30	OR	3
4	IN-03	IR	0.3
5	IN-10	IR	1
6	IN-30	IR	3
7	Misalign	Shaft Misalignment	0
8	Unbalance	Rotor Unbalance	0

To provide an intuitive understanding of the typical operating states considered in this study, Figure 8 presents schematic illustrations of the healthy state, inner-race fault, outer-race fault, shaft misalignment, and rotor unbalance. Under the healthy state, no evident structural abnormalities are present in the bearing raceways or the rotating system. Inner- and outer-race faults are primarily characterized by localized damage to the corresponding raceways, which tends to generate periodic impacts during operation. Shaft misalignment refers to an angular or positional deviation between the rotational centerlines of the driving and driven shafts, whereas rotor unbalance is caused by an offset between the rotor mass center and its rotational axis.

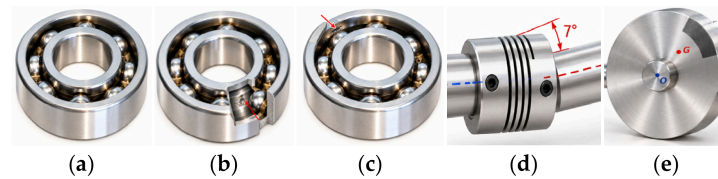


Figure 8. Schematic illustrations of five typical states of the rolling bearing and rotating system. (a) Healthy state; (b) inner-race fault; (c) outer-race fault; (d) shaft misalignment; (e) rotor unbalance.

4.1.2. Experimental Environment

We ran all tests under a unified environment to guarantee reproducible outputs. The hardware system runs on Windows 10 with an Intel i9-10900X CPU and 128 GB of RAM. An NVIDIA GeForce RTX 4070 GPU (12 GB VRAM; NVIDIA Corporation, Santa Clara, CA, USA) accelerates the computations. For software, we selected Python 3.8 and PyCharm IDE 2024.2.4. The deep learning framework is implemented based on PyTorch 2.0, which is integrated with CUDA 11.7 to fully leverage the GPU for parallel computational acceleration.

4.1.3. Training Strategy

To thoroughly evaluate the performance of the proposed MSFEAF-Net model, a comprehensive end-to-end training strategy is developed, spanning the entire pipeline from raw signal acquisition to the final diagnostic evaluation. Initially, synchronous vibration and current signals are captured from the bearing test rig. Subsequently, artificial background noise is superimposed onto the raw signals, and all noisy samples are further subjected to segmentation, labeling, and normalization. During the data preprocessing stage, the overlapping sliding window technique is implemented individually for the vibration and current sequences, with a window size of 2048 and a stride of 1024. The segmented samples are then partitioned into training, validation, and testing subsets according to a strict ratio of 7:2:1. Consequently, the number of samples per health category is configured as 240 for the PU dataset and 200 for the KAIST dataset. To guarantee a rigorous and fair benchmark comparison, identical hyperparameter profiles are applied across all comparative models. During training, the training set updates the network parameters. Concurrently, we evaluate the validation set at each epoch. This monitoring helps avoid potential overfitting. We choose the cross-entropy loss function for multi-class classification. After training, we save the checkpoint with the best validation accuracy. Finally, the testing set provides a systematic evaluation of the trained model. The detailed hyperparameter configurations designated for network training are summarized in Table 7.

Table 7. Hyperparameter Configurations.

Parameter	Value
Initial Learning Rate	0.00001
Minimum Learning Rate	0.000001
Batch size	16
Optimizer	AdamW
Epochs	100
Dropout	0.5

4.2. Comparative Study

Next, we execute comprehensive comparative trials on both the PU and KAIST datasets. These investigations confirm the validity and superiority of our framework. To ensure a rigorous and unbiased benchmark, all comparative models share identical

profiles regarding data partitioning, input sequence length, training epochs, optimizer configurations, and evaluation metrics. Each method is executed independently for 10 trials, with the average results reported to mitigate the stochastic effects induced by random initialization and sample perturbations. Tailored to the diagnostic tasks of this study, five representative deep learning frameworks are selected as comparative baselines. To facilitate an intuitive comparison of the network architectures of the models evaluated in the experiments, Table 8 lists their main components in the order of feature extraction, feature fusion, and classification output. The arrows indicate the direction of feature propagation through the network, and the number preceding each module denotes the number of times that module is repeated. Only the key modules that account for structural differences among the models are retained in the table to prevent excessive parameter details from obscuring the comparison. Specifically, the baselines encompass the following variants: MSFEAF-Net-V, a variant utilizing solely vibration signal inputs to benchmark the diagnostic capacity of a single vibration modality; MSFEAF-Net-C, a variant taking exclusively current signal inputs to quantify the tracking performance of a single electrical modality; ResNet, a dual-branch network employing fixed-scale residual blocks for cross-modal feature extraction, which completes modality fusion via straightforward vector concatenation for fault identification; FEMFN [39], a framework implementing a dual-branch Res-CBAM architecture to carry out feature extraction and subsequent screening for fault recognition; and MSF-DFormer [40], an advanced architecture executing multi-sensor fusion diagnosis by virtue of multi-scale convolutions, adaptive dense fusion, and a deformable Transformer. The aforementioned benchmarks span diverse methodology paradigms, including uni-modal modeling, conventional feature concatenation, multimodal attention-guided fusion, and Transformer-based integration, thereby facilitating a holistic validation of the performance edges of the proposed method.

Table 8. Experimental Model Network Structure.

Experimental Model	Network Structure
MSFEAF-Net-V	Conv1d + BN + ReLU + MaxPool → 4 × MSRB → Fusion → Upsampling + Fusion → MC + CBAM → AvgPool → FC + Dropout → FC
MSFEAF-Net-C	Conv1d + BN + ReLU + MaxPool → 4 × MSRB → Fusion → Upsampling + Fusion → MC + CBAM → AvgPool → FC + Dropout → FC
ResNet	Conv1d + BN + ReLU + MaxPool → 4 × RB → AvgPool → Flatten → Concat → FC + BN + ReLU + Dropout → FC
FEMFN	Conv1d + BN + ReLU + MaxPool → 3 × ResCBAM → Cross-attention + SE + Residual → AMF → Upsampling → Concatenation → PC + CA → AvgPool → FC + Dropout → FC
MSF-DFormer	MSRB + BN + ReLU + SCRM + SFFM → Upsampling + Weight + Fusion → Transformer → AvgPool → FC + Dropout → FC
MSFEAF-Net	Conv1d + BN + ReLU + MaxPool → 4 × MSRB → Transformer + CBAM → Fusion → Upsampling + Fusion → MC + CBAM → AvgPool → FC + Dropout → FC

Note: FC stands for Fully Connected layer; MC corresponds to Multi-scale Convolution; C indicates Channel; AMF represents Adaptive Multimodal Fusion; PC denotes Parallel Convolution; CA refers to Channel Attention; RB signifies Residual Blocks; SCRM denotes Spatial-Channel Reorganization Module; and SFFM represents Sensor Feature Fusion Module.

