machines

MDPI

Article

Fault Diagnosis of Gearbox Bearings Under Extreme Class Imbalance Based on Multi-Resolution Windows and Density-Aware Cross-Modal Fusion

Hao Wei [1], Minghui Liang [2], Gang Lan [3,4,*] and Xueyi Li [5]

[1] College of Automotive Engineering, Guangxi Transport Vocational and Technical College, Nanning 530023, China; 123560295@gxjtc.edu.cn
[2] Fengting River Reservoir Management Center, Department of Water Resources of Guangxi Zhuang Autonomous Region, Nanning 530232, China; mhliang@gxfthwrd.cn
[3] Guangxi Water Development Group Co., Ltd., Nanning 530023, China
[4] Guangxi Water Development Group Trading Co., Ltd., Nanning 530023, China
[5] Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China; phm@mail.tsinghua.edu.cn
* Correspondence: langang@gxfthwrd.cn

Abstract

Robust bearing fault diagnosis under variable speeds and extreme class imbalance remains challenging due to the difficulty in decoupling transient-periodic features. To address this, we propose a Dual-Branch Weighted Convolution Network with Cross-Attention Fusion (DBWC). Specifically, a dual-branch architecture with spatial density-weighted kernels is proposed to decouple high-frequency transients from low-frequency periodic trends. Subsequently, a Cross-Attention Fusion module synthesizes these heterogeneous features by using global contexts to filter local noise. Additionally, an imbalance-aware strategy integrating Focal Loss and composite augmentation is developed to mitigate model bias. Extensive experiments on MCC5-THU and HUST benchmarks demonstrate that DBWC achieves an accuracy of 91.00% and 90.12%, respectively. The proposed method outperforms state-of-the-art models by an average margin of 5%, providing a data-efficient paradigm for complex industrial monitoring.

Keywords: fault diagnosis; intelligent operation and maintenance; multi-channel fusion; bearings failure

Check for updates

Academic Editors: Ronghua Zhu, Cailiang Zhang, Chaoge Wang, Zepeng Liu and Dimitrios Giagopoulos

Received: 19 May 2026
Revised: 16 July 2026
Accepted: 21 July 2026
Published: 23 July 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) license.

1. Introduction

Wind turbine gearbox bearings are pivotal components within the mechanical drivetrain, underpinning the operational reliability of wind energy conversion systems. Due to their operation under non-stationary loads and harsh environmental conditions, these bearings are susceptible to progressive degradation, which, if undetected, can culminate in catastrophic gearbox failures, prohibitively high maintenance costs, and extended system downtime [1]. According to industrial surveys, bearing-related failures account for approximately 40–50% of all unplanned downtime in wind turbine drivetrain systems, with average corrective maintenance costs per event ranging from $15,000 to $50,000 depending on turbine size and accessibility [2]. From a tribological perspective, the wind turbine gearbox bearing constitutes a highly complex elastohydrodynamic lubrication (EHL) system. Under ideal stationary conditions, a stable lubricating film separates the rolling elements and raceways, distributing contact stresses evenly. However, under the non-stationary

loading profiles inherent to wind farm operations—characterized by unpredictable aerodynamic turbulence, sudden wind gusts, and abrupt braking events—this delicate equilibrium is severely disrupted. These dynamic external forces cause instantaneous fluctuations in radial and axial loads, leading to the frequent breakdown of the minimum EHL film thickness [3]. Consequently, the system transitions into mixed or boundary lubrication regimes, resulting in direct asperity-to-asperity contact between the metallic surfaces. This severe non-stationary loading induces localized stress concentrations, micro-sliding, and accelerated subsurface fatigue. It is precisely these localized tribological degradation mechanisms that physically generate the highly transient, high-frequency shock impulses, which are subsequently modulated by the fluctuating global rotational speeds to form the complex, coupled periodic trends observed in the vibration signals [4]. Consequently, the development of robust fault diagnosis frameworks for wind turbine gearbox bearings has become a paramount objective within the realm of prognostic and health management (PHM). Despite the significant strides made in data-driven methodologies, real-world wind power monitoring presents formidable challenges: vibration signals are frequently modulated by unpredictable non-stationary operational regimes, and fault-related data remains extremely scarce compared to health status records [5–7]. These complexities limit the direct application of conventional signal processing techniques under highly non-stationary conditions, thereby encouraging the integration of advanced deep learning architectures capable of high-precision feature extraction under extreme constraints. The dynamics of fault occurrence in wind turbine bearings are inextricably linked to the viscous properties and thermo-dynamic stability of the lubricant within the gearbox housing. Changes in operating temperatures, caused by fluctuations in wind turbine speed and load, lead to the degradation of the oil film, thereby increasing the friction coefficient and accelerating the adhesive and abrasive wear of the elements [8].

Prior literature on deep learning-based diagnosis initially focused on single-sensor Convolutional Neural Networks (CNNs) and basic multi-scale frameworks to capture features from raw vibration signals. While demonstrably effective in laboratory settings, these methods often exhibit limited efficacy in balancing feature diversity with noise robustness in wind farm environments. For instance, incorporating channel attention into multi-scale networks has been shown to diversify feature capture, yet performance degrades noticeably under the low signal-to-noise ratios (SNRs) typical of industrial gearboxes [9]. Similarly, although attention-guided multi-branch structures have enhanced classification accuracy, they frequently falter when confronted with the extreme class imbalance inherent in real-world maintenance logs [10–12]. Early compensation techniques, such as weighted cross-entropy and oversampling, have attempted to mitigate this imbalance but fail to maintain diagnostic stability when fault samples constitute only a minute fraction of the total dataset [13]. Furthermore, recent advancements in Graph Neural Networks (GNNs) [14] and Graph Transformers [15] have expanded feature representation capabilities, albeit at the cost of significantly increased computational complexity.

The period from 2024 to 2026 has witnessed a transition toward more sophisticated architectures specifically targeting multi-sensor fusion and temporal dynamics. To capture the heterogeneous fault signatures of wind turbine components, researchers have explored dual-branch designs, such as SENet-SwinT [16] and HawkFish [17] optimized networks, to integrate multi-sensor information. Concurrently, holistic fusion strategies—ranging from physics-informed fusion to multi-domain feature tensor fusion—have been investigated to enhance class separability [18–20]. Regarding temporal dependencies, hybrid models combining WDCNN with BiLSTM [21] or multi-scale attention-based XLSTM [22] have been developed to track dynamic feature interactions. Additionally, novel frameworks like VFQB [23] and Cross-Attention Networks [24] have further refined feature representation,

while meta-learning and ensemble modeling have boosted minority class recognition in few-shot scenarios [25,26]. Meanwhile, complementary research has explored physics-based diagnostic modeling [27], domain adaptation techniques under variable operating conditions [28], and CNN-based feature learning directly from raw signals [29]. Despite these advancements, maintaining generalization under severe domain shifts and variable operating regimes remains a persistent hurdle [30]. Despite these advancements, maintaining generalization under severe domain shifts and variable operating regimes remains a persistent hurdle [30].

Notwithstanding this flourishing research landscape, three fundamental bottlenecks continue to impede the deployment of reliable diagnostic systems for wind turbine gearboxes. First, existing multi-sensor methods often rely on shallow feature fusion strategies, such as elementary concatenation, which fail to facilitate the deep, non-linear cross-modal interactions required to decipher complex fault couplings in gearboxes [31–33]. Second, current multi-scale extractors typically process signal windows indiscriminately, neglecting the physical reality that transient impulses and periodic degradation trends require specialized, decoupled receptive fields [34–36]. Third, existing frameworks exhibit extreme fragility under class imbalance: in scenarios where fault samples represent less than 5% of the data, generative and attention-based approaches frequently suffer from mode collapse or overfit to the dominant healthy classes, leading to unsatisfactory F1 scores in noisy environments [37–39].

To bridge these gaps, this study proposes a Dual-Branch Weighted Convolutional Network with Cross-Attention Fusion (DBWC), specifically engineered for wind turbine gearbox bearing diagnosis under extreme class imbalance. The primary contributions of this work are summarized as follows:

(1) We design a specialized bifurcation structure that treats temporal scales differentially. One branch employs weighted short-time window convolutions to isolate high-frequency transient impulses, while the parallel branch utilizes dilated long-time window convolutions to capture global periodic patterns, ensuring a comprehensive representation of bearing fault signatures.

(2) Departing from conventional shallow fusion, we introduce a Cross-Attention module that enforces deep, non-linear interaction between sensor modalities. This mechanism dynamically weighs feature importance, effectively amplifying fault-sensitive characteristics while suppressing the uncorrelated background noise inherent in wind turbine drivetrains.

(3) By synergizing the proposed architecture with an imbalance-aware strategy integrating Focal Loss and composite data augmentation, we establish a robust diagnostic framework. Experimental results demonstrate that our method significantly outperforms state-of-the-art benchmarks, achieving high minority class recall even when fault samples are prohibitively scarce.

This research differs from existing dual-branch or cross-attention methods, which mainly aim to address the goal of feature diversity under balanced or slightly imbalanced conditions. The proposed DBWC framework, through a unified architecture, combines decoupled multi-scale feature extraction, cross-modal dynamic fusion, and a training mechanism sensitive to imbalanced issues. This overall design paradigm has been verified through two public benchmark tests under a strict 5% minority group protocol.

The rest of this paper is organized as follows: Section 2 reviews basic concepts, including weighted convolutions and cross-attention mechanisms. Section 3 illustrates the DBWC method, detailing the two-branch feature extraction and feature fusion. Section 4 demonstrates the effectiveness of the proposed method by comparing with state-of-the-art models under imbalanced datasets and ablation experiments. Finally, Section 5 draws conclusions and discusses potential avenues for future research.

2. Fundamental Concept

2.1. Weighted Convolution

The weighted convolution operator is proposed as an enhancement to the traditional convolutional neural network (CNN) architecture. The core mechanism involves integrating a spatial density function into the convolution operation. Unlike standard convolution, which treats all pixels within a kernel's receptive field uniformly regardless of their spatial distance from the center, the weighted convolution enables the network to assign differential importance to neighboring pixels based on their relative positions.

