

Article

A Novel Framework Based on Complementary Views for Fault Diagnosis with Cross-Attention Mechanisms

Xiaorong Liu ¹, Zhonghan Chen ¹, Dongfeng Hu ^{1,2} and Liansong Zong ^{1,3,*}¹ School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China² Shudao Investment Group Co., Ltd., Chengdu 610094, China³ School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

* Correspondence: lszong@my.swjtu.edu.cn; Tel.: +86-028-87729580

Abstract: Bearing fault diagnosis is critical for the reliability and safety of rotating machinery in industrial applications. Traditional fault diagnosis methods and single-view deep learning models often fail to capture the complex, multi-dimensional nature of vibration signals, limiting their effectiveness in accurately identifying faults under varying conditions. To address this, we propose a novel multi-view framework that leverages complementary views—time, frequency, and wavelet domains—of vibration signals for robust fault diagnosis. Our framework integrates a cross-attention mechanism (CAM) that combines self-attention and cross-attention to capture both intra-view and inter-view dependencies, enabling the effective fusion of multi-domain information. By modeling the interactions between different views, the proposed approach enhances feature representation, leading to improved diagnostic accuracy even under noisy industrial conditions. Experimental results on public bearing datasets demonstrate the superior performance of our method compared to state-of-the-art approaches, with significant improvements in robustness and accuracy. This framework provides a promising solution for intelligent fault diagnosis in complex industrial environments.

Keywords: bearing fault diagnosis; multi-view learning; cross-attention



Academic Editors: Suping Xu,
Hengrong Ju, Keyu Liu and
Dah-Jye Lee

Received: 23 November 2024

Revised: 23 January 2025

Accepted: 25 January 2025

Published: 24 February 2025

Citation: Liu, X.; Chen, Z.; Hu, D.; Zong, L. A Novel Framework Based on Complementary Views for Fault Diagnosis with Cross-Attention Mechanisms. *Electronics* **2025**, *14*, 886. <https://doi.org/10.3390/electronics14050886>

Copyright: © 2025 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Bearings are essential components in rotating machinery, playing a crucial role in various industries, including manufacturing, aerospace, energy, and transportation [1,2]. Due to their function of facilitating smooth and stable rotation of mechanical parts, they endure significant stresses over prolonged operational periods. Operating environments for bearings are often harsh and complex, involving conditions such as heavy loads, varying temperatures, and contamination from dust or lubricants. These challenging conditions accelerate the wear and tear process, gradually impacting the performance and structural integrity of bearings [3,4]. Failure of these critical components can have severe consequences, from unexpected shutdowns to costly repairs or replacements, and even catastrophic system failures if not addressed in time [5]. Therefore, accurate and timely fault diagnosis is essential for maintaining reliability, reducing unexpected downtime, and minimizing maintenance costs. Predictive maintenance approaches, using advanced signal processing, machine learning, and data-driven techniques, have been increasingly applied to enhance bearing health monitoring. Early identification of faults allows for planned maintenance and mitigates the risk of severe failures [6].

With the advancements in rolling bearing fault diagnosis research, feature-extraction-based techniques have gained widespread adoption, owing to their relative robustness

against noise in certain conditions [7,8]. Commonly utilized time–frequency analysis methods, such as the short-time Fourier transform (STFT) and wavelet transform, are frequently employed to address fault detection tasks [9,10]. Moreover, approaches like variational mode decomposition (VMD) [11], in conjunction with refined composite multi-scale dispersion entropy (RCMDE) [12], have been applied effectively in diagnosing faults in components such as stators and rotors. Despite their utility, traditional feature extraction techniques are limited in their capacity to provide a comprehensive representation of fault information, often failing to capture the full spectrum of underlying fault characteristics.

Recent advancements in deep learning have paved the way for data-driven approaches in fault diagnosis, enabling models to learn complex patterns from large amounts of sensor data. However, bearing vibration signals are inherently multi-dimensional and contain valuable information across multiple domains, such as the time domain, the frequency domain, and the time–frequency domain. Each domain provides unique information into the condition of the bearing. Many existing fault diagnosis models, particularly single-view approaches, use methods such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) to extract features from the vibration signals [13–15]. CNNs are widely applied to capture spatial features in the time or frequency domains, while LSTMs are effective in modeling temporal dependencies within the time domain. While these single-view methods have shown promising results, they are limited by their focus on only one domain at a time, which can result in incomplete feature extraction and reduced diagnostic accuracy under complex conditions [16].

To capture the complementary characteristics of multi-domain information, some multi-view fusion strategies achieve success [17,18]. Recent studies have explored multi-view fault diagnosis, where features from multiple domains are combined to form a more comprehensive representation of the signal [19,20]. For instance, some methods integrate time and frequency domain features, while others use time–frequency analysis to leverage both temporal and spectral information [21–23]. Although these multi-view approaches improve diagnostic performance by providing a richer representation, they often lack mechanisms to effectively model the interactions between views or within each view. Most existing multi-view models treat each domain independently, without capturing the dependencies and correlations between views, which can lead to incomplete feature fusion and limit the model's ability to accurately diagnose complex faults. Consequently, there remains a need for advanced multi-view frameworks that not only extract features from multiple domains but also integrate information from these domains in a way that fully exploits their inter-view and intra-view dependencies.