4.2.1. Comparative Analysis of Multimodal Fusion

To verify the diagnostic performance of the proposed MSFEAF-Net under baseline operating scenarios, multimodal fusion comparative experiments are first executed under the first operating condition of the PU dataset. As illustrated in Figures 9 and 10, for the uni-modal variants, the accuracy curve of MSFEAF-Net-V exhibits a rapid initial ascent and asymptotically stabilizes in the later stages, accompanied by a swift convergence of the

loss curve. This implies that the single vibration modality intrinsically possesses substantial diagnostic capability. Conversely, MSFEAF-Net-C exhibits a sluggish growth rate in accuracy with pronounced fluctuations, and its loss reduction is insufficient, demonstrating that the diagnostic performance is relatively deficient when relying solely on the current modality. Regarding the multimodal comparative frameworks, the accuracy curves of ResNet, FEMFN, and MSF-DFormer all experience a prompt rise before entering a steady state, while their corresponding losses decrease monotonically, demonstrating that multimodal fusion can effectively enhance the diagnostic efficacy. In comparison, the proposed MSFEAF-Net delivers the steepest ascent in accuracy, achieves the most stable convergence, and exhibits the most thorough and well-behaved loss decay in the later training phases.

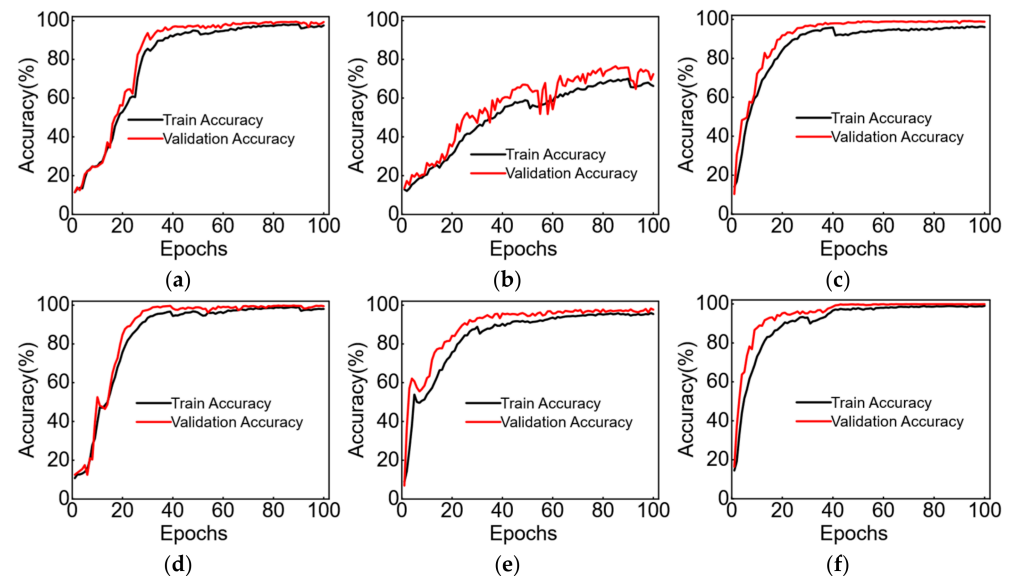


Figure 9. Accuracy Curve. (a) MSFEAF-Net-V; (b) MSFEAF-Net-C; (c) ResNet; (d) FEMFN; (e) MSF-DFormer; (f) MSFEAF-Net.

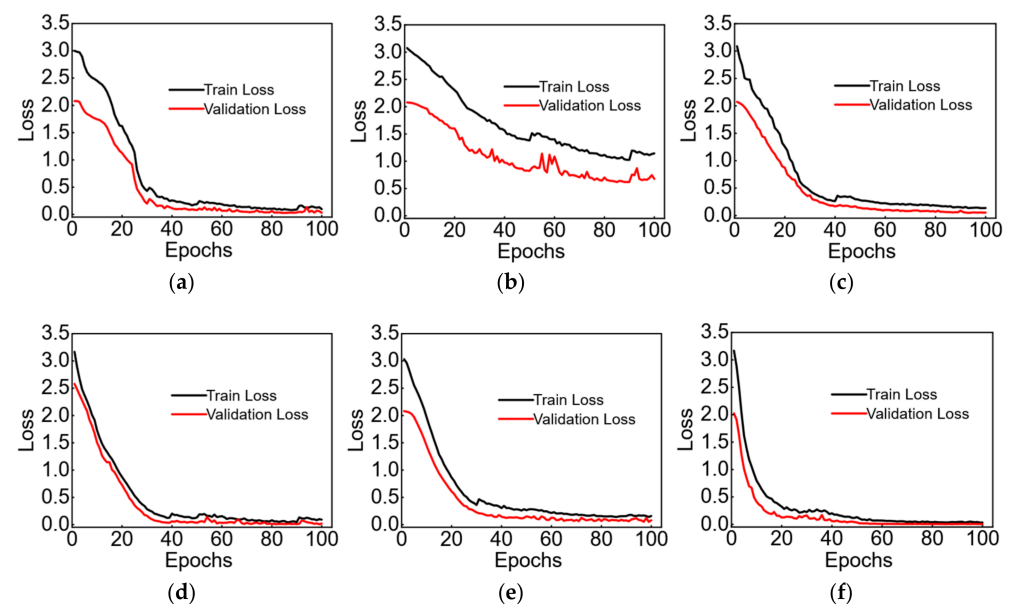


Figure 10. Loss Curve. (a) MSFEAF-Net-V; (b) MSFEAF-NetC; (c) ResNet; (d) FEMFN; (e) MSF-DFormer; (f) MSFEAF-Net.

As shown in Figure 11, MSFEAF-Net achieves the highest ultimate testing accuracy of 99.88%, which indicates its superior proficiency in mining the complementary information between vibration and current signals, thereby yielding the optimal diagnostic performance. Compared with related studies under the same experimental conditions, FEMFN and MSF-DFormer achieve test accuracies of 98.11% and 97.92%, respectively, whereas MSFEAF-Net achieves 99.88%, representing improvements of 1.77 and 1.96 percentage points, respectively. Compared with existing methods, the proposed approach enables more effective exploitation of the complementary information contained in vibration and current signals through early cross-modal interaction, hierarchical dynamic weighting, and multi-level feature aggregation. These findings firmly demonstrate that the rational cross-modal interaction and adaptive fusion strategy embedded in MSFEAF-Net can effectively bolster both the accuracy and stability of rolling bearing fault diagnosis.