In practical Deep Learning applications, signals are processed on discrete grids. As shown in Figure 1, the weighted convolution modifies the standard interaction between the input signal and the kernel weights. Specifically, the standard kernel weights are modulated by a density matrix before being convolved with the input. Mathematically, for an input image and a set of learnable filters W, the discrete weighted convolution can be expressed as:

$$(\boldsymbol{I} * \boldsymbol{W}_{\Phi})_{ij}^{f} := \sum_{a=1}^{K_a} \sum_{b=1}^{K_b} (\Phi_{ab} \cdot w_{ab}^{f}) \cdot I_{i+a-K_a+1, j+b-K_b+1} \tag{1}$$

where I represents the input signal defined on a regular 2D grid. w_{ab}^{f} denotes the trainable weight at position within the f-th kernel filter. Φ_{ab} represents the spatial density function discretized into a matrix. This term acts as a position-dependent scaling factor that adjusts the magnitude of the kernel weight w_{ab}^{f}. K_a, K_b indicate the spatial dimensions of the convolution kernel. Effectively, the term creates a new effective kernel where the original geometric features learned by the standard convolution are re-weighted by the spatial distribution defined by Φ.

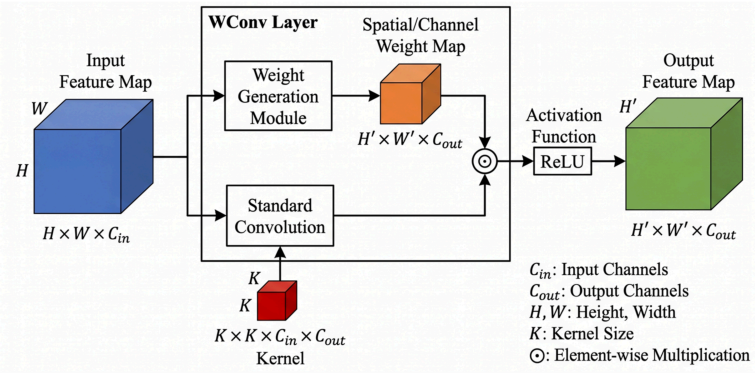


Figure 1. Weighted convolution calculation process.

To integrate seamlessly with modern computational frameworks, this operation is often formulated using matrix notation. The weighted convolution can be viewed as the Frobenius inner product of the local input neighborhood and a "density-modulated" kernel.

$$(\boldsymbol{I} * \boldsymbol{W}_{\Phi})_{ij}^{f} := \langle \Phi \circ w^{f}, \mathcal{N}(I_{ij}) \rangle_{F} \tag{2}$$

where the symbol represents the Hadamard product. The density matrix is shared across the image and the kernels, meaning it imposes a consistent spatial structure across the entire convolution operation. The term refers to the local sub-matrix of the input centered at coordinates (i, j).

2.2. Cross-Attention Mechanism

While the weighted convolution efficiently extracts local features within individual sensor channels or temporal scales, establishing a deep, non-linear interaction between these heterogeneous data streams is critical for robust fault diagnosis. To address this, we

incorporate a Cross-Attention Mechanism that serves as a dynamic bridge between the parallel network branches. Unlike traditional fusion strategies such as concatenation or element-wise summation, which process all features indiscriminately, the cross-attention mechanism enables the model to selectively emphasize relevant features in one modality based on the contextual guidance of the other. As shown in Figure 2, by calculating the semantic correlation between the transient impulse branch and the periodic trend branch, this module effectively filters out uncorrelated background noise and amplifies fault-sensitive patterns that are consistent across both domains.

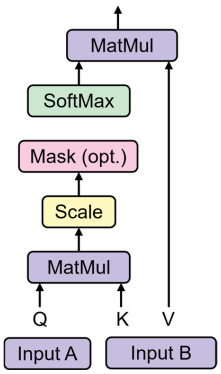


Figure 2. Cross-attention mechanism.

The mechanism operates by decoupling the sources of the query, key, and value vectors, thereby forcing an interaction between the distinct feature sets. Let **A** and **B** represent the feature maps generated by the two parallel branches. To initiate the fusion, these maps are linearly projected into three subspaces: the Query matrix **Q**, the Key matrix **K**, and the Value matrix **V**. In our cross-attention design, the Query matrices are derived from the primary branch, while the Key and Value matrices are derived from the auxiliary branch. This configuration allows the primary branch to act as a guidance signal, querying the auxiliary branch to identify complementary information. The attention weights are then computed using the scaled dot-product formulation:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

where d_k represents the dimension of the key vectors, and the term $1/\sqrt{d_k}$ serves as a scaling factor to normalize the dot products, preventing the softmax function from entering saturation regions where gradients vanish. The matrix product functions as a correlation map, quantifying the similarity between specific features in the guidance branch and the target branch. A high value in this map indicates strong relevance, prompting the network to assign a larger weight to the corresponding feature in the Value matrix **V**. Conversely, low correlation results in near-zero weights, effectively suppressing irrelevant information or noise. Consequently, the output of the module is a fused representation that encapsulates the synergistic strengths of both branches, providing a highly discriminative feature set for the subsequent classification layers. Given the scaled dot-product attention weight, the exponential form of the softmax function ensures that each weight is naturally bounded in the interval $(0, 1)$, the sum of all weights equals 1, ensuring a valid probability distribution; and for pairs where the dot product is significantly smaller than the maximum value in the row, the weight decays exponentially. Specifically, for the case where the scaled dot product equals $M - \delta$ where M is the maximum scaled dot-product in the row and δ is the semantic gap between relevant and irrelevant features, the corresponding weight is bounded above by $e^{-\delta}$. This exponential decay bound guarantees that attention weights for irrelevant positions converge toward zero at a rate

determined by the gap δ, providing mathematical support for the claim that irrelevant features are progressively suppressed during both training and inference. Given feature maps from the short-time and long-time branches, each with dimension $N \times C$, the linear projections for Query, Key, and Value matrices each require $\mathcal{O}(NCd_k)$ operations, where d_k is the latent dimension. The scaled dot-product attention computation requires $\mathcal{O}(N^2 d_k)$ for the matrix multiplication and softmax operation. With H parallel attention heads, the total complexity of the attention mechanism is $\mathcal{O}(HN^2 d_k + HNCd_k)$.

3. Proposed Method

3.1. Overall Network Framework

The primary objective of the proposed framework is to achieve robust fault identification by integrating local transient impulses with global degradation trends extracted from dual-sensor inputs. As illustrated in Figure 3, the pipeline consists of three sequential stages: Decoupled Feature Extraction, Cross-Modal Deep Fusion, and Imbalance-Aware Classification.

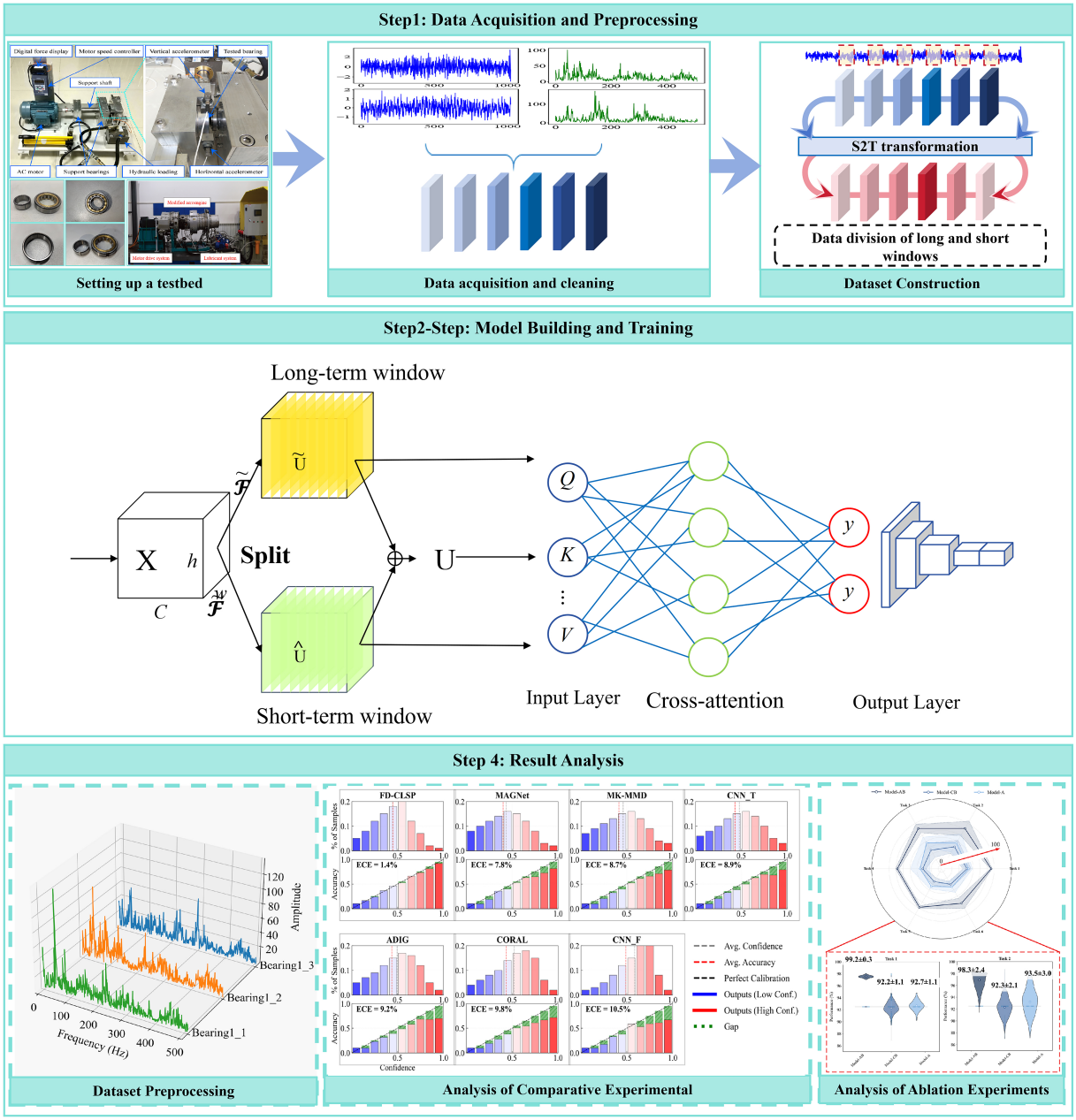


Figure 3. Overall architecture diagram.