To address these challenges, we propose a novel framework for bearing fault diagnosis that leverages complementary views of vibration signals using cross-attention mechanisms. This framework integrates multi-view features from time and wavelet representations, capturing the unique characteristics of each view while enhancing inter-view interactions. The core of this framework is the cross-attention mechanism (CAM), which combines self-attention and cross-attention to effectively model intra-view and inter-view dependencies. By leveraging CAM, the proposed framework can better capture the complex relationships within and between domains, leading to improved fault diagnosis performance.

The contributions of this paper are summarized as follows:

- We propose a novel multi-view framework for bearing fault diagnosis, utilizing complementary information from frequency and wavelet views to improve feature extraction and robustness.
- We introduce a cross-attention mechanism that combines self-attention and cross-attention, enabling the model to capture both intra-view and inter-view dependencies and enhancing its ability to distinguish between fault types.

- We conduct extensive experiments on two public bearing datasets, demonstrating the superior performance of the proposed framework compared to existing fault diagnosis methods, particularly in the presence of noise and varying operational conditions.

The remainder of this paper is organized as follows. Section 2 reviews related work on bearing fault diagnosis and attention mechanisms. Section 3 details the proposed multi-view framework and the cross-attention mechanism. Section 4 presents experimental results and analysis, and Section 5 concludes the paper with a discussion on future research directions.

2. Related Work

Data-driven methods have gained significant traction in bearing fault diagnosis, with deep learning models such as ResNet [24], convolutional neural networks (CNNs) [13,25], long short-term memory networks (LSTMs) [14], and attention mechanisms (AMs) [26] demonstrating promising results. Classic CNN architectures are extensively used due to their strong ability to extract spatial features from vibration signals, making them well-suited for identifying fault patterns in complex data. On the other hand, LSTMs excel at capturing temporal dependencies in sequential data, enabling effective analysis of time-series information commonly found in vibration signals [27]. More recently, attention models have been introduced into fault diagnosis tasks, leveraging their powerful attention mechanisms to selectively focus on the most relevant parts of the input data, thereby improving feature representation.

Moreover, improved versions of these networks, which incorporate advanced architectural modifications and hybrid approaches, have significantly enhanced diagnostic accuracy. Hou et al. [28] introduced an attention-driven multi-feature parallel fusion model for rolling bearing fault diagnosis, leveraging the capabilities of transformer networks. In a parallel effort, Fu et al. [15] developed a fault diagnosis framework that integrates convolutional neural networks (CNN) with long short-term memory (LSTM) networks to simultaneously capture spatial and temporal features. Meanwhile, Song et al. [27] proposed an enhanced approach combining optimized CNNs with Bi-directional LSTM (BiLSTM) to address bearing fault diagnosis under diverse operating conditions. Peng et al. proposed a Transformer convolutional network (TCN) based on transfer learning, which addresses the challenge of achieving high-precision fault diagnosis under data scarcity by splitting signal data into patches, utilizing a Transformer encoder, and decoding and classifying patterns with a convolutional neural network [29]. However, these models often focus on single-domain features and lack mechanisms to integrate multi-domain information. This limitation restricts their ability to fully exploit the complementary information from different feature spaces, making them less robust in complex, variable environments [23].

Recent studies in multi-view fault diagnosis attempt to address this limitation by combining features from multiple domains, such as time, frequency, and time–frequency representations, to improve diagnostic performance [16]. Multi-view learning approaches aim to leverage the complementary information across views, thus providing a more comprehensive representation of the vibration signal. However, most of these methods primarily focus on feature extraction in separate views and do not adequately consider the interaction of information both between and within views. Hosameldin et al. [30] developed the CFCNet model, which utilizes feature concatenation for fusing multi-view data during the fusion stage. In a different approach, Peng et al. [22] designed a multi-view feature construction technique leveraging genetic programming (GP) combined with ensemble learning, specifically tailored for rolling bearing fault diagnosis. Meanwhile, canonical correlation analysis (CCA), a well-established method, focuses on identifying correlations between two variables by projecting them into a shared subspace. Building

on this, Jiang et al. [21] introduced a harmonic correlation feature learning framework to enhance bearing fault diagnosis.

Despite the advancements in multi-view fault diagnosis techniques, significant challenges remain in effectively integrating information from multiple views. For instance, feature concatenation approaches often neglect the nuanced semantic relationships between features, and CCA-based methods struggle to model high-level interactions within multi-view data [31]. Consequently, these methods often fail to fully exploit the dual aspects of consensus and complementarity inherent in multi-view representations. Moreover, existing models often fail to capture the complex interdependencies across diverse views, which hinders their ability to accurately diagnose intricate fault patterns in real-world industrial scenarios. This limitation underscores the need for more sophisticated frameworks capable of addressing these shortcomings.