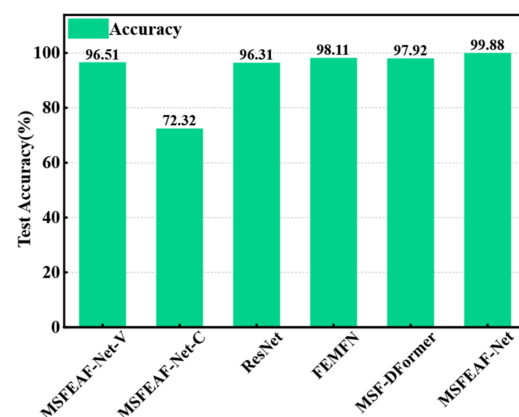


Figure 11. Bar Chart of Test Accuracy.

4.2.2. Comparative Analysis Under Diverse Signal-to-Noise Ratios

We evaluate network robustness across harsh noise scenarios. For this purpose, different levels of white Gaussian noise are mixed into the original PU and KAIST signals. This generates diverse noisy benchmarks for anti-noise validation. The SNR levels are specifically designated as -2 dB, -4 dB, -6 dB and -8 dB to simulate real-world industrial scenarios where sensor signals are heavily contaminated by background noise. All comparative models are executed repeatedly for 10 trials under identical experimental conditions, with the average accuracy captured as the final metric. As tabulated in Table 9 and Figure 12, with the progressive decrease in SNR, the classification accuracies of all models exhibit varying degrees of degradation, which indicates that intense noise significantly distorts fault feature expressions and exacerbates the difficulty of classification. Nevertheless, compared with other benchmarking methods, the proposed MSFEAF-Net consistently maintains the highest or near-highest diagnostic precision across all noise levels, demonstrating superior anti-noise robustness. Particularly under low SNR conditions, the uni-modal baseline models, MSFEAF-Net-V and MSFEAF-Net-C, undergo more severe performance drop-offs, demonstrating that a single modality is highly susceptible to feature distortion in intense noise backgrounds. Meanwhile, although the straightforward concatenation-based ResNet integrates dual-modality information, its performance enhancement remains limited due to the lack of dynamic modeling regarding the shifting contributions of individual modalities under noisy conditions. In contrast, FEMFN and MSF-DFormer achieve a certain edge in noise environments by virtue of leveraging attention mechanisms and Transformer structures to augment feature interaction capability; however, they still exhibit performance fluctuations under extremely low SNR scenarios. The outstanding performance delivered by the MSFEAF-Net is primarily attributed to the following three cornerstones: First, the

dual-branch multi-scale residual block architecture facilitates the extraction of more robust localized and global fault characteristics from diverse receptive fields. Second, the cross-modal Transformer encoder is capable of continuously mining the complementary relationship between vibration and current modulations even amid heavy noise interference. Third, the dynamic feature fusion module generates modality weights based on the joint features of the current sample at different feature levels, helping to mitigate the adverse effects of disturbances affecting a single modality on the fused features and thereby enhancing the representation of critical fault information.

Table 9. Accuracy Comparison of Different Methods Under Various SNR Levels.

Method	PU Dataset (%)				KAIST Dataset (%)			
	−2	−4	−6	−8	−2	−4	−6	−8
MSFEAF-Net-V	95.67	93.21	88.54	83.09	96.12	91.43	87.76	82.35
MSFEAF-Net-C	69.15	65.83	59.27	52.68	67.92	63.44	57.81	51.09
ResNet	94.02	89.46	84.77	78.93	94.88	90.12	85.33	79.67
FEMFN	97.88	94.15	91.66	84.21	97.73	93.99	91.43	83.02
MSF-DFormer	98.95	96.33	93.01	88.56	98.24	95.78	92.13	87.89
MSFEAF-Net	99.12	97.56	95.07	91.78	98.87	97.01	94.55	91.32

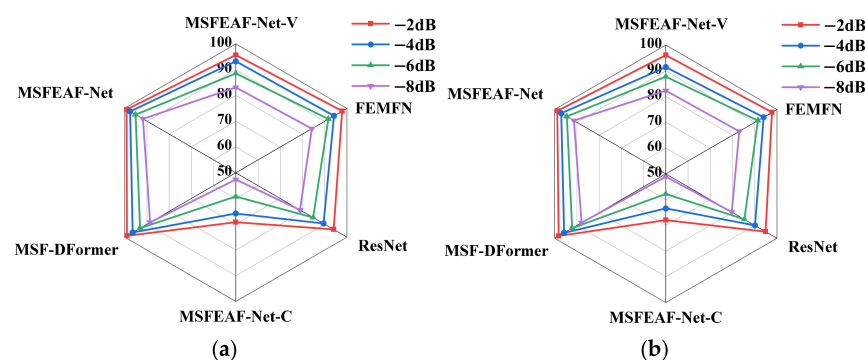


Figure 12. Accuracy Radar Chart. (a) PU Dataset (%); (b) KAIST Dataset (%).

In summary, the proposed MSFEAF-Net sustains a high diagnostic accuracy under harsh noise environments, firmly validating that this methodology possesses excellent anti-interference capacity and massive potential for real-world engineering applications.

4.2.3. Comparative Analysis Under Different Sample Sizes

To evaluate the applicability of the proposed model in data-scarce scenarios, comparative experiments were conducted across varying sample sizes using both the PU and KAIST datasets. By modulating the volume of available training samples, multiple experimental scenarios ranging from sample-abundant to sample-scarce were constructed to investigate the feature learning efficacy and generalization performance of each model under small-sample regimes.

As illustrated in Figure 13, the diagnostic performance of all methods exhibits an overall downward trend as the training sample size decreases. This degradation stems from the fact that insufficient data restricts the comprehensive learning of fault feature distributions while exacerbating the risk of overfitting. Nevertheless, compared with benchmarking models, MSFEAF-Net consistently maintains stable classification performance across different sample scales and retains a significant accuracy advantage even under small-sample conditions. This demonstrates that the proposed model possesses superior feature representation capabilities and high sample utilization efficiency.