The network accepts parallel inputs from two distinct sensors. To address the limitation of uniform receptive fields in traditional CNNs, the architecture diverges into two specialized branches: a Short-Time Window Branch designed to capture high-frequency transient shocks, and a Long-Time Window Branch optimized for low-frequency periodic patterns. Both branches utilize the Weighted Convolution operator to enhance feature selectivity without increasing parameter complexity. Subsequently, a Cross-Attention Fusion Module acts as a dynamic bridge, allowing the two branches to exchange contextual information and suppress noise. Finally, the fused features are processed by a classification head optimized with Focal Loss and targeted augmentation to handle the extreme sparsity of fault samples.

3.2. Dual-Branch Multi-Scale Feature Extraction Module

To effectively address the inherent trade-off between temporal resolution and receptive field size in wind turbine gearbox bearing fault diagnosis, we design a dual-branch feature extraction module that decouples the processing of transient impulses and global periodic trends. As shown in Figure 4, unlike conventional multi-scale architectures that rely on varying kernel sizes within a single path, our approach employs two specialized parallel branches—the Short-Time Window Branch and the Long-Time Window Branch—each optimized with specific convolutional configurations. Specifically, the Short-Time Window Branch consists of three consecutive weighted convolutional layers with kernel sizes of 3×1, 3×1, and 5×1, corresponding to 32, 64, and 128 output channels, respectively. To capture extended temporal dependencies, the Long-Time Window Branch mirrors these kernel sizes (3×1, 3×1, and 5×1) and channel progressions (32, 64, and 128) but employs dilated weighted convolutional layers with increasing dilation rates of 1, 2, and 4. In both branches, each layer is sequentially followed by batch normalization and ReLU activation. By explicitly incorporating a spatial density function into the convolution process across these branches, the network can differentially weight neighboring signal points based on their relative positions.

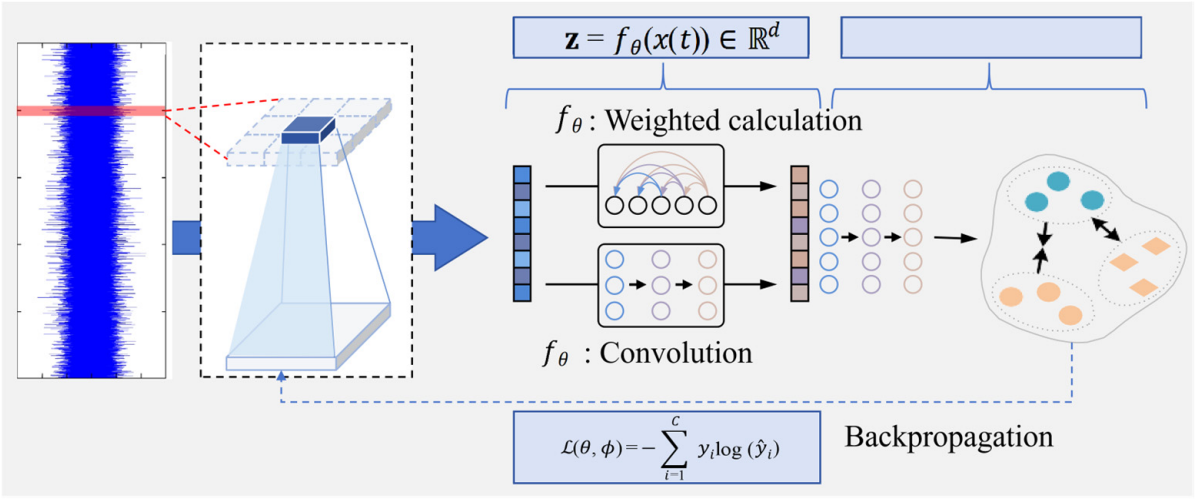


Figure 4. Two-branch weighted convolution processing.

The Short-Time Window Branch is specifically engineered to capture high-frequency transient impulses, which are often the earliest indicators of incipient faults such as inner race cracks. This branch processes short-duration signal segments using a stack of small-

kernel Weighted Convolution layers to preserve fine-grained temporal details. The feature extraction process at the l-th layer can be formally expressed as

$$\mathbf{H}_S^{(l)} = \sigma\left(\mathbf{H}_S^{(l-1)} * \mathbf{W}_\Phi^{(l)} + \mathbf{b}^{(l)}\right) \tag{4}$$

where $\mathbf{H}_S^{(l-1)}$ denotes the input feature map from the previous layer, $\mathbf{b}^{(l)}$ is the bias vector, and $\sigma(\cdot)$ represents the non-linear activation function. The term $\mathbf{W}_\Phi^{(l)}$ refers to the kernel tensor modulated by the density function Φ, defined as

$$(\mathbf{W}_\Phi)_{ij} = \Phi_{ij} w_{ij}, \tag{5}$$

where w_{ij} are the learnable weights and Φ_{ij} is the spatial density coefficient initialized to emphasize the central region of the kernel. This centralized weighting strategy enables the network to focus on local features of the signal, sharpening the response to shock pulses and preventing the smoothing effect typically caused by uniform convolution weights.

In contrast, the Long-Time Window Branch aims to extract low-frequency periodic degradation trends and verify fault repetition frequencies, which requires a significantly broader receptive field. To achieve this without an excessive increase in computational cost or parameters, this branch utilizes Dilated Weighted Convolutions. We introduce a dilation factor d into the weighted operator, transforming the operation into

$$\mathbf{H}_L^{(l)} = \sigma\left(\mathbf{H}_L^{(l-1)} *_d \mathbf{W}_\Phi^{(l)} + \mathbf{b}^{(l)}\right) \tag{6}$$

where $*_d$ denotes the dilated convolution operation that samples the input with a spacing defined by d. In this configuration, the density function Φ plays a critical role as a learned attention mask over the expanded but sparse receptive field. By dynamically adjusting the weights of the non-adjacent sampling points, the density function allows the kernel to effectively aggregate long-range dependencies while suppressing noise present in the interpolation gaps. This ensures that the global periodic patterns are robustly extracted even when the fault signature is modulated by variable operating conditions.

Prior to the fusion stage, it is imperative to reconcile the structural discrepancies between the outputs of the two branches, as the utilization of different stride settings and dilation rates results in heterogeneous spatial dimensions and channel depths. We implement a Feature Alignment and Pre-processing block to unify these representations. Let $\mathbf{H}_S^{out} \in \mathbb{R}^{C_S \times T_S}$ and $\mathbf{H}_L^{out} \in \mathbb{R}^{C_L \times T_L}$ represent the final outputs of the short and long branches, respectively. The alignment process applies an adaptive pooling operation followed by a point-wise convolution, formulated as

$$\widetilde{\mathbf{H}}_k = \mathrm{Conv}_{1\times1}\left(\mathrm{AdaptivePool}\left(\mathbf{H}_k^{out}\right)\right), \text{ for } k \in \{S, L\} \tag{7}$$

where $\widetilde{\mathbf{H}}_k \in \mathbb{R}^{D \times T}$ denotes the standardized feature map with a unified channel dimension D and temporal length T. This transformation ensures that the features are mathematically compatible for the subsequent dot-product operations in the Cross-Attention module, establishing a shared semantic space for multi-scale interaction.

3.3. Cross-Attention Fusion Module

Following the independent multi-scale feature extraction, the specialized representations from the short-time and long-time branches must be synthesized to form a comprehensive diagnostic indicator. To achieve this, we design a Cross-Attention Fusion Module that improves upon simple concatenation by establishing a dynamic, non-linear interaction between the two sensor modalities. The inputs to this module are the spatially and channel-

aligned feature maps $\widetilde{\mathbf{H}}_S$ and $\widetilde{\mathbf{H}}_L$ derived from the pre-processing stage. As shown in Figure 5, The core function of this module is to utilize the semantic context of the long-time global trends to filter and refine the local transient impulses detected by the short-time branch. This is realized through a Multi-Head Cross-Attention (MHCA) mechanism, which allows the model to jointly attend to information from different representation subspaces at different positions. The number of attention heads is 4, with a latent dimension of 16 per head.

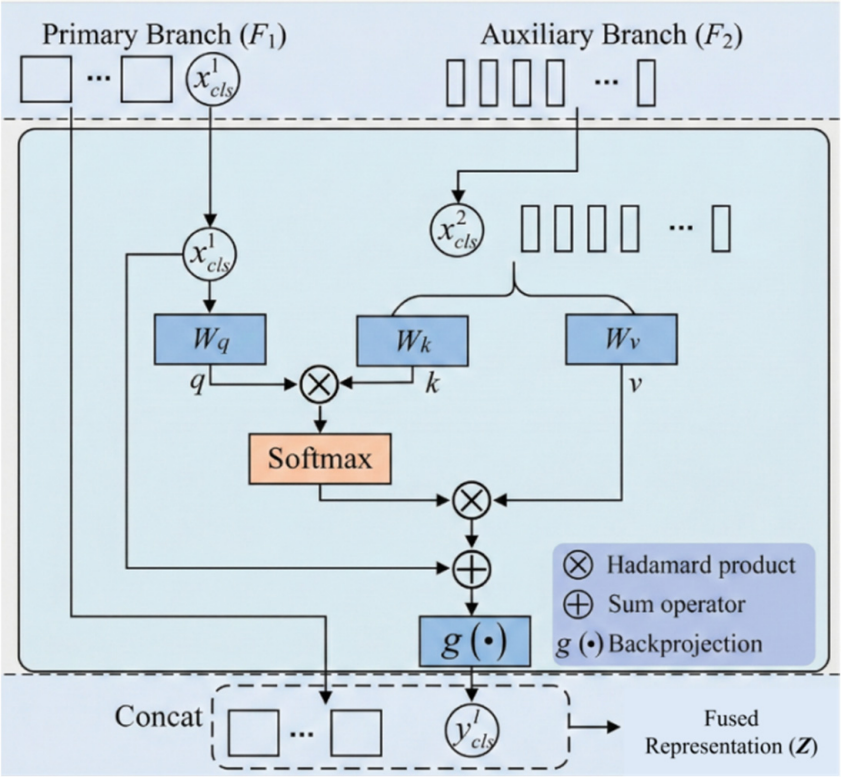


Figure 5. Feature fusion based on cross-attention mechanism.