To overcome these specific limitations, we introduce a novel framework for bearing fault diagnosis that leverages complementary views enhanced by the CAM. Distinct from existing approaches, our method adeptly captures both intra-view and inter-view dependencies, significantly improving the interaction and integration of information across multiple domains. The CAM employs self-attention to model relationships within individual views (e.g., frequency and wavelet features) and cross-attention to establish dependencies between different views, culminating in a unified and robust multi-view representation.

3. The Proposed Model

3.1. Problem Formulation

Given a vibration signal $\mathbf{x}$, we perform signal transformations to generate a multi-view signal set $S = (\mathbf{f}, \mathbf{w})$, where $\mathbf{f}$ and $\mathbf{w}$ denote the frequency domain and wavelet domain features, respectively. The goal of our model is to fuse view-specific information from these two inputs using the CAM, creating a unified multi-view representation that enables precise fault category prediction.

We frame this task as a classification problem, learning a mapping function, $\mathcal{F} : S \rightarrow y$, that associates each multi-view representation, $s \in S$, with its corresponding fault category, y . The CAM employs self-attention and cross-attention mechanisms to enrich feature interactions across views, producing a robust and integrated representation for accurate fault diagnosis.

3.2. Overview

The proposed model leverages multi-view data from vibration signals for fault diagnosis, as shown in Figure 1, incorporating frequency and wavelet domain representations. The model is structured into three main components: multi-view data representation, cross-attention module (CAM), and classification.

3.3. Multi-View Data Representation

Given a vibration signal, $\mathbf{x} \in \mathcal{R}^{1 \times n}$, we transform it to obtain a set of multi-view representations, $S = \{\mathbf{f}, \mathbf{w}\}$, where $\mathbf{f}$ and $\mathbf{w}$ represent frequency and wavelet domain data, respectively.

The frequency domain data $\mathbf{f}$ is obtained by applying the Fast Fourier Transform (FFT) to the input signal $\mathbf{x}$:

$$\mathbf{f} = \mathcal{F}(\mathbf{x}) = \int_{-\infty}^{\infty} \mathbf{x}(t) e^{-j2\pi ft} dt \quad (1)$$

where $\mathcal{F}$ denotes the FFT operation. This transformation captures periodic information in the signal.

The frequency view encoder E_f uses a BiLSTM (bidirectional long short-term memory) network to extract temporal dependencies in the frequency domain features. Given the input sequence $\mathbf{f}$, the BiLSTM computes the latent feature representation $\mathbf{h}_F$:

$$\mathbf{h}_F = \text{BiLSTM}(\mathbf{f}) \quad (2)$$

The wavelet domain feature $\mathbf{w}$ is obtained by applying the continuous wavelet transform (CWT) to $\mathbf{x}$:

$$\mathbf{w} = \mathcal{W}(\mathbf{x}) = \int_{-\infty}^{\infty} \mathbf{x}(t) \psi^* \left(\frac{t - \tau}{s} \right) dt \quad (3)$$

where $\mathcal{W}$ represents the CWT operation, ψ is the wavelet function, s is the scale, and τ is the translation parameter. This time–frequency representation is encoded through a convolutional neural network (CNN), denoted as E_w , which captures local patterns and spatial hierarchies in the wavelet domain. The CNN encoder computes the latent feature $\mathbf{h}_W$:

$$\mathbf{h}_W = \text{CNN}(\mathbf{w}) \quad (4)$$

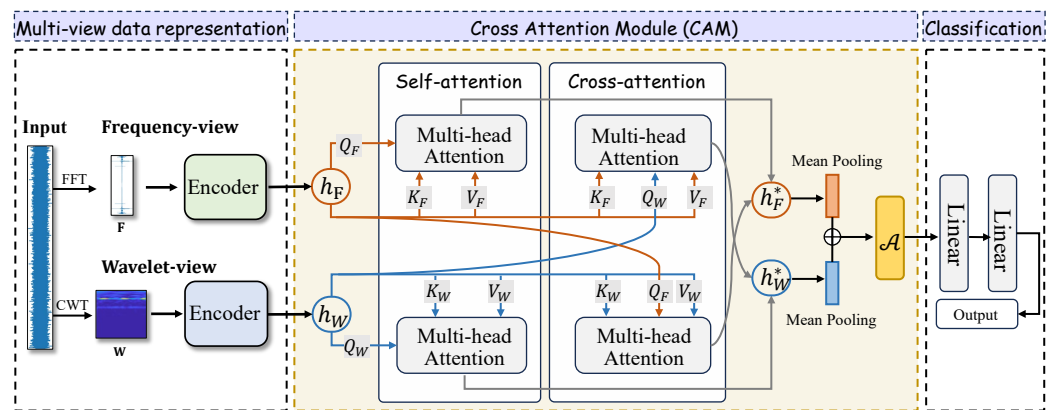


Figure 1. Illustration of the proposed cross-attention module (CAM) network, integrating frequency and wavelet features through self-attention and cross-attention for fault classification.