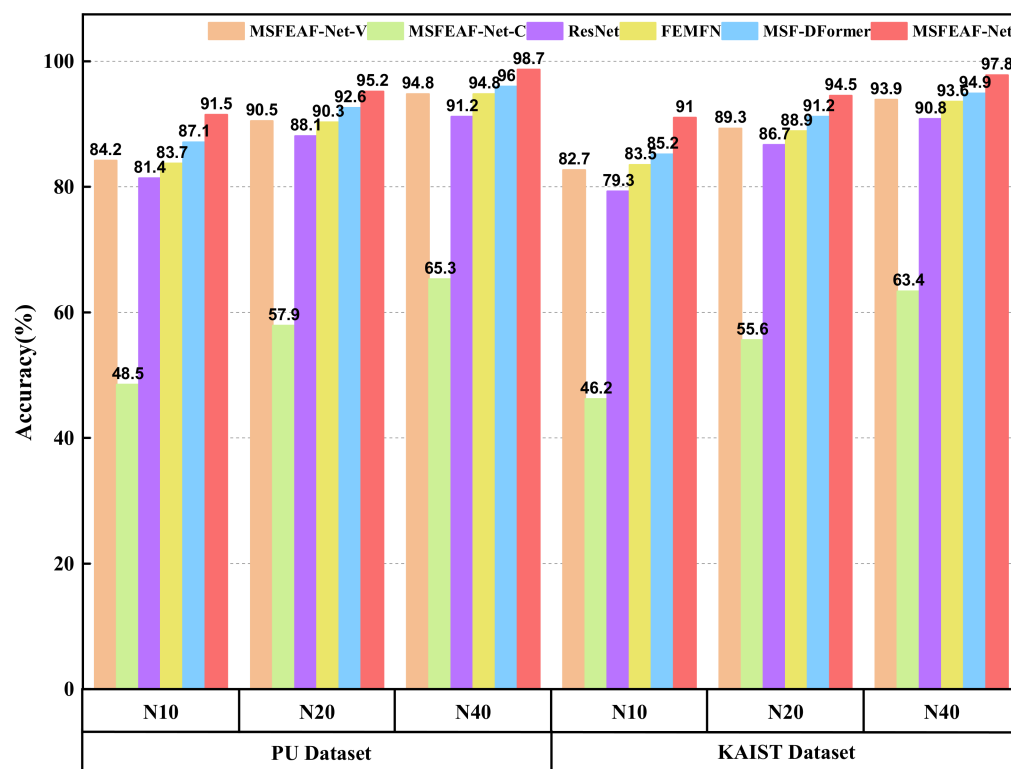


Figure 13. Diagnostic accuracies of various methods under different sample size scales.

Specifically, uni-modal models suffer more pronounced performance drops as samples decrease, indicating that models relying solely on single-sensor information are more sensitive to shifts in sample distribution. Although ResNet utilizes dual-modality information, its straightforward concatenation strategy proves insufficient for mining cross-modal complementarity under limited data constraints. While FEMFN and MSF-DFormer outperform traditional fusion models in small-sample scenarios, their relatively complex architectures still exhibit a certain degree of dependency on the scale of the training set. In contrast, through the synergetic effects of multi-scale residual feature extraction, cross-modal correlation modeling, and the dynamic fusion mechanism, MSFEAF-Net effectively preserves critical fault information and suppresses redundant features under data constraints, thereby yielding more stable and reliable diagnostic results.

Consequently, the varying sample-size experiments further validate that MSFEAF-Net not only excels under conventional data scales but also maintains robust generalization capabilities in real-world scenarios where training samples are restricted.

4.2.4. Comparative Analysis of Different Feature Fusion Methods

To verify the effectiveness of hierarchical input-dependent weighting, comparative experiments were conducted on the PU dataset using direct concatenation (Concatenation) and conventional attention-based fusion (Attention Fusion), while keeping the backbone network and training settings unchanged. Concatenation connects the features of the two modalities along the channel dimension at the end of the network, whereas Attention Fusion generates modality weights at a single feature level. In contrast, the DFFM generates input-dependent modality weights separately at the 128-, 256-, and 512-channel levels and aggregates the fusion results from all three levels. The accuracies of the three feature fusion methods under different signal-to-noise ratios (SNRs) are presented in Table 10. As the SNR decreases from -2 dB to -8 dB, the accuracies of all three fusion methods decline. At -2 dB, the DFFM achieves an accuracy of 99.12%, exceeding Concatenation and Attention Fusion

by 3.70 and 2.04 percentage points, respectively. At -8 dB, it achieves an accuracy of 91.78%, representing improvements of 12.35 and 5.53 percentage points, respectively. These results indicate that, compared with feature concatenation at the end of the network and single-level attention-based fusion, modality weighting conditioned on both the input sample and feature level can more effectively integrate the complementary information contained in vibration and current signals while maintaining favorable diagnostic performance under the tested noisy conditions.

Table 10. Performance Comparison of Different Feature Fusion Methods Under Various SNR Levels.

SNR (dB)	Fusion Method	Accuracy (%)
-2 dB	Concatenation	95.42%
	Attention Fusion	97.08%
	DFFM	99.12%
-4 dB	Concatenation	91.65%
	Attention Fusion	94.82%
	DFFM	97.56%
-6 dB	Concatenation	86.12%
	Attention Fusion	91.34%
	DFFM	95.07%
-8 dB	Concatenation	79.43%
	Attention Fusion	86.25%
	DFFM	91.78%

To further compare the classification performance of the three fusion methods under noisy conditions, Figure 14 presents their confusion matrices at an SNR of -2 dB. Direct concatenation merely stacks the features from the two modalities without dynamically adjusting their relative contributions; consequently, relatively evident misclassifications remain between certain fault categories. Attention Fusion assigns weights according to the feature responses, resulting in a greater concentration of elements along the main diagonal than that achieved by direct concatenation. However, because fusion is performed primarily at a single feature level, its ability to exploit differences in modal contributions across different network depths remains limited. In contrast, the classification results of the DFFM are more concentrated along the main diagonal, with fewer misclassifications in the off-diagonal regions, indicating that the fused features exhibit better class separability. These results demonstrate that generating modality weights based on the current sample at different feature levels facilitates the exploitation of complementary information between vibration and current features.

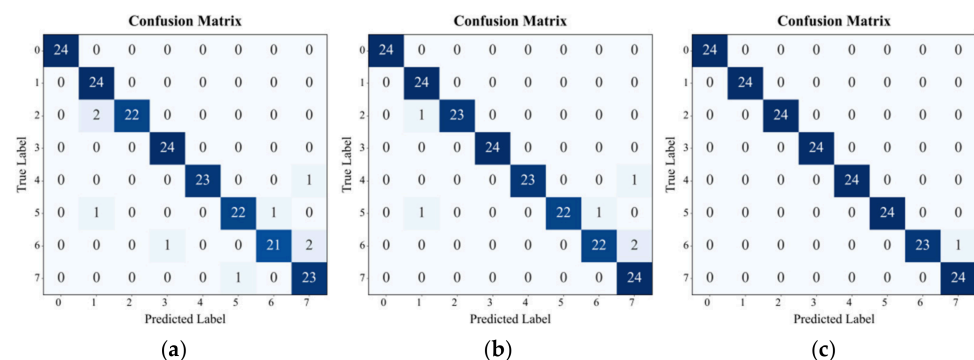


Figure 14. Confusion matrices of different feature fusion methods under the -2 dB noise environment. (a) Concatenation; (b) Attention Fusion; (c) DFFM.