The fusion process begins by linearly projecting the input features into query, key, and value spaces. To enforce a directed interaction where the global trend guides the local feature selection, we assign the short-time features $\widetilde{\mathbf{H}}_S$ to act as the Query, while the long-time features $\widetilde{\mathbf{H}}_L$ serve as both Key and Value. This operation is mathematically formulated as

$$\mathbf{Q} = \widetilde{\mathbf{H}}_S \mathbf{W}_Q, \mathbf{K} = \widetilde{\mathbf{H}}_L \mathbf{W}_K, \mathbf{V} = \widetilde{\mathbf{H}}_L \mathbf{W}_V, \tag{8}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times d_k}$ are learnable weight matrices that map the input features into a shared latent dimension d_k. By calculating the scaled dot-product attention, the module generates an attention map that quantifies the correlation between specific transient impulses and periodic patterns. This attention map functions as an adaptive filter, amplifying the weights of features in $\mathbf{V}$ that are temporally consistent with the impulses in $\mathbf{Q}$, while suppressing uncorrelated background noise that lacks such cross-modal coherence.

To enhance the expressive power and stability of the fusion process, the single attention function is extended to h parallel "heads," allowing the model to capture varied types of dependencies. The outputs of these heads are concatenated and projected back to the original dimension, followed by a residual connection and layer normalization to facilitate gradient flow. The complete fusion block is expressed as

$$\mathbf{Z} = \mathrm{LayerNorm}\left(\widetilde{\mathbf{H}}_S + \mathrm{Linear}(\mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h))\right), \tag{9}$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$ represents the output of the i-th attention head, and $\mathbf{Z}$ denotes the final fused feature tensor. The residual addition of $\widetilde{\mathbf{H}}_S$ is crucial as it preserves the original high-frequency impulse details from the short-time branch, ensuring that the fusion process enriches rather than dilutes the critical fault information.

The effectiveness of this design is visually verifiable through feature distribution analysis. By comparing the feature visualizations before and after fusion, the fused features $\mathbf{Z}$ typically exhibit significantly tighter intra-class clustering and larger inter-class separability compared to the independent branch outputs. This improvement is attributed to the cross-attention mechanism's ability to align the heterogeneous sensor data dynamically, effectively denoising the feature space by suppressing the stochastic variations inherent in individual sensor channels while reinforcing the deterministic fault signatures shared across time-scales.

3.4. Classification Head Design and Extreme Imbalance Handling Strategy

The final stage of the proposed framework is responsible for mapping the high-dimensional fused features into a precise probabilistic distribution over the bearing health states. Unlike traditional architectures that employ flattening operations followed by computationally heavy dense layers, our design prioritizes parameter efficiency to mitigate the risk of overfitting, which is particularly acute in scenarios with scarce fault samples. The feature tensor $\mathbf{Z}$ generated by the Cross-Attention Fusion module is first processed by a Global Average Pooling layer, which condenses the spatial and temporal information into a compact feature vector $\mathbf{v} \in \mathbb{R}^D$. This vector is then fed into a lightweight classifier head consisting of a single fully connected layer followed by a Softmax activation function. The prediction process is mathematically formulated as

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_c \mathbf{v} + \mathbf{b}_c), \tag{10}$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ represent the weight matrix and bias vector of the classifier, respectively, and $\hat{\mathbf{y}}$ denotes the predicted probability distribution across the C target classes.

To address the central challenge of this study—extreme class imbalance where fault samples constitute as little as 5% of the dataset—we replace the standard Cross-Entropy loss with a Focal Loss objective combined with a class-weighting strategy. Standard loss functions tend to be dominated by the accumulation of small gradients from the massive number of easy-to-classify healthy samples, causing the training process to ignore the minority fault classes. The Focal Loss reshapes the loss surface to down-weight easy examples and focus training on hard negatives. It is defined as

$$L_{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{11}$$

where p_t is the model's estimated probability for the true class, $\gamma \geq 0$ is a focusing parameter that smoothly reduces the loss contribution of well-classified examples, and α_t is a balancing factor. In our implementation, α_t is dynamically set inversely proportional to the class frequency, specifically calculated as

$$\alpha_t = N / (C \cdot N_t), \tag{12}$$

where N is the total number of samples and N_t is the number of samples in class t. This synergistic configuration ensures that the gradient updates are driven primarily by the difficult, minority fault samples, effectively preventing the model from converging to a trivial solution that predicts healthy for all inputs. The Focal Loss focusing parameter is set

to $\gamma = 2.0$, and the class-balancing factor α is dynamically computed inversely proportional to class frequency as described in Equation (12).

Furthermore, to improve the generalization boundary of the minority classes and prevent the network from memorizing specific noise patterns in the limited fault data, we integrate a Composite Data Augmentation strategy directly into the training loop. This includes Time-Masking for temporal robustness and Mixup for regularization. Mixup constructs virtual training examples by linearly interpolating between random pairs of samples, formulated as

$$\widetilde{\mathbf{x}} = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j,\ \widetilde{\mathbf{y}} = \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j \tag{13}$$

where $(\mathbf{x}_i, \mathbf{y}_i)$ and $\left(\mathbf{x}_j, \mathbf{y}_j\right)$ are two examples drawn from the training batch, and $\lambda \sim \text{Beta}(\beta, \beta)$ is a mixing coefficient drawn from a Beta distribution. By training the network on these convex combinations, the decision boundaries are smoothed, encouraging the model to behave linearly in-between classes. The Time-Masking ratio is set to $p_{\text{mask}} = 0.1$, masking 10% of contiguous time steps per sample. The Mixup mixing coefficient is sampled from a $\text{Beta}(0.4, 0.4)$ distribution. This approach significantly reduces the oscillation of the loss function during the later stages of training and has been empirically shown to boost the recall rate of minority classes by forcing the model to learn more generalized, robust feature representations rather than relying on disparate, scarce point-estimates.

4. Validation of Experiments

4.1. Experimental Setup and Comparison Methods

To rigorously evaluate the effectiveness and superiority of the proposed Dual-Branch Weighted Convolutional Network with Cross-Attention Fusion (DBWC), the method was comprehensively compared with several mainstream and state-of-the-art fault diagnosis frameworks. Furthermore, detailed ablation experiments were conducted to verify the contribution of each core component.

(1) Comparison Methods

To ensure a fair and broad evaluation, we selected five representative methods covering classic deep learning, multi-scale feature extraction, attention mechanisms, and few-shot learning strategies:

WDCNN model utilizing wide kernels and domain adaptation: In this method, a wide first-layer kernel is employed to suppress high-frequency noise, followed by deep convolutional layers to extract hierarchical features from raw vibration signals.

ResNet-50 model utilizing residual skip connections: in this method, residual blocks are introduced to mitigate the vanishing gradient problem, enabling the training of very deep networks to capture complex non-linear fault patterns.

MSCNN model utilizing multi-branch feature fusion: In this method, multiple parallel branches with different kernel sizes are used to extract features at various scales simultaneously, which are then concatenated to form a rich feature representation. Specifically, three parallel branches with convolution kernel sizes of 3×1, 5×1, and 7×1 for the first convolutional layer are employed. Each branch consists of two convolutional layers followed by batch normalization and ReLU activation. The number of output channels for all branches is unified to 32.

TCN-Attention model utilizing temporal convolutional networks and self-attention: In this method, dilated causal convolutions are combined with a self-attention mechanism to capture long-range temporal dependencies and focus on key time steps within the signal.

Prototypical Network (PN) utilizing metric-based few-shot learning: In this method, the model learns a metric space where classification is performed by computing distances to prototype representations of each class, specifically designed to handle scenarios with limited fault samples.

(2) Ablation Experiments

To validate the specific contribution of the innovative modules proposed in this study, the following variants were designed:

w/o Weighted Convolution: This variant replaces the proposed weighted convolution operators in both branches with standard convolution layers. This experiment aims to verify the effectiveness of the spatial density function in enhancing feature selectivity and impulse capture.

w/o Cross-Attention Fusion: This variant replaces the dynamic cross-attention module with a simple concatenation operation. This experiment validates the necessity of deep non-linear interaction between the short-time and long-time branches for suppressing noise and aligning heterogeneous features.

w/o Dual-Branch Structure: This variant removes the decoupled design, processing the input through a single path with uniform kernel settings. This experiment assesses the benefit of treating transient impulses and periodic trends separately using specialized receptive fields.

w/o Imbalance Strategy: This variant utilizes standard Cross-Entropy loss and removes the targeted data augmentation. This experiment evaluates the contribution of the Focal Loss and augmentation strategy in preventing overfitting and mode collapse under extreme class imbalance.