3.4. Cross-Attention Module (CAM)

The CAM allows the model to capture interactions not just within a single view (self-attention), but across multiple views. The cross-attention is applied to two types of feature representations: frequency features h_F and wavelet features h_W . This inter-view communication helps the model learn from complementary information that each view provides.

Self-attention is applied separately to h_F and h_W to capture intra-view dependencies. For each view $h \in \{h_F, h_W\}$, the self-attention mechanism operates as follows:

$$\text{Self-Attention}(h) = \text{Softmax} \left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d}} \right) \quad (5)$$

where $\mathbf{Q}_h = W_q h$, $\mathbf{K}_h = W_k h$, and $V_h = W_v h$ are the query, key, and value projections of h , respectively, and W_q , W_k , and W_v are learnable weight matrices. This process yields self-attended features h_F^* and h_W^* .

To capture inter-view interactions, cross-attention is performed between the frequency and wavelet features. Cross-attention allows each view to attend to the other view's features. For the frequency view attention on wavelet features:

$$\text{Cross-Attention}_{FW} = \text{Softmax}\left(\frac{\mathbf{Q}_F \mathbf{K}_W^T}{\sqrt{d}}\right) \mathbf{V}_W \quad (6)$$

Similarly, for the wavelet view attention on frequency features:

$$\text{Cross-Attention}_{WF} = \text{Softmax}\left(\frac{\mathbf{Q}_W \mathbf{K}_F^T}{\sqrt{d}}\right) \mathbf{V}_F \quad (7)$$

where $\mathbf{Q}_F, \mathbf{K}_W, \mathbf{V}_W$ (for cross-attention between frequency and wavelet views) and $\mathbf{Q}_W, \mathbf{K}_F, \mathbf{V}_F$ (for cross-attention between wavelet and frequency views) are query, key, and value projections based on the features $\mathbf{h}_F$ and $\mathbf{h}_W$, respectively. The output is an integrated multi-view representation combining information across both views.

After self-attention and cross-attention, mean pooling is applied to the resulting attention outputs $\mathbf{h}_F^*$ and $\mathbf{h}_W^*$ to generate a pooled representation for each view. The combined representation, $\mathcal{A}$, is obtained by concatenating these pooled features:

$$\mathbf{h}_F^* = \frac{1}{2} [\text{MHA}(\mathbf{Q}_F, \mathbf{K}_F, \mathbf{V}_F) + \text{MHA}(\mathbf{Q}_F, \mathbf{K}_W, \mathbf{V}_W)] \quad (8)$$

$$\mathbf{h}_W^* = \frac{1}{2} [\text{MHA}(\mathbf{Q}_W, \mathbf{K}_W, \mathbf{V}_W) + \text{MHA}(\mathbf{Q}_W, \mathbf{K}_F, \mathbf{V}_F)] \quad (9)$$

$$\mathcal{A} = \mathbf{h}_F^* \oplus \mathbf{h}_W^* \quad (10)$$

where $\oplus$ denotes the concatenation operation. This fused feature $\mathcal{A}$ serves as the unified multi-view representation for fault diagnosis.

3.5. Classification Layer

The fused representation $\mathcal{A}$ is passed through a fully connected layer for classification. The objective of the model is to learn a function $\mathcal{F} : S \rightarrow y$ that maps each multi-view representation $s \in S$ to a possible fault category y . The final classification output $\hat{y}$ is given by:

$$\hat{y} = \text{Softmax}(W_c \mathcal{A} + b_c) \quad (11)$$

where W_c and b_c are the weights and biases of the classification layer, and Softmax is used to produce probabilities for each fault category.

3.6. Loss Function

The model is trained using a cross-entropy loss function to minimize the difference between the predicted labels $\hat{y}$ and the true labels y :

$$\mathcal{L} = - \sum_{i=1}^N y_i \log(\hat{y}_i) \quad (12)$$

where N is the total number of training samples, y_i is the true label for the i -th sample, and $\hat{y}_i$ is the predicted probability for the correct category.

This proposed model enables effective feature integration from multi-view representations, allowing for accurate fault classification while preserving the unique characteristics of each view. The Algorithm 1 summarizes the main steps of the proposed multi-view

cross-attention model for fault diagnosis. This includes feature extraction using BiLSTM and CNN encoders, attention-based fusion, and classification.

Algorithm 1: Multi-view cross-attention model for fault diagnosis.

Input: Vibration signal $\mathbf{x} \in \mathcal{R}^{1 \times n}$, number of classes C , learning rate η , number of training epochs T .