4.3. Ablation Study

To comprehensively evaluate the individual contributions of the core architectural modules within the proposed multi-scale feature enhancement and adaptive fusion network (MSFEAF-Net), a systematic ablation study was conducted on the PU dataset based on the experimental schemes summarized in Table 11. The full MSFEAF-Net model (Case 1), incorporating all designed components, serves as the baseline to quantitatively measure the performance gains yielded by each individual module. Under identical experimental settings, each of the five architectural configurations was executed independently for 5 trials, and the resulting multi-run diagnostic accuracies are graphically displayed in Figure 15.

Table 11. Architectural Configuration Matrix for the Ablation Study.

Experiment ID	Model Modification	Purpose
Case 1	Complete MSFEAF-Net	Serve as the baseline for evaluating the diagnostic performance of the complete model
Case 2	Use a Single-Scale Convolution Kernel	Verify the effectiveness of multi-scale feature extraction
Case 3	Remove the Feature Enhancement Module	Verify the effectiveness of early cross-modal feature enhancement
Case 4	Replace adaptive weights with fixed weights	Verify the effectiveness of hierarchical dynamic weighting
Case 5	Remove the entire Adaptive Feature Selection Module	Verify the effectiveness of additional feature refinement before final classification

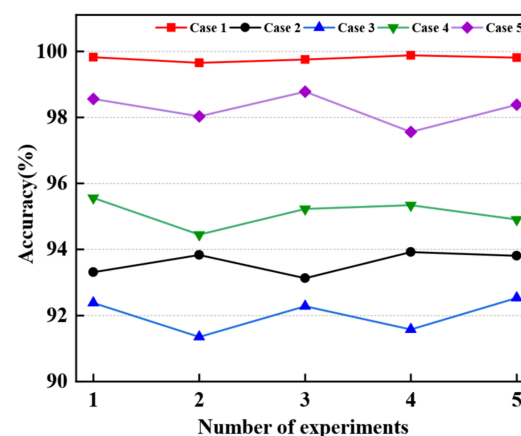


Figure 15. Diagnostic accuracy curves of ablation study.

Synthesizing the data from the ablation study, the omission of any constituent component from the full model results in a discernible degradation in diagnostic accuracy, which robustly substantiates the necessity of each integrated module. Notably, a deep architectural coupling is observed between the multi-scale residual block (MSRB) and the feature enhancement module (FEM). When the multi-scale feature extraction is removed (Case 2) and the network degenerates into a single-scale architecture, the average accuracy plummets from the baseline 99.78% to 93.88%. This implies that a fixed receptive field fails to adapt to the heterogeneous frequency-band characteristics of bearing signals; furthermore, the insufficient resolution of the input representations directly restricts the FEM

from aligning cross-modal feature correlations. Conversely, when the feature enhancement module is severed (Case 3), the average accuracy drops further to 92.02%. This indicates that the absence of the global long-range temporal modeling provided by the FEM leads to insufficient cross-modal feature correlations and prevents deep nonlinear interactions between modalities, demonstrating that both modules contribute to performance improvement. In addition, when the DFFM is replaced with fixed equal-weight fusion (Case 4), the model accuracy decreases to 95.10%. This result indicates that generating modality weights based on the current sample at different feature levels helps improve the utilization of complementary information from vibration and current signals. Lastly, removing the adaptive feature selection module (Case 5) reduces the diagnostic average accuracy to 98.26%, which underscores the indispensable value of the AFSM in filtering out information redundancy and executing the refined screening of critical fault components to fortify the stability of the final classification outputs.

In summary, the full model (Case 1) consistently sustains an exceptionally high diagnostic accuracy (with a mean value of 99.78%) and exhibits the most stable convergence curve across all epochs, which is heavily credited to the synergetic operation of these functional modules along the complete diagnostic pipeline. This thoroughly validates the methodological rationality, soundness, and superiority of the network architecture engineered in the MSFEAF-Net.

4.4. Feature Visualization Analysis

To analyze changes in the class representations learned by MSFEAF-Net during feature propagation, equal numbers of samples from eight bearing condition categories were selected from the PU dataset test set. t-SNE was then employed to sequentially visualize the raw inputs, the outputs of each MSRB in the two branches, the DFFM outputs at different feature levels, the MSFP output, and the final classification output in two dimensions. The same samples and class colors were used at all stages, and the results are presented in Figure 16.

In the raw input space, samples from different categories are intermixed without clear class boundaries. After MSRB1 and the FEM, some fault categories in the vibration branch begin to form clusters. As the network deepens, most categories become progressively separated, although local overlaps remain between minor-damage conditions and faults occurring at the same location. In contrast, the features extracted at different layers of the current branch exhibit a higher degree of mixing. This is because vibration signals directly reflect the transient impulses generated by localized damage, whereas fault responses in current signals are formed indirectly through electromechanical coupling. These responses have relatively low amplitudes and are easily affected by the fundamental frequency, harmonics, and noise. After DFFM-128, DFFM-256, and DFFM-512, the within-class distributions of the fused features become progressively more compact, while the inter-class boundaries become generally clearer, indicating that hierarchical dynamic fusion can exploit the complementary information contained in vibration and current signals. After the MSFP further aggregates the fused features from different network depths, categories that overlap at the shallow stages become more clearly separated. The final classification features form eight relatively compact clusters with clear boundaries, demonstrating that the constructed processes of early cross-modal interaction, hierarchical dynamic fusion, and cross-level feature aggregation progressively improve feature class separability.