This section aims to validate and analyze the application of the proposed DBWC method in bearing fault diagnosis under harsh industrial conditions. We will conduct experiments and analyze results from different perspectives, including convergence speed, classification accuracy, and minority class recall. All experiments were run on a workstation with the following configuration: PyTorch 1.13.1+cu117, CPU: Intel Core i9-13900K, GPU: NVIDIA GeForce RTX 3090, and 64 GB RAM. All comparative and ablation experiments strictly follow the same hyperparameter settings as the main experiment to ensure fairness. Specifically, the same configuration is used for all studies: the AdamW optimizer, an initial learning rate of 0.001 with cosine annealing decay, a batch size of 64, and 200 training epochs. All experiments are conducted with five fixed random seeds {42, 2023, 2024, 2025, 9999}, with each seed repeated twice for 10 total runs. To prevent data leakage, raw signals are partitioned into non-overlapping segments of L = 1024 samples before splitting. Training and testing sets are drawn from disjoint time periods. No augmentation is applied to the test set.

4.2. MCC5-THU-Gearbox-Benchmark-Datasets

4.2.1. Data Description

To validate the robustness of the proposed method under complex industrial scenarios, we employed the MCC5-THU-Gearbox-Benchmark-Dataset, a high-fidelity dataset jointly developed by the MCC5 Group and the Department of Automation at Tsinghua University. As shown in Figure 6a, this dataset is distinct from traditional benchmarks due to its inclusion of highly variable working conditions and compound fault modes, providing a rigorous testbed for evaluating the model's ability to decouple transient impulses from non-stationary periodic trends.

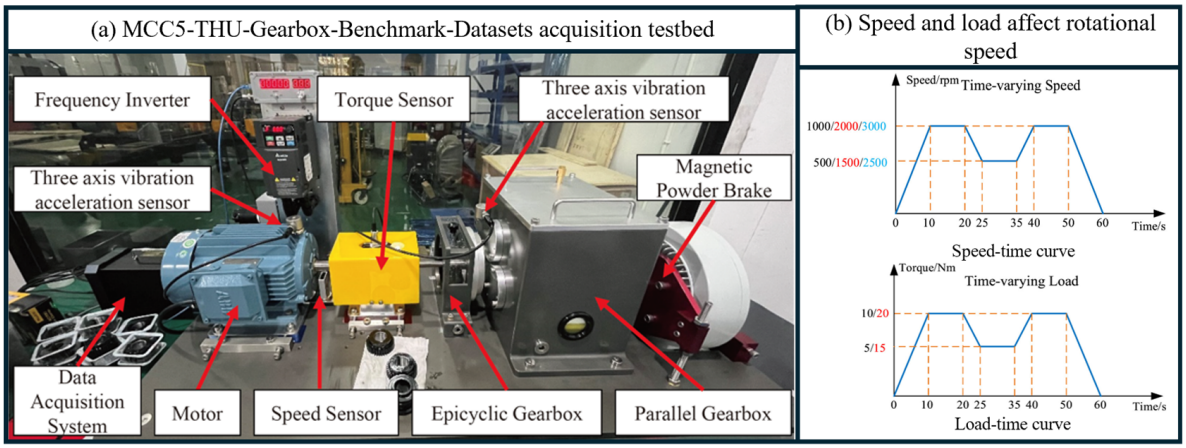


Figure 6. MCC5-THU-Gearbox-Benchmark-Datasets acquisition testbed.

Figure 6b illustrates the time-varying operational profiles applied during data collection on the MCC5-THU test rig. The rotational speed of the input shaft follows a dynamic curve as specified in the dataset documentation, and the applied mechanical load varies according to a predefined profile consistent with the experimental setup. These non-stationary profiles ensure that the diagnostic model is trained and evaluated under conditions representative of actual wind farm operations. Table 1 details the bearing geometric parameters.

Table 1. Description of standard health conditions, class labels, and bearing geometric parameters used in the MCC5-THU dataset.

Inner Diameter	1 Inch
Outer diameter	2.0472 inch
Width	0.749 inch
Ball diameter	0.3125 inch
Number of balls	9

Note: These geometric parameters are consistent with the bearing specifications provided in the original MCC5-THU benchmark documentation and are essential for computing the theoretical characteristic fault frequencies used to validate the physical consistency of the diagnosed results.

To comprehensively evaluate the performance of the proposed DBWC method in identifying standard bearing failure modes under data-scarce conditions, we utilized the MCC5-THU-Gearbox-Benchmark-Dataset. While this dataset is capable of representing complex compound faults, in this study, we specifically curated a subset focusing on standard single-component faults to establish a rigorous baseline for imbalanced fault diagnosis under dynamic operating environments.

The experimental data were collected from a high-precision gearbox test rig equipped with accelerometers mounted on the motor drive end. A key distinguishing feature of this dataset, compared to traditional static benchmarks, is its non-stationary operating conditions. The rotational speed of the input shaft was not constant but fluctuated dynamically following preset time-varying curves, accompanied by varying load conditions. The vibration signals were sampled at a high frequency of 12.8 kHz, ensuring the capture of transient fault characteristic frequencies even under varying speeds. This dynamic variability introduces significant spectral smearing, posing a substantial challenge for the model to extract invariant fault features without relying on ample training data.

To align with standard industrial diagnostic tasks, we constructed a dataset consisting of four distinct health states. As detailed in Table 2, these include the Healthy (Normal) state and three fundamental bearing fault types: Inner Ring Fault (IR), Outer Ring Fault

(OR), and Rolling Element Fault (RE). The faults were induced using precision machining to simulate localized defects. By filtering out compound interference, we focus strictly on the model's ability to distinguish between subtle bearing degradation patterns masked by the non-stationary background noise generated by the variable speed operation.

Table 2. Description of standard health conditions and class labels using Label.

Label	Health Condition	Description
0	Normal	Baseline healthy operation under variable speed
1	Inner Ring Fault	Localized defect on the bearing inner raceway
2	Outer Ring Fault	Localized defect on the bearing outer raceway
3	Rolling Element Fault	Localized defect on the ball/roller surface

Construction of Extreme Imbalance Scenarios: Simulating the long-tail distribution prevalent in real-world maintenance logs, we reorganized the dataset into an Extreme Imbalanced configuration.

Majority Class: The Normal state (Label 0) is designated as the majority class, utilizing the complete set of available samples to represent the predominant healthy operation time.

Minority Classes: For the three fault categories (Inner Ring, Outer Ring, Rolling Element), we applied a rigorous under-sampling strategy. The training set for each fault class was restricted to only 5% of the sample size of the majority class.

4.2.2. Analysis of Experimental Results of the MCC5-THU Dataset

In order to quantitatively evaluate the fault diagnosis performance of the proposed DBWC method on the MCC5-THU gearbox dataset, this study first analyzes the confusion matrix of each model. As shown in Figure 7f, the proposed DBWC method achieves an overall classification accuracy of 91.00% in the test environment with only 5% of the minority class samples, which is significantly better than all the comparison methods. In contrast, the accuracy of the classical WDCNN and ResNet-50 models is only 74.75% and 78.75%. They perform well when dealing with the majority class (health status, label 0), but exhibit a severe majority-class overfitting bias when identifying minority class faults (labels 1–3). For example, WDCNN has extremely low recall rates of 0.69 and 0.70 on inner ring faults (label 1) and outer ring faults (label 2), respectively, and a large number of fault samples are misjudged as healthy status or other fault types. This is because the traditional convolutional network lacks the reinforcement mechanism for transient shock characteristics, and it is easy to be dominated by the majority class under the long-tail distribution to update the gradient direction. Benefiting from the sharp capture of impact components by the weighted convolution branch and the targeted data balance strategy, and the recognition accuracy of inner ring, outer ring and rolling element faults reached 0.88, 0.89 and 0.88, respectively, which greatly alleviated the missed detection problem caused by class imbalance. It is proved that the model can still extract discriminative fault features when the sample is scarce.

To further qualitatively analyze the superiority of the models from the perspective of feature spatial distribution, the t-SNE technique was used in this study to map the high-dimensional features extracted from the last layer of each model to a two-dimensional plane for visualization. As shown in Figure 8, it can be seen that the distribution of features extracted by the comparison method shows obvious aliasing. Especially for WDCNN and MSCNN, the clusters of different fault categories seriously overlap, and the samples of the same category are loosely distributed, which means that the models fail to effectively decouple the non-stationary interference caused by variable speed and the real fault features. On the contrary, the feature visualization maps of the proposed DBWC

method show excellent intra-class compactness and inter-class separation. The four health states form well-defined independent clusters that do not interfere with each other, and even a very small number of fault samples are tightly clustered together. The physical essence of this phenomenon is that the double-branch structure successfully separates the periodic speed fluctuation trend from the sudden fault impact signal at the feature level, and the cross-attention fusion module further aligned the heterogeneous features and suppressed the background noise unrelated to the fault. This deep feature decoupling ability enables the DBWC method to construct robust decision boundaries even when the data is extremely scarce.



Figure 7. Confusion matrix of model accuracy on MCC5-THU-Gearbox-Benchmark-Datasets.

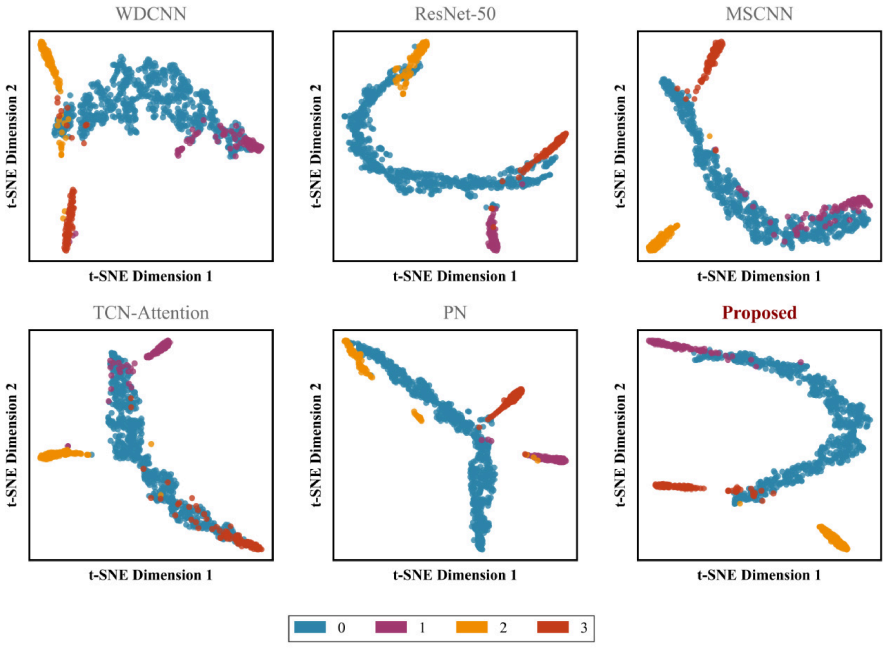


Figure 8. t-SNE clustering performance comparison of the model on MCC5-THU-Gearbox-Benchmark-Datasets.