Output: Trained model parameters θ .

for all epochs do

for all batches in S do

$\mathbf{f} = \mathcal{F}(\mathbf{x});$

 ► Equation (1)

$h_F \leftarrow \text{BiLSTM}(\mathbf{f});$

 ► Equation (2)

$\mathbf{w} = \mathcal{W}(\mathbf{x});$

 ► Equation (3)

$h_W \leftarrow \text{CNN}(\mathbf{w});$

 ► Equation (4)

$(\mathbf{h}_F^*, \mathbf{h}_W^*) \leftarrow \text{Self-Attention}(\mathbf{h}_F, \mathbf{h}_W);$

 ► Equation (5)

$\mathbf{h}_F^* \leftarrow \text{Cross-Attention}_{FW}(\mathbf{h}_F, \mathbf{h}_W);$

 ► Equation (6)

$\mathbf{h}_W^* \leftarrow \text{Cross-Attention}_{WF}(\mathbf{h}_F, \mathbf{h}_W);$

 ► Equation (7)

$\mathcal{A} \leftarrow \text{Concat}(\text{MenaPool}(h_F^*, h_W^*));$

 ► Equations (8)–(10)

$\hat{y} \leftarrow \text{Softmax}(W_c h^* + b_c);$

$\mathcal{L} \leftarrow -\sum_{i=1}^N y_i \log(\hat{y}_i);$

$\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L};$

end

end

Return parameters θ .

4. Experiment

4.1. Datasets

To evaluate the effectiveness of the proposed multi-view cross-attention module (CAM) framework, we conduct experiments on two public datasets, including CWRU and SEU_bearing.

CWRU Bearing Dataset: The Case Western Reserve University (CWRU) Bearing Data Center provides the CWRU dataset, with the experimental setup comprising an electric motor, a torque sensor, a dynamometer, and two accelerometers mounted on the driving-end and fan-end housings to capture vibration signals. The data were collected at sampling frequencies of 12 kHz or 48 kHz under four different motor load conditions. Faults are categorized into three types: inner ring faults, rolling element faults, and outer ring faults. These have diameters of 0.007, 0.014, and 0.021 inches. In this study, we utilize driver-side data sampled at 12 kHz. Table 1 further classifies the dataset into ten distinct states, including one healthy condition and nine fault conditions, based on fault severity and type.

The SEU_bearing dataset, derived from the gearbox dataset provided by Southeast University (SEU), serves as a valuable resource for bearing fault diagnosis research. The experimental setup includes a motor, a brake, a controller module, and two test transmissions. The dataset includes vibration signals collected from eight distinct signal channels. For this study, the second vibration signal channel was utilized. Acceleration sensors were strategically installed on the drive motor, the planetary gearbox, and the parallel gearbox to capture vibration signals. The experiments were conducted under two operating conditions, with the rotational speed and load set to 20 Hz-0 V and 30 Hz-2 V, respectively. For each operating condition, the dataset contains five fault types: cracked inner race, cracked comb, cracked ball, cracked outer race, and healthy state. Detailed descriptions of these fault types and experimental conditions are provided in Table 2.

Table 1. Detailed description of CWRU datasets.

Fault Type	Description (1791 rpm—0 HP)	Label	N_{train}	N_{test}
Health state	No fault	0	240	48
	0.007-inch inner ring fault	1	120	24
Inner ring	0.014-inch inner ring fault	2	120	24
	0.021-inch inner ring fault	3	120	24
	0.007-inch rolling element	4	120	24
Rolling	0.014-inch rolling element	5	120	24
	0.021-inch rolling element	6	120	24
	0.007-inch outer ring fault	7	120	24
Outer ring	0.014-inch outer ring fault	8	120	24
	0.021-inch outer ring fault	9	120	24

Table 2. The SEU_bearing dataset includes experimental data collected under two operating conditions: 20_0 and 30_2, where the notation represents the rotating speed (RS) and load condition (LC) settings of 20 Hz-0 V and 30 Hz-2 V, respectively.

Load	Fault Type	Description	Label	N_{train}	N_{test}
30_2	inner	Crack occurs in the inner ring	0	819	205
30_2	comb	Crack occurs in the inner and outer ring	1	819	205
20_0	ball	Crack occurs in the ball	2	819	206
30_2	outer	Crack occurs in the outer ring	3	819	205
30_2	health	No fault	4	819	205
20_0	outer	Crack occurs in the outer ring	5	819	205
30_2	ball	Crack occurs in the ball	6	819	205
20_0	health	No fault	7	819	205
20_0	inner	Crack occurs in the inner ring	8	819	205
20_0	comb	Crack occurs in the inner and outer ring	9	819	205

4.2. Baselines

- CNN [5]: A classic fault diagnosis network that employs one-dimensional (1D) convolutional neural network to extract local features.
- CNN-LSTM [32]: It integrates multi-scale convolutional neural networks and long short-term memory, and has been designed for rolling bearing fault diagnosis.
- MCCA [33]: A multi-view learning framework utilizing canonical correlation analysis (CCA) is introduced to optimize the correlation between vibration and current features. To ensure a fair comparison, its original multi-level feature extraction mechanism has been replaced with the view-specific encoder employed in our approach.
- BAN [34]: This multi-view fusion method employs bilinear attention mechanisms to effectively model complex interactions between features by enabling each element of one feature vector to interact with every element of another.
- Concatenation (Cat): A simple multi-view fusion strategy where high-level representations from different views are directly concatenated. Fault type identification is subsequently performed through a linear transformation.
- Dot: A multi-view fusion approach that performs element-wise multiplication of high-level representations across different views, followed by fault type identification through a linear transformation layer.
- Add: A multi-view fusion strategy that integrates high-level representations from different views through element-wise addition, with subsequent fault type identification achieved via a linear transformation.