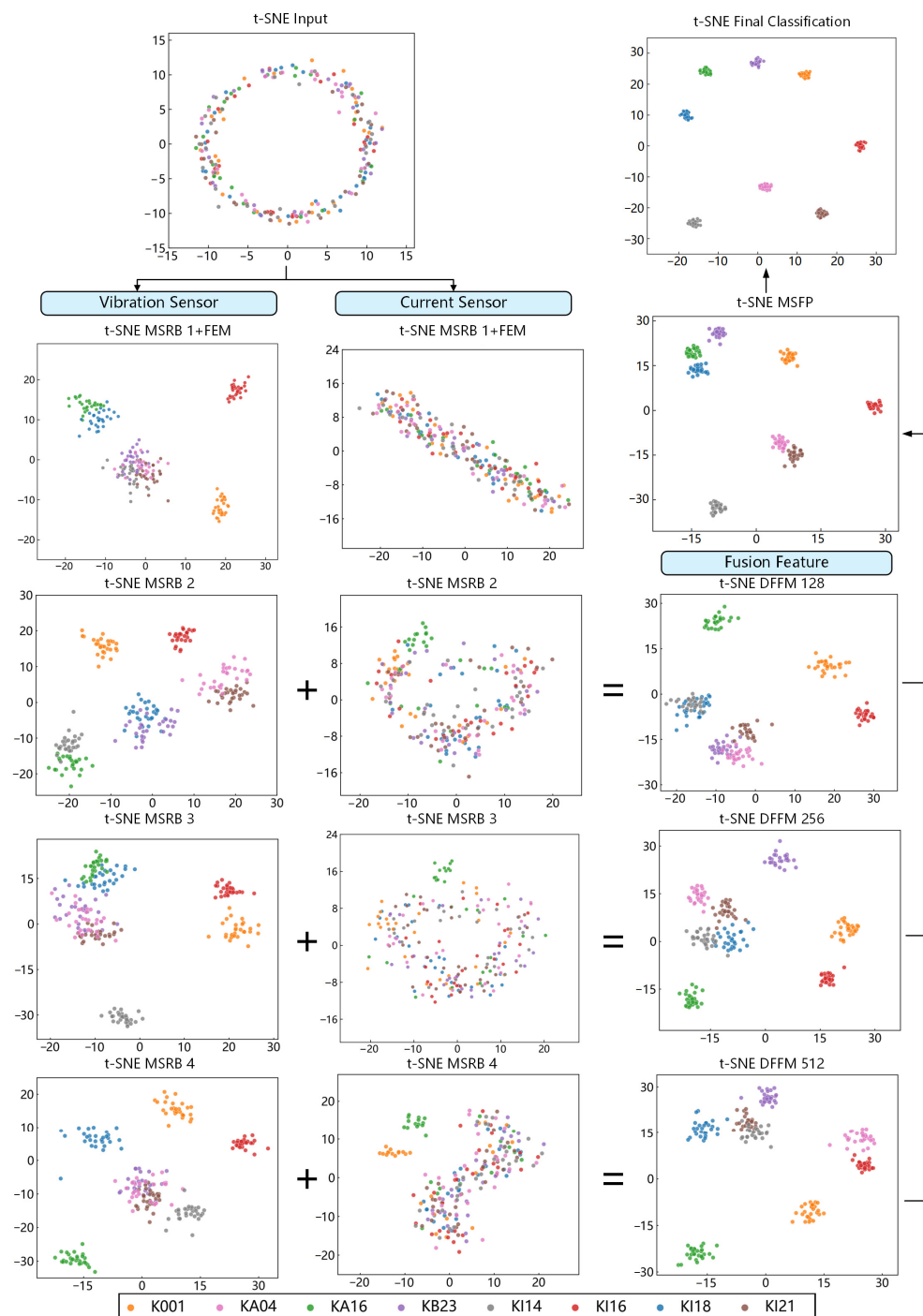


Figure 16. t-SNE visualization of feature distributions at different stages of MSFEAF-Net.

4.5. Cross-Condition Generalization Performance Evaluation

To evaluate the diagnostic performance of the proposed MSFEAF-Net under changing operating conditions, closed-set cross-condition diagnosis tasks were designed based on the four operating conditions of the PU dataset, as listed in Table 3. All four operating conditions contain the same bearing states, and the label indices of each category remain consistent across the different conditions. Therefore, the source and target operating conditions share the same fault-category space, and the experiments constitute a closed-set cross-condition generalization evaluation. For each task, only samples from the source operating condition were used for model training and parameter selection. The source-condition data were divided into training and validation sets at a ratio of 8:2, while

all samples from the target operating condition were used exclusively for final testing. Neither the target-condition samples nor their labels were involved in model training, hyperparameter selection, or model checkpointing. The data normalization parameters were calculated from the source-condition training set and directly applied to the target-condition test set. The experimental tasks were named according to the ($S \rightarrow T$) convention, where (S) denotes the source domain providing the training data and (T) denotes the target domain providing the test data. The specific task settings are presented in Table 12.

Table 12. Settings of the Cross-Condition Generalization Tasks.

Task ID	Transfer Path	Varying Parameter
Task A	1 $\rightarrow$ 2	Rotational speed variation
Task B	2 $\rightarrow$ 3	Torque variation
Task C	3 $\rightarrow$ 4	Radial load variation
Task D	1 $\rightarrow$ 4	Multi-parameter hybrid variation

MSFEAF-Net was compared with the baseline models selected in Section 4.2. The average diagnostic accuracies of all models under the cross-condition tasks are presented in Table 13.

Table 13. Accuracies of Different Methods for Each Cross-Condition Generalization Task (%).

Method	Task A	Task B	Task C	Task D
MSFEAF-Net-V	91.24	88.56	90.32	81.02
MSFEAF-Net-C	62.45	58.91	60.15	50.34
ResNet	89.56	85.34	87.12	78.56
FEMFN	93.88	91.12	92.45	84.21
MSF-DFormer	95.02	93.45	94.18	87.45
MSFEAF-Net	97.85	96.23	96.88	92.45

Experimental results indicate that MSFEAF-Net exhibits outstanding cross-condition diagnostic performance. Specifically, under the extreme operating scenario characterized by multi-parameter hybrid variations (Task D), the uni-modal baseline models (MSFEAF-Net-V and MSFEAF-Net-C) undergo severe performance degradation due to their high susceptibility to operating condition fluctuations, with the single current modality exhibiting particularly weak capability. Conversely, leveraging multi-source information complementarity, the proposed MSFEAF-Net significantly bolsters the robustness under varying operating conditions, sustaining a classification accuracy of over 92%. Compared with ResNet, which employs fixed-scale convolutional kernels, MSFEAF-Net can flexibly capture the characteristic frequency shifts induced by rotational speed variations (Task A) through the multi-scale residual blocks (MSRBs), thereby extracting more discriminative fault features. Furthermore, owing to the adaptive adjustment of modality weights at different feature levels by the dynamic feature fusion module, MSFEAF-Net outperforms MSF-DFormer in complex cross-condition tasks and effectively suppresses interference from non-informative features caused by complex operating conditions.