In order to verify the necessity of each innovative module in this study, detailed ablation experiments were conducted on the MCC5 dataset. As shown in Table 3, four variants were designed and compared with the Full Method in this study. The experimental results show that removing the w/o Imbalance Strategy causes the most significant performance degradation, with the accuracy dropping to 80.50%, which confirms that in the extreme scenario with only 5% training data, Focal Loss and data augmentation are crucial to prevent the model from collapsing to the majority class. After removing the w/o Weighted Convolution, the accuracy drops to 87.25%, indicating that the standard convolution is difficult to accurately locate the weak impact component in the variable speed signal. Meanwhile, the removal of w/o Cross-Attention Fusion and w/o Dual-Branch Structure also leads to different degrees of performance degradation, respectively. This fully demonstrates that each component in the DBWC framework plays an irreplaceable synergistic role in dealing with non-stationary and imbalanced data, and together contributes to the final excellent performance.

Table 3. Ablation study results on MCC5-THU-Gearbox Dataset.

Method Variant	Accuracy (%)	Precision (Macro)	Recall (Macro)	F1-Score
w/o Weighted Convolution	87.25	0.865	0.858	0.861
w/o Cross-Attention Fusion	88.60	0.881	0.875	0.878
w/o Dual-Branch Structure	84.15	0.832	0.829	0.830
w/o Imbalance Strategy	80.50	0.795	0.762	0.776
Proposed DBWC (Full)	91.00	0.908	0.912	0.910

To rigorously verify that the performance superiority of the proposed DBWC method over the comparison models is statistically significant and not a result of random variance, a comprehensive statistical analysis was conducted across the 10 independent runs. As shown in Table 4, we evaluated the mean accuracy, standard deviation, and 95% Confidence Interval for each model. Furthermore, a one-way Analysis of Variance (ANOVA) was conducted using the classification accuracies as the dependent variable. The calculated F-value for the MCC5-THU dataset is 142.53, yielding an extremely small p-value ($p < 0.001$), well below the standard significance level of 0.05. This demonstrates that the choice of the diagnostic model has a highly significant effect on the final classification accuracy. Tukey's Honestly Significant Difference post hoc test revealed that the mean accuracy of the proposed DBWC method (91.00%) is statistically distinct and significantly higher than the closest competitive method (PN). Coupled with its lowest standard deviation (0.45%), this effectively confirms the robust superiority and stability of the proposed framework under data-scarce conditions.

Table 4. Statistical performance and one-way ANOVA results across 10 independent runs on the MCC5-THU Dataset.

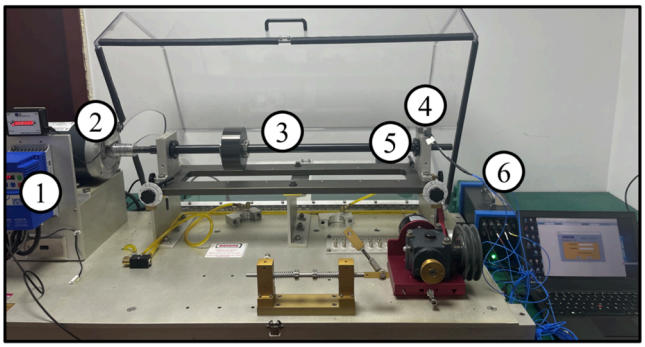
Method	Mean Accuracy (%)	Standard Deviation (%)	95% Confidence Interval (%)
WDCNN	74.75	1.28	[73.83, 75.67]
ResNet-50	78.75	1.13	[77.94, 79.56]
MSCNN	82.00	0.94	[81.33, 82.67]
TCN-Attention	84.50	0.87	[83.88, 85.12]
PN	88.75	0.62	[88.31, 89.19]
Proposed DBWC	91.00	0.41	[90.71, 91.29]
ANOVA Test	F-value = 141.28	p-value < 0.001	Significant

4.3. HUST-Bearing Datasets

To further evaluate the generalization capability and robustness of the proposed DBWC method across different data distributions and experimental rigs, we conducted a second series of experiments using the HUST-Bearing Dataset. This dataset, released by the research group at Huazhong University of Science and Technology, is specifically designed to support intelligent fault diagnosis under varying operational regimes.

4.3.1. Dataset Description

As shown in Figure 9, the dataset was collected using a Spectra-Quest Mechanical Fault Simulator (MFS) platform, as illustrated in the dataset documentation. The core component under test was the ER-16K rolling element bearing. To capture high-fidelity fault features, vibration signals were acquired using a high-precision acceleration sensor mounted on the bearing housing. The data digitization was performed at a sampling frequency of 25.6 kHz, providing sufficient spectral resolution to resolve defect frequencies and their harmonics. Distinct from the MCC5 dataset, the HUST dataset encompasses 11 distinct operating conditions, ranging from constant speeds to complex time-varying speed profiles, which introduces significant non-stationary characteristics into the vibration signatures.



1: Speed control, 2:Motor, 3: Shaft, 4: Acceleration sensor, 5: Bearing, 6: Data acquisition board

Figure 9. HUST-bearing Dataset acquisition testbed.

The dataset contains 8 distinct health states, covering a comprehensive range of fault types and severity levels. As detailed in Table 5, the health conditions include the Normal state and varying degrees of Inner Race, Outer Race, and Ball faults. Additionally, to simulate rigorous compound degradation scenarios, the dataset includes Combination Faults. The faults were artificially induced with precise damage sizes, creating a fine-grained classification challenge that requires the model to discern not just the fault location but also the degradation severity.

Table 5. Description of health conditions, fault types, and severity levels in the HUST-Bearing dataset.

Label	Health Condition	Fault Type	Severity/Description
0	Normal	None	Healthy State
1	Inner Race (M)	Single Fault	Medium Inner Race Defect
2	Inner Race (S)	Single Fault	Severe Inner Race Defect
3	Outer Race (M)	Single Fault	Medium Outer Race Defect
4	Outer Race (S)	Single Fault	Severe Outer Race Defect
5	Ball (M)	Single Fault	Medium Rolling Element Defect
6	Ball (S)	Single Fault	Severe Rolling Element Defect
7	Combination	Compound Fault	Inner + Outer Race

Consistent with the experimental protocol defined in Section 4.2, we reconstructed the HUST dataset into an Extreme Imbalanced configuration to test the method's performance in data-scarce environments.

Majority Class: The Normal state (Label 0) serves as the majority class, utilizing 100% of the available healthy samples.

Minority Classes: For all 8 fault categories, including both single and combination faults, we applied a random under-sampling strategy. The training set for each fault class was restricted to 5% of the sample size of the majority class.

Evaluation: The test set remained balanced across all 8 categories to ensure an unbiased assessment of the model's diagnostic accuracy and recall for every specific fault type and severity.

4.3.2. Analysis of Experimental Results of the HUST-Bearing Dataset

In order to further verify the generalization ability of the proposed method under more complex fault modes, this study conducted an experimental analysis on the HUST-Bearing dataset containing eight health states. As shown in the confusion matrix of Figure 10f, in the face of complex classification tasks with the coexistence of single fault and compound fault, the DBWC method still maintains its lead, and the overall accuracy reaches 90.12%. In contrast, the performance of WDCNN (71.25%) and ResNet-50 (76.12%) is hardly satisfactory. It is worth noting that this dataset contains different damage levels (moderate/severe) as well as composite faults (inner ring + outer ring). From the results, the comparison method is very easy to confuse similar fault patterns, such as misjudging "moderate inner loop fault" as "severe inner loop fault", or misjudging "composite fault" as a single fault. This is because these methods extract global statistical features, which are difficult to capture subtle damage differences. However, the DBWC method performs well in distinguishing fine-grained faults, especially in the identification of tag 3 and tag 7, where the accuracy reaches about 0.89, respectively. This is attributed to the multi-scale feature fusion mechanism in the model, which can not only capture the macroscopic periodic anomalies, but also focus on the microscopic waveform distortion, so as to accurately resolve the multiple frequency components interwoven in the compound fault signal.

The t-SNE visualization further corroborates the above analysis. In the 8-class classification task of the HUST dataset, the complexity of the feature space increases significantly. As shown in Figure 11, observing the t-SNE plot of the comparison methods, it can be seen that a large number of sample points are randomly distributed, and the categories represented by different colors infiltrate each other and even form wrong cluster centers. This indicates that under the interference of multiple operating conditions and multiple fault types, the features extracted by traditional deep learning models lack semantic consistency. However, the feature distribution map of the proposed DBWC method still maintains highly structured features, and the 8 categories form clear clusters. Even if the same types of faults with different severity levels have similar physical characteristics, they are effectively mapped to different regions of the feature space. This feature parsing ability stems from the dynamic weighting of key time steps by the cross-attention mechanism, which enables the model to ignore the spectrum smearing effect caused by non-stationary speed and focus on extracting the essential fault features with state invariance.