4.3. Experiment Setting

4.3.1. Normalization

Outlier sample data can significantly increase training time and hinder model convergence. To address this issue, normalization is applied to the preprocessed data before training. The normalization process used in this study is defined as:

$$\mathbf{x}' = \frac{\mathbf{x} - \mu}{\sigma} \quad (13)$$

where $\mathbf{x}$ represents the original data, μ is the mean of the dataset, σ is the standard deviation, and $\mathbf{x}'$ is the normalized data. This transformation standardizes the data to have a mean of zero and a variance of one, thereby facilitating faster model convergence and improving training efficiency.

4.3.2. Data Split

Data partitioning plays a crucial role in intelligent diagnostics, where random splitting is frequently adopted. However, this method carries substantial risks, as potential data overlap during preparation may lead to test set leakage. Such leakage undermines the reliability and validity of the classification algorithm's evaluation, leading to overly optimistic performance estimates [5]. For the order split, a sliding window is applied to segment the vibration signal, ensuring no overlap between the data samples. Each data sample consists of 1024 points. The first 80% of the time series is used as the training set, while the last 20% is allocated as the testing set.

4.3.3. Evaluation Metrics

Three evaluation metrics are employed to assess the performance of the proposed model: accuracy (ACC), precision (P), and recall (R). Accuracy quantifies the proportion of correctly classified samples out of the total samples, providing a comprehensive measure of the model's classification performance. Precision evaluates the proportion of correctly classified positive samples among all samples predicted as positive, reflecting the model's reliability in identifying positive instances. Recall, in contrast, measures the proportion of correctly classified positive samples relative to the total actual positive samples, emphasizing the model's sensitivity to positive cases. The corresponding formulas are defined as follows.

Accuracy (ACC) represents the proportion of correctly classified samples among all samples, providing an overall measure of the model's performance.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Precision (P) measures the proportion of true positives among all samples classified as positive, indicating the model's reliability in identifying positives.

$$P = \frac{TP}{TP + FP} \quad (15)$$

Recall (R) represents the proportion of true positives among all actual positive samples, highlighting the model's sensitivity to positive samples.

$$R = \frac{TP}{TP + FN} \quad (16)$$

4.3.4. Parameter Setting

The Adam optimizer is employed during the model training process, with a fixed learning rate of 0.001 and a batch size of 16 for all experiments. The experiments are conducted on a Linux system using PyTorch 2.1, executed on a machine equipped with a GeForce RTX 3090 graphics card (manufactured by NVIDIA Corporation, headquartered in Santa Clara, CA, USA) and 16 GB of RAM (manufactured by Micron Technology, Inc., headquartered in Boise, ID, USA). Other critical hyperparameters, such as the learning rate and batch size, are optimized through a grid search strategy on a limited subset of the validation dataset.

5. Results

5.1. Compared Results with Different Methods

Table 3 presents a comparative analysis of the diagnostic performance of various methods across two datasets: CWRU and SEU_bearing. The evaluation metrics include accuracy (Acc), precision (P), and recall (R), which are critical for assessing the effectiveness of fault diagnosis models. The results indicate that the proposed CAM module consistently outperforms the other methods across all metrics, achieving near-perfect or perfect performance on both datasets.

Table 3. Comparative evaluation of diagnostic performance among various approaches.

Model	CWRU			SEU_Bearing		
	Acc	P	R	Acc	P	R
CNN	95.56	96.24	95.60	90.76	90.28	90.59
MCNN-LSTM	98.86	98.73	98.77	98.36	98.42	98.44
MCCA	99.23	99.29	99.24	99.36	99.27	99.32
BAN	99.38	99.41	99.37	99.49	99.55	99.27
CAM (Cat)	98.54	98.87	98.72	98.78	98.15	98.34
CAM (Dot)	97.42	97.79	97.82	97.48	97.75	97.49
CAM (Add)	98.47	98.44	98.41	98.53	98.80	98.54
CAM	100	100	100	99.98	99.98	99.98

The CAM model outperforms other methods, achieving 100% accuracy, precision, and recall on the CWRU dataset and near-perfect results on the SEU_bearing dataset. This highlights its ability to effectively capture complex multi-view feature correlations and adapt to diverse datasets. In comparison, methods like MCCA and BAN show lower performance, indicating that simpler correlation-based techniques or bilinear attention mechanisms fall short in optimal feature integration. Furthermore, CAM significantly surpasses single-view approaches such as CNN, which achieved 95.56% Accuracy on the CWRU dataset and 90.76% on the SEU_bearing dataset, demonstrating the clear advantage of multi-view feature utilization in fault diagnosis tasks.