4.6. Computational Complexity Analysis

To evaluate the computational costs and engineering application potential of different models, their parameter counts, floating-point operations (FLOPs), and single-sample inference times were measured under the software and hardware environment described in Section 4.1.2. An input length of 2048 and a batch size of 1 were used for all models. The inference time is reported as the mean and standard deviation obtained from multiple measurements. The detailed results are presented in Table 14.

Table 14. Computational complexity and diagnostic performance of different models.

Model	Params/M	FLOPs/G	Inference Time/ms	Accuracy/%
ResNet	19.83	0.9212	6.4852 (± 0.0635)	96.31
FEMFN	21.32	0.9221	6.5752 (± 0.1316)	98.11
MSF-DFormer	21.34	0.9203	6.9823 (± 0.2835)	97.92
MSFEAF-Net	21.37	0.9216	6.9452 (± 0.1236)	99.88

As shown in Table 14, ResNet has the lowest parameter count and inference time, but its diagnostic accuracy is only 96.31%. MSFEAF-Net contains 21.37 M parameters, requires 0.9216 G FLOPs, and achieves a single-sample inference time of 6.9452 ms and a test accuracy of 99.88%. Compared with ResNet, MSFEAF-Net increases the parameter count by 1.54 M and the inference time by 0.4600 ms, while improving accuracy by 3.57 percentage points. Compared with FEMFN and MSF-DFormer, MSFEAF-Net increases the parameter count by only 0.05 M and 0.03 M, respectively, while requiring a comparable number of FLOPs and improving accuracy by 1.77 and 1.96 percentage points, respectively. In addition, the inference time of MSFEAF-Net is slightly shorter than that of MSF-DFormer. These results demonstrate that the proposed model achieves higher diagnostic accuracy with only a modest increase in computational cost, thereby providing a favorable balance between computational efficiency and diagnostic performance.

5. Conclusions and Future Work

To address the limitations of single end-stage fusion of vibration and current features, fixed modal contributions, and insufficient utilization of intermediate-layer information, this study proposes a multi-scale feature enhancement and adaptive fusion network (MSFEAF-Net). Through the coordinated operation of the dual-branch multi-scale residual feature extraction framework and the feature enhancement module, the network performs in-depth extraction of temporal and complementary features from vibration and current signals. The dynamic fusion strategy then generates sample-dependent modality weights at different feature levels, while the multi-scale feature pyramid aggregates the fusion results from different network depths. Finally, the adaptive feature selection module refines critical feature information for fault classification. Experimental results demonstrate that MSFEAF-Net achieves extremely high diagnostic accuracy under conventional conditions and exhibits excellent robustness and generalization performance under strong noise interference, small-sample constraints, and complex Cross-Condition scenarios. Nevertheless, this study has certain limitations. The current experiments are primarily based on publicly available test-rig datasets, and the performance of the proposed method under compound faults and complex disturbances in real-world operating environments has not yet been fully validated. Future work will validate the proposed method using real-world operational data and investigate unsupervised domain adaptation to enhance its practical deployability and adaptability to unseen operating conditions.

Author Contributions: Conceptualization, W.Z. and C.Z.; methodology, W.Z. and F.J.; software, W.Z. and Z.D.; validation, F.J.; investigation, Z.D.; resources, C.Z.; data curation, T.L. and J.H.; writing—original draft preparation, W.Z.; writing—review and editing, Z.D., T.L., F.J. and C.Z.; supervision, C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Science and Technology Program of Inner Mongolia Autonomous Region under Grant No. 2024SKYPT0032 and the Basic Scientific Research Fund of Directly Governed Universities of Autonomous Regions No. 2024QNJS007 and Achievement Transformation Program Project under Grant No. 2025KJHZ0033.

Data Availability Statement: The data presented in this study are openly available in the Paderborn University Bearing Data Center at <https://mb.uni-paderborn.de/en/kat/research/bearing-datacenter/data-sets-and-download>, accessed on 24 September 2026, reference number [37]; and in the KAIST dataset at <https://doi.org/10.17632/ztmf3m7h5x.6>, reference number [38].

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

References

1. Kestel, K.; Chesterman, X.; Zappalá, D.; Watson, S.; Li, M.; Hart, E.; Carroll, J.; Vidal, Y.; Nejad, A.R.; Sheng, S.; et al. Condition monitoring of wind turbine drivetrains: State-of-the-art technologies, recent trends, and future outlook. *Wind Energy Sci.* **2026**, *11*, 2103–2155. [\[CrossRef\]](#)
2. Sun, Y.; Wang, J.; Wang, X. Fault diagnosis of mechanical equipment in high energy consumption industries in China: A review. *Mech. Syst. Signal Process.* **2023**, *186*, 109833. [\[CrossRef\]](#)
3. Amirat, Y.; Choqueuse, V.; Benbouzid, M.E.H. EEMD-based wind turbine bearing failure detection using the generator stator current homopolar component. *Mech. Syst. Signal Process.* **2013**, *41*, 667–678. [\[CrossRef\]](#)
4. Liu, Z.; Zhang, L. A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings. *Measurement* **2020**, *149*, 107002. [\[CrossRef\]](#)
5. Liu, T.; Lu, J.; Li, J.; Huang, J.; Zhang, C.; Dai, Y.; Zhang, F. Application of tensor singular spectrum decomposition based on Morlet wavelet ridges in weak fault diagnosis of bearings. *Nondestruct. Test. Eval.* **2025**, 1–30. [\[CrossRef\]](#)
6. Yang, G.; Xu, W.Y.; Deng, Q.; Wei, Y.; Li, F. A review of rolling bearing compound fault diagnosis based on vibration signals. *J. Xihua Univ. (Nat. Sci. Ed.)* **2024**, *43*, 48–69.
7. Chen, K.; An, K.; Wang, X.X.; Song, J.C.; Lu, S.L. Motor speed estimation and bearing fault diagnosis based on deep fusion of vibration and current signals. *Bearing* **2023**, *2*, 54–60.
8. Ferrando Chacon, J.L.; Kappatos, V.; Balachandran, W.; Gan, T.-H. A novel approach for incipient defect detection in rolling bearings using acoustic emission technique. *Appl. Acoust.* **2015**, *89*, 88–100. [\[CrossRef\]](#)
9. Jiang, X.; Song, Q.; Wang, H.; Du, G.; Guo, J.; Shen, C.; Zhu, Z. Central frequency mode decomposition and its applications to the fault diagnosis of rotating machines. *Mech. Mach. Theory* **2022**, *174*, 104919. [\[CrossRef\]](#)
10. Sheng, Z.; Xu, Y.; Zhang, K. Applications in bearing fault diagnosis of an improved Kurtogram algorithm based on flexible frequency slice wavelet transform filter bank. *Measurement* **2021**, *174*, 108975. [\[CrossRef\]](#)
11. He, D.; Liu, C.; Jin, Z.; Ma, R.; Chen, Y.; Shan, S. Fault diagnosis of flywheel bearing based on parameter optimization variational mode decomposition energy entropy and deep learning. *Energy* **2022**, *239*, 122108. [\[CrossRef\]](#)
12. Dong, Z.; Zhao, D.; Cui, L. An intelligent bearing fault diagnosis framework: One-dimensional improved self-attention-enhanced CNN and empirical wavelet transform. *Nonlinear Dyn.* **2024**, *112*, 6439–6459. [\[CrossRef\]](#)
13. Yu, P.; Zhang, J.; Zhang, B.; Cao, J.; Peng, Y. Research on small sample rolling bearing fault diagnosis method based on mixed signal processing technology. *Symmetry* **2024**, *16*, 1178. [\[CrossRef\]](#)
14. 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. [\[CrossRef\]](#)
15. Sze, V.; Chen, Y.H.; Yang, T.J.; Emer, J.S. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* **2017**, *105*, 2295–2329. [\[CrossRef\]](#)
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Cheng, J.; Dong, L.; Lapata, M. Long short-term memory-networks for machine reading. *arXiv* **2016**, arXiv:1601.06733.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, L. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
19. Tang, Y.; Zhang, C.; Wu, J.; Xie, Y.; Shen, W.; Wu, J. Deep learning-based bearing fault diagnosis using a trusted multiscale quadratic attention-embedded convolutional neural network. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 3513215. [\[CrossRef\]](#)
20. Liao, W.; Fu, W.; Yang, K.; Tan, C.; Huang, Y. Multi-scale residual neural network with enhanced gated recurrent unit for fault diagnosis of rolling bearing. *Meas. Sci. Technol.* **2024**, *35*, 056114. [\[CrossRef\]](#)
21. Chen, X.; Zhang, B.; Gao, D. Bearing fault diagnosis base on multi-scale CNN and LSTM model. *J. Intell. Manuf.* **2021**, *32*, 971–987. [\[CrossRef\]](#)
22. Guo, Y.; Zhou, J.; Dong, Z.; She, H.; Xu, W. Research on bearing fault diagnosis based on novel MRSVD-CWT and improved CNN-LSTM. *Meas. Sci. Technol.* **2024**, *35*, 095003. [\[CrossRef\]](#)
23. Song, B.; Liu, Y.; Fang, J.; Liu, W.; Zhong, M.; Liu, X. An optimized CNN-BiLSTM network for bearing fault diagnosis under multiple working conditions with limited training samples. *Neurocomputing* **2024**, *574*, 127284. [\[CrossRef\]](#)