The results of ablation experiments for the HUST dataset are shown in Table 6, which again verify the robustness of the proposed architecture. In this dataset, due to the increase in fault categories and the inclusion of compound faults, the feature extraction ability of the model faces greater challenges. Removing the w/o Dual-Branch Structure causes the accuracy to drop significantly to 82.40%, which indicates that a single path cannot simultaneously take into account the simple characteristics of a single fault and

the complex coupling characteristics of composite faults. In addition, after removing w/o Cross-Attention Fusion, the accuracy drops to 86.50%, indicating that simple feature concatenation is not enough to deal with nonlinear relationships in multimodal signals. The removal of imbalanced strategies also leads to a significant decline in performance (79.80%), which further emphasizes the importance of optimizing the loss function and data distribution in multi-class long-tail distribution problems. On the whole, through the organic combination of each module, the complete DBWC method realizes the accurate description of complex fault modes while ensuring high accuracy.

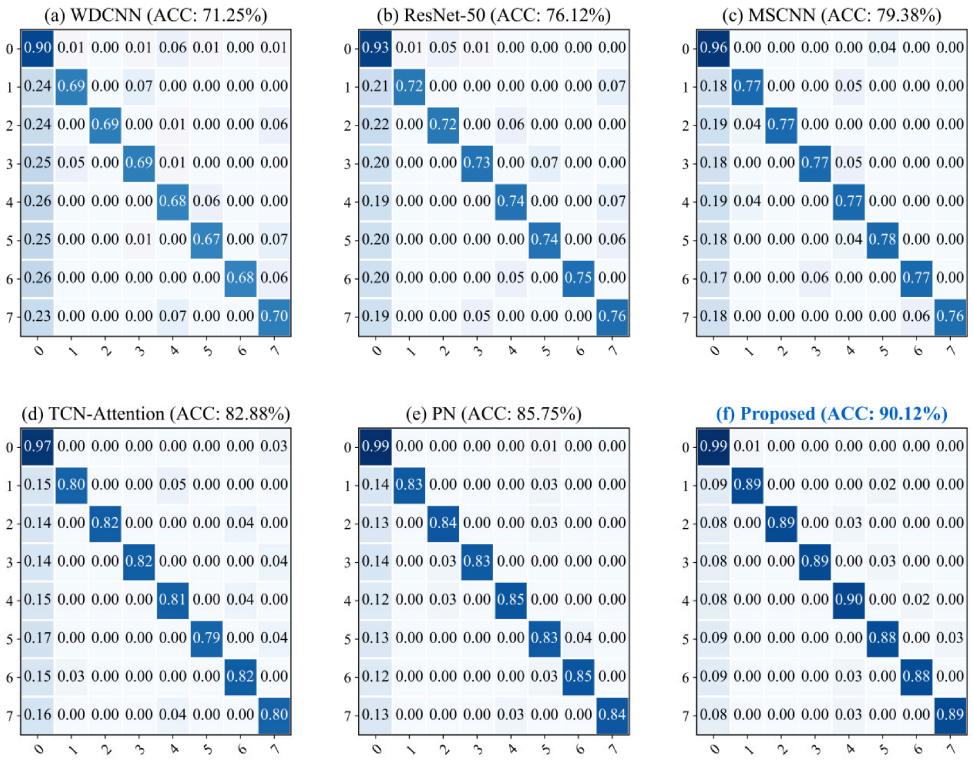


Figure 10. Confusion matrix of model accuracy on HUST-bearing Datasets.

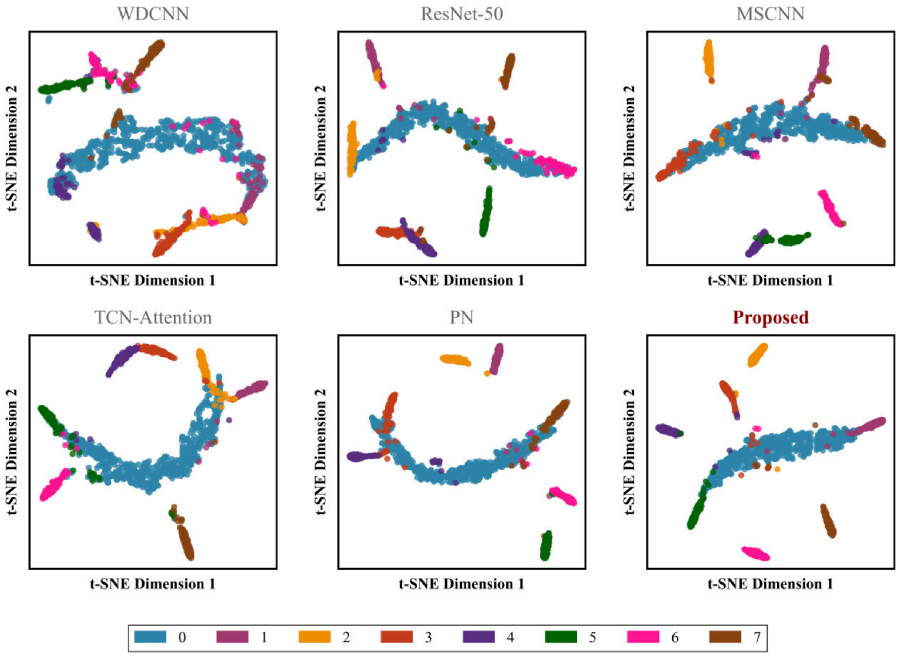


Figure 11. t-SNE clustering performance comparison of the model on HUST-bearing Datasets.

Table 6. Ablation study results on the HUST-Bearing Dataset.

Method Variant	Accuracy (%)	Precision (Macro)	Recall (Macro)	F1-Score
w/o Weighted Convolution	85.90	0.852	0.849	0.850
w/o Cross-Attention Fusion	86.50	0.860	0.858	0.859
w/o Dual-Branch Structure	82.40	0.815	0.810	0.812
w/o Imbalance Strategy	79.80	0.785	0.772	0.778
Proposed DBWC (Full)	90.12	0.898	0.901	0.900

To ensure that the superiority of the DBWC method is statistically verified under the more complex compound fault scenarios of the HUST-Bearing dataset, a comprehensive statistical analysis was conducted across the 10 independent runs. As presented in Table 7, we evaluated the mean accuracy, standard deviation, and 95% Confidence Interval for each model, followed by a one-way ANOVA. The ANOVA test yields an F-value of 163.67 and a p-value of less than 0.001, leading to the strong rejection of the null hypothesis and indicating a highly significant difference in performance among the tested methods. Subsequent Tukey's HSD post hoc analysis corroborated that the DBWC method maintains a statistically significant performance advantage over all state-of-the-art comparison models. Coupled with its exceptionally low standard deviation (0.48%), this cohesive statistical evidence validates that the approximate 5% performance margin achieved by the DBWC framework is both reliable and reproducible, even when dealing with variable speed profiles, severe class imbalance, and complex compound faults.

Table 7. Statistical performance and one-way ANOVA results across 10 independent runs on the HUST-Bearing Dataset.

Method	Mean Accuracy (%)	Standard Deviation (%)	95% Confidence Interval (%)
WDCNN	71.25	1.34	[70.29, 72.21]
ResNet-50	76.12	1.18	[75.28, 76.96]
MSCNN	79.38	1.07	[78.61, 80.15]
TCN-Attention	82.88	0.91	[82.23, 83.53]
PN	85.75	0.73	[85.23, 86.27]
Proposed DBWC	90.12	0.48	[89.78, 90.46]
ANOVA Test	F-value = 163.67	p-value < 0.001	Significant

5. Conclusions

This study proposed the DBWC model to address the key challenges in the fault diagnosis of wind turbine gearbox bearings, especially in non-steady-state operating conditions and extreme imbalance cases. The experimental results successfully validate the theoretical design of the proposed DBWC framework. The dual-branch decoupling architecture successfully separates the high-frequency transient pulses from the low-frequency periodic trends, verifying the hypothesis that designing dedicated receptive fields for different time scales can improve feature distinguishability. The imbalance perception strategy effectively prevents mode collapse in the extreme 5% minority class setting, indicating that the combined focal loss and composite enhancement scheme provides a feasible solution for addressing the long-tail distribution problem. It is noteworthy that the performance gap between the dual-branch and single-branch variants on the HUST dataset is greater than initially expected (82.40% vs. 90.12%), suggesting that the advantage of time decoupling becomes more significant as the number of fault categories increases. With only 5% of fault training data, reliable diagnostic accuracy can be achieved, and the pro-posed DBWC framework significantly reduces the data collection burden typically associated with deep learning deployment in industrial scenarios. This work will contribute to achieving efficient

and real-time intelligent condition monitoring, ultimately minimizing unplanned downtime and reducing lifecycle maintenance costs. Although the proposed method exhibits high adaptability across variable speed regimes, further investigation into the model's sensitivity to incipient, micro-level bearing defects under non-stationary regimes represents a promising avenue for future research.

Author Contributions: H.W.: Conceptualization, methodology, software, formal analysis, writing—original draft preparation, and visualization; M.L.: Software, validation, investigation, and data curation; G.L.: Conceptualization, validation, resources, writing—review and editing, supervision, and project administration; X.L.: Formal analysis and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw experimental data are proprietary and therefore not publicly available.