Variants of the CAM model employing simpler fusion strategies—concatenation (Cat), Dot, and Add—demonstrate moderate performance improvements over traditional methods such as MCCA and BAN. However, these variants are unable to match the effectiveness of the original CAM model. This observation indicates that while basic fusion techniques can utilize multi-view features to some extent, they lack the ability to capture the intricate interactions between features from different views. In contrast, the CAM effectively models these complex dependencies, leading to substantial performance gains.

The CAM model exhibits robust generalization capabilities, delivering high performance across two datasets with markedly different characteristics. Specifically, the CWRU dataset encompasses rolling bearing faults, whereas the SEU_bearing dataset focuses on gearbox faults. This consistency in performance underscores the CAM model's adaptability and robustness to diverse fault scenarios and operating conditions.

5.2. Performance Comparison

The CAM model's unique design provides several advantages over other methods, as reflected in the experimental results.

The CAM model achieves an accuracy of 100% on the CWRU dataset and 99.98% on the SEU_bearing dataset, outperforming other methods such as MCCA (99.23% and 99.36%) and BAN (99.38% and 99.49%). This substantial improvement is attributed to CAM's ability to construct comprehensive multi-view representations and leverage attention mechanisms, allowing it to capture a broader and more detailed set of diagnostic features.

The CAM model achieves perfect precision and recall on the CWRU dataset and near-perfect scores on the SEU_bearing dataset, demonstrating its exceptional ability to accurately identify fault patterns while avoiding false positives and omissions of relevant features. In comparison, methods such as CAM(Cat), CAM(Dot), and CAM(Add) achieve precision and recall values in the range of 97–98%, highlighting the superior diagnostic capability of CAM's attention-based architecture, which effectively captures and integrates critical features across multiple views.

The CAM model facilitates effective mutual learning between frequency view and wavelet view representations. By integrating information across these views, the model enhances its ability to detect subtle and complex patterns that might otherwise be overlooked when each view is analyzed separately. This advanced interaction between features is crucial to the CAM model's exceptional diagnostic performance.

Overall, the CAM model, distinguished by its multi-view data representation, self-attention mechanism for capturing intra-view dependencies, and cross-attention for integrating inter-view information, has demonstrated exceptional effectiveness in fault diagnosis. Experimental results demonstrate that CAM achieves superior accuracy, precision, and recall compared to other methods, establishing it as a robust and reliable solution for complex diagnostic tasks.

5.3. Performance of CAM

To systematically examine the diagnostic performance for each fault category, we utilized confusion matrices to analyze the model's classification results. As depicted in Figure 2, the confusion matrices provide a detailed representation of the diagnostic and classification outcomes across various datasets, capturing both accurate classifications and misclassifications. For the CWRU dataset, the model achieved high classification accuracy across all fault categories, with all true labels correctly identified and no misclassification errors observed. In contrast, as shown in Figure 3, a minor misclassification was noted for the SEU_bearing dataset, where fault category 9 was misidentified as category 2. These results indicate that the CAM model exhibits a potential for fault diagnosis, especially with the datasets examined. However, the observed misclassification in the SEU_bearing dataset suggests an area for improvement, indicating that further optimization or fine-tuning may enhance the model's capacity to distinguish closely related fault types, particularly in more complex diagnostic scenarios.

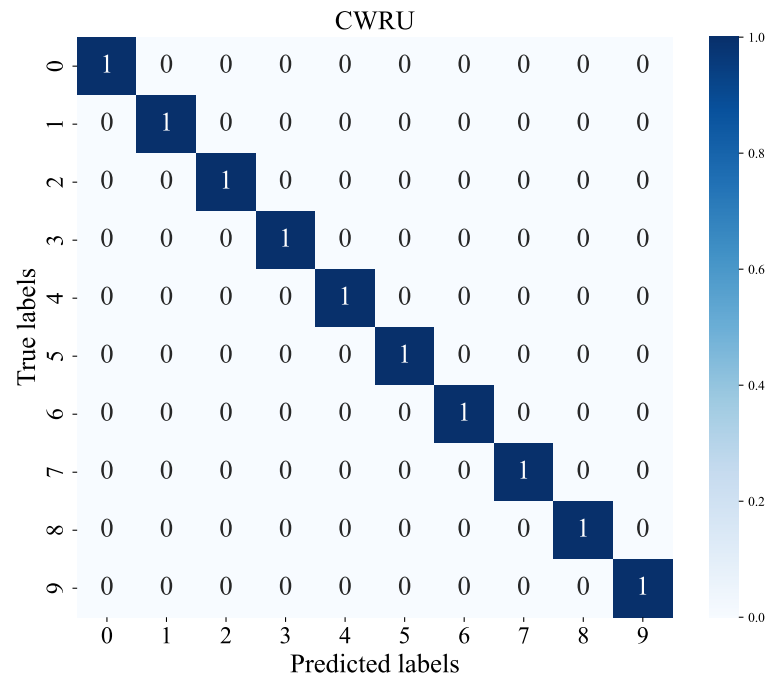


Figure 2. The confusion matrices of CAM on CWRU.