24. Chen, H.; Wei, J.; Huang, H.; Wen, L.; Yuan, Y.; Wu, J. Novel imbalanced fault diagnosis method based on generative adversarial networks with balancing serial CNN and Transformer (BCTGAN). *Expert Syst. With Appl.* **2024**, *258*, 125171. [[CrossRef](#)]
25. Zhao, L.; Yang, S.; Liu, Y. Dimension reduction graph-based sparse subspace clustering for intelligent fault identification of rolling element bearings. *Int. J. Mech. Syst. Dyn.* **2021**, *1*, 207–219. [[CrossRef](#)]
26. Spirto, M.; Melluso, F.; Nicoletta, A.; Malfi, P.; Cosenza, C.; Savino, S.; Niola, V. Rolling bearing fault detection using symmetrized dot pattern indices and feedforward neural network. *Struct. Health Monit.* **2026**, 1–21. [[CrossRef](#)]
27. Chen, Y.Q.; Zhang, R.D.; Gao, F.R. Multi-scale residual attention networks with spatio-temporal attention co-optimization for industrial fault diagnosis. *Comput. Chem. Eng.* **2026**, *205*, 109482. [[CrossRef](#)]
28. Yang, J.J.; Han, H.F.; Dong, X.; Wang, G.; Zhang, S. Bearing fault diagnosis grounded in the multi-modal fusion and attention mechanism. *Appl. Sci.* **2025**, *15*, 1531. [[CrossRef](#)]
29. Cui, Y.; Dong, Z.; Gao, W.; Chang, C.; Wang, J. A rolling bearing fault diagnosis framework based on multi-modal feature fusion and marginal-conditional alignment. *Meas. Sci. Technol.* **2026**, *37*, 036112. [[CrossRef](#)]
30. Liu, T.; Liu, C.; Huang, J.; Zhang, C.; Jin, F.; Zhang, K.; Du, Z.; Guo, X.; Zhang, X. Fault diagnosis of wind turbine rolling bearings with two-stage fusion of multi-channel adaptation and multi-domain correlation. *Meas. Sci. Technol.* **2026**, *37*, 105102. [[CrossRef](#)]
31. Randall, R.B.; Antoni, J. Rolling element bearing diagnostics—A tutorial. *Mech. Syst. Signal Process.* **2011**, *25*, 485–520. [[CrossRef](#)]
32. El Bouchikhi, E.H.; Choqueuse, V.; Benbouzid, M.E.H. Current frequency spectral subtraction and its contribution to induction machines' bearings condition monitoring. *IEEE Trans. Energy Convers.* **2013**, *28*, 135–144. [[CrossRef](#)]
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
34. Zhou, J.; Yang, X.; Li, J. Deep residual network combined with transfer learning based fault diagnosis for rolling bearing. *Appl. Sci.* **2022**, *12*, 7810. [[CrossRef](#)]
35. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
36. Shehzadi, T.; Hashmi, K.A.; Liwicki, M.; Afzal, M.Z. Object detection with transformers: A review. *Sensors* **2025**, *25*, 6025. [[CrossRef](#)]
37. Lessmeier, C.; Kimotho, J.K.; Zimmer, D.; Sextro, W. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In Proceedings of the 3rd European Conference of the Prognostics and Health Management Society, Bilbao, Spain, 5–8 July 2016; pp. 152–156.
38. Jung, W.; Kim, S.H.; Yun, S.H.; Bae, J.; Park, Y.-H. Vibration, acoustic, temperature, and motor current dataset of rotating machine under varying operating conditions for fault diagnosis. *Data Brief.* **2023**, *48*, 109049. [[CrossRef](#)]
39. Chen, H.X. *Research on Bearing Fault Diagnosis Based on Multimodal Transfer Learning*; Yanshan University: Qinhuangdao, China, 2024.
40. Jiang, L.; Ding, C.L.; Guo, Y.X.; Hua, G.; Zhang, Y. MSF-DFormer: A multisensor multi-scale fusion network with deformable Transformer for fault diagnosis under complex working conditions with limited samples. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 3562418. [[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.