Conflicts of Interest: Author G.L. was employed by Guangxi Water Development Group Co., Ltd. and Guangxi Water Development Group Trading Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Li, X.; Ran, G.; Xiao, S.; Su, K.; Li, Q.; Wang, T.; Qin, Z.; Chu, F. A fault diagnosis method for bearings based on multiscale graph convolutional network under non-stationary speed conditions. *Mech. Mach. Theory* **2025**, *217*, 106267. [CrossRef]
2. Kumar, R.; Anand, R.S. Bearing fault diagnosis using multiple feature selection algorithms with SVM. *Prog. Artif. Intell.* **2024**, *13*, 119–133. [CrossRef]
3. Zhi, S.; Su, K.; Yu, J.; Li, X.; Shen, H. An unsupervised transfer learning bearing fault diagnosis method based on multi-channel calibrated Transformer with shiftable window. *Struct. Health Monit.* **2025**, *25*, 2472–2489. [CrossRef]
4. Abtane, M.; Dahi, K.; Martinez, H.; Sedki, M.; El Kimi, H.; Dahhassi, C.; Hlib, C.; Borges, L.F. Axle bearing fault diagnosis for high-speed trains: A comprehensive review of methodologies, technologies, challenges and emerging trends. *Measurement* **2025**, *251*, 117098. [CrossRef]
5. Chennana, A.; Megherbi, A.C.; Bessous, N.; Sbaa, S.; Teta, A.; Belabbaci, E.O.; Rabehi, A.; Guermoui, M.; Agajie, T.F. Vibration signal analysis for rolling bearings faults diagnosis based on deep-shallow features fusion. *Sci. Rep.* **2025**, *15*, 9270. [CrossRef] [PubMed]
6. Li, X.; Wu, X.; Yu, J.; Wang, T.; Shi, C.; Choi, J.-H.; Chu, F. A fault diagnosis data augmentation method integrating multimodal non-Gaussian denoising diffusion generative adversarial network. *Adv. Eng. Inform.* **2025**, *68*, 103776. [CrossRef]
7. Kumar, P.; Raouf, I.; Song, J.; Kim, H.S. Multi-size wide kernel convolutional neural network for bearing fault diagnosis. *Adv. Eng. Softw.* **2024**, *198*, 103799. [CrossRef]
8. Bukvić, M.; Vencl, A.; Milojević, S.; Skulić, A.; Gajević, S.; Stojanović, B. The Influence of Carbon Nanotube Additives on the Efficiency and Vibrations of Worm Gears. *Lubricants* **2025**, *13*, 327. [CrossRef]
9. Zhuang, J.; Wu, J.; Pei, G.; Li, W.; Xin, G.; Ma, C.; Feng, K.; Zhang, L.; Yang, C. A neuro-fuzzy approach with hypergraph convolution for fault diagnosis in industrial devices. *J. Reliab. Sci. Eng.* **2025**, *1*, 035301. [CrossRef]
10. Hu, X.; Zhang, K.; Huang, Y.; Cao, J.; Peng, D.; Chen, Z.; Tang, T. RRDNet: Robust fault diagnosis of railway switch machines under imperfect sensing. *IEEE Trans. Instrum. Meas.* **2026**, *75*, 2507215. [CrossRef]
11. 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]
12. Li, X.; Li, G.; Wang, T.; Dong, Z.; Qin, Z.; Chu, F. Knowledge extraction and retrieval-augmented generation for intelligent maintenance of wind power equipment based on graph attention networks. *Chin. J. Mech. Eng.* **2025**, *39*, 100141.
13. Li, S.; Peng, Y.; Shen, Y.; Zhao, S.; Shao, H.; Bin, G.; Guo, Y.; Yang, X.; Fan, C. Rolling bearing fault diagnosis under data imbalance and variable speed based on adaptive clustering weighted oversampling. *Reliab. Eng. Syst. Saf.* **2024**, *244*, 109938. [CrossRef]
14. Karuppasamy, L.; Manivannan, K.; Kosalairaman, T.; Jeya Prasanna, A.; Kaviya, R.; Kaviya, S. A Novel Deep Learning Approach for Automated Rolling Bearing Fault Diagnosis (ARBFD) Using Graph Neural Networks and Physics Informed Deep Learning. In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025)*; SCITEPRESS: Setúbal, Portugal, 2025.
15. Wei, L.; Peng, X.; Cao, Y. Multi-Scale Graph Transformer for Rolling Bearing Fault Diagnosis. *Eksploat. Niezawodn.-Maint. Reliab.* **2025**, *27*, 194779.

16. Jiang, M.; Li, Z.; Ma, J. Research on a Dual-Branch Feature Fusion Method for Rolling Bearing Fault Diagnosis Based on SENet-SwinT. In *Proceedings of the 2025 8th International Conference on Intelligent Robotics and Control Engineering (IRCE)*; IEEE: New York, NY, USA, 2025; pp. 276–282.
17. Wang, H.; Wang, H.; Xie, J. Fault diagnosis of rolling bearings based on HFMD and dual-branch parallel network under acoustic signals. *Sensors* **2025**, *25*, 5338. [CrossRef] [PubMed]
18. Jin, M.; Zhou, Q.; Yan, L.; Dong, W. Bearing Fault Diagnosis across Working Conditions via Temporal-Spectral Attention and Feature Adaptation. *IEEE Sens. J.* **2025**, *25*, 41789–41800. [CrossRef]
19. Zhang, Y.; Zhang, Z.; Xia, Z.; Wen, H.; Dai, M. A Physics Informed Fusion CNN for Bearing Fault Diagnosis Under Variable Rotation Speed. In *Proceedings of the 2024 30th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*; IEEE: New York, NY, USA, 2024; pp. 1–6.
20. Ran, G.; Li, X.; Li, Q.; Wang, T.; Tan, J.; Li, M.; Chu, F. Semantics-Guided Contrastive Pre-training with Cluster-Aware Decision Making for Zero-Shot Bearing Fault Diagnosis. *Sci. China Technol. Sci.* **2026**, *62*, 1–16.
21. Zou, Y.; Zhang, X.; Liu, T.; Zhang, Y.; Li, L.; Zhao, W. Rolling Bearing Fault Diagnosis Based on Cross-Attention Fusion WDCNN and BILSTM. *Comput. Mater. Contin.* **2025**, *83*, 4699–4723. [CrossRef]
22. Li, Y.; Zhang, H.; Zhang, X.; Feng, H. Multi-scale attention-based xLSTM for rolling bearing fault diagnosis. *Meas. Sci. Technol.* **2025**, *36*, 066116. [CrossRef]
23. Su, K.; Kong, X.; Li, X.; Wang, T.; Chu, F. BI-Sandwich: A cross-domain model for fault diagnosis of aircraft engine main shaft bearings. *Struct. Health Monit.* **2026**, *25*, 14759217261426251.
24. Hu, X.; Jiang, C.; Huang, Y.; Peng, D.; Su, H.; He, Y.; Chen, Z. SMNet: A Novel Compositional Generalization Model for Industrial Robot Multi-Joint Fault Diagnosis. *IEEE Internet Things J.* **2026**, *72*, 12345–12356.
25. Zeng, L.; Jian, J.; Chang, X.; Wang, S. A meta-learning method for few-shot bearing fault diagnosis under variable working conditions. *Meas. Sci. Technol.* **2024**, *35*, 056205. [CrossRef]
26. He, D.; He, M.; Yoon, J. Full Ceramic Bearing Fault Diagnosis with Few-Shot Learning Using GPT-2. *Comput. Model. Eng. Sci. (CMES)* **2025**, *143*, 1955–1969. [CrossRef]
27. Randall, R.B.; Antoni, J. Rolling element bearing diagnostics—A tutorial. *Mech. Syst. Signal Process.* **2011**, *25*, 485–520. [CrossRef]
28. Koutsoupakis, J.; Giagopoulos, D. A simulation-enhanced deep learning framework for bearing fault diagnosis using an improved bearing force model. *Mech. Syst. Signal Process.* **2025**, *241*, 113519. [CrossRef]
29. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Trans. Ind. Electron.* **2017**, *65*, 5990–5998.
30. Liang, K.; Zhang, S.; Xiao, W. Bearing Fault Diagnosis Based on Multi-domain Feature Fusion. In *Proceedings of the 2025 44th Chinese Control Conference (CCC)*; IEEE: New York, NY, USA, 2025; pp. 5464–5469.
31. Su, K.; Ran, G.; Li, X.; Wang, T.; Yang, Y.; Chu, F.; Kong, X. Harmonising convolutional networks and Transformer for transfer learning in bearing fault diagnosis. *Nondestruct. Test. Eval.* **2026**, *41*, 1–25. [CrossRef]
32. Hu, M.; He, Y.; Lin, X.; Lu, Z.; Jiang, Z.; Ma, B. Digital twin model of gas turbine and its application in warning of performance fault. *Chin. J. Aeronaut.* **2023**, *36*, 449–470. [CrossRef]
33. Ge, L.; Zhao, D.; Wang, T.; Cui, L. High-dimensional optimized extraction chirplet transform: Algorithm and applications. *ISA Trans.* **2026**, *169*, 520–534. [CrossRef] [PubMed]
34. Jin, Z.; Hu, X.; Wang, H.; Guan, S.; Liu, K.; Fang, Z.; Wang, H.; Wang, X.; Wang, L.; Zhang, Q. Rolling Bearing Fault Diagnosis Model Based on Multi-Scale Depthwise Separable Convolutional Neural Network Integrated with Spatial Attention Mechanism. *Sensors* **2025**, *25*, 4064. [CrossRef] [PubMed]
35. Zhao, B.; Zhang, X.; Zhan, Z.; Wu, Q. Deep multi-scale adversarial network with attention: A novel domain adaptation method for intelligent fault diagnosis. *J. Manuf. Syst.* **2021**, *59*, 565–576. [CrossRef]
36. Liu, Y.; Zheng, W.; Du, Y.; Wang, Y.; Jin, J.; Yu, M. Bearing fault diagnosis based on cross image multi-attention mechanism. *Sci. Rep.* **2025**, *15*, 21100. [CrossRef] [PubMed]
37. Chen, H.; Wei, J.; Huang, H.; Yuan, Y.; Wang, J. Review of imbalanced fault diagnosis technology based on generative adversarial networks. *J. Comput. Des. Eng.* **2024**, *11*, 99–124. [CrossRef]
38. Liu, Z.; Peng, Z. Few-shot bearing fault diagnosis by semi-supervised meta-learning with graph convolutional neural network under variable working conditions. *Measurement* **2025**, *240*, 115402. [CrossRef]
39. Li, J.; Liu, Y.; Li, Q. Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method. *Measurement* **2022**, *189*, 110500. [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.