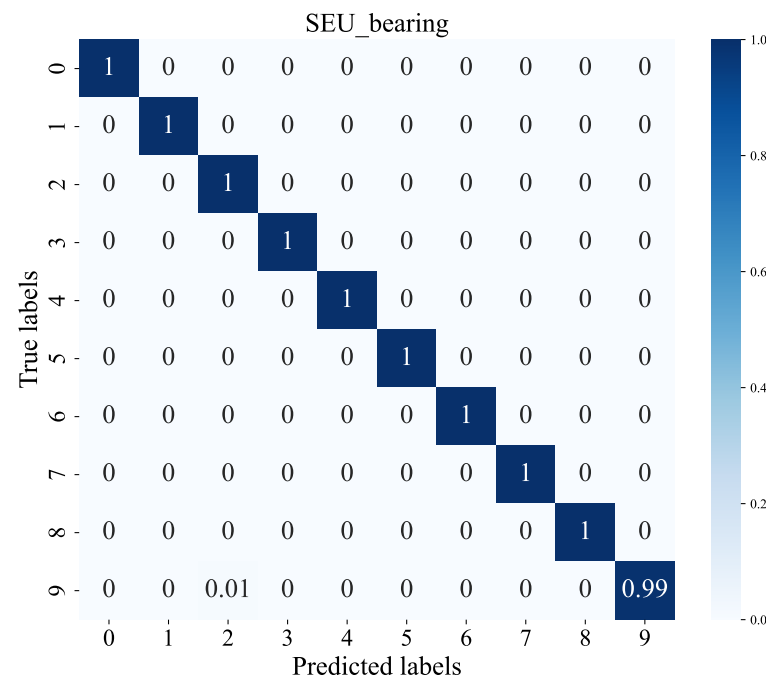


Figure 3. The confusion matrices of CAM on SEU_bearing.

5.4. Impact of Noise on Model Performance

In industrial settings, vibration signals are often corrupted by noise due to complex operating conditions. To assess model robustness against noise interference, artificial noise is added to the original signals, and performance is evaluated across different signal-to-noise ratios (SNR), calculated as:

$$\text{SNR} = 10 \cdot \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \quad (17)$$

where P_{signal} represents the power of the original signal, and P_{noise} represents the power of the noise. Lower SNR values indicate higher levels of noise interference.

To assess the noise resilience of the models, experiments are conducted across signal-to-noise ratio (SNR) values ranging from -2 dB to 6 dB. The experimental results, illustrated in the Figures 4 and 5, demonstrate the model's performance under varying noise conditions. Specifically, the graphs depict the accuracy of the models across SNR levels of -2 dB, 0 dB, 2 dB, 4 dB, and 6 dB for two datasets: CWRU and SEU_bearing. The evaluated models include the Dot, Cat, Add, and CAM mechanisms, enabling a comprehensive comparison of their robustness in noisy environments.

For both datasets, the CAM mechanism consistently achieves the highest accuracy across all SNR levels, demonstrating its robustness to noise. Notably, CAM maintains relatively stable and high accuracy even under challenging conditions, such as an SNR of -2 dB, highlighting its superior capability to manage significant noise interference compared to other mechanisms.

The performance of the Dot, Cat, and Add mechanisms varies significantly with changes in noise levels. As the SNR increases, all mechanisms demonstrate improved accuracy, which is expected since a higher SNR corresponds to a cleaner signal with reduced noise interference. However, the CAM mechanism exhibits a notably stronger performance at lower SNR levels, highlighting its superior noise resistance and robustness under challenging conditions.

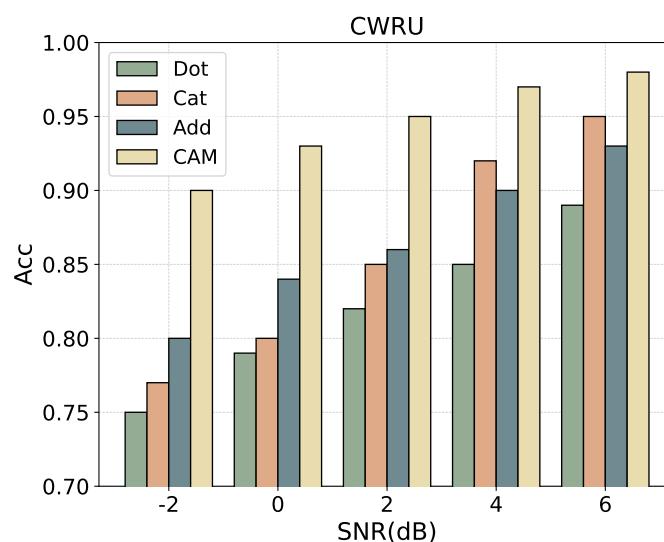


Figure 4. Accuracy of CWRU under different SNR levels.

At SNR levels of 4 dB and 6 dB, the performance gap between CAM and the other mechanisms narrows slightly, indicating that the advantages of CAM are most prominent under high-noise conditions. The Dot and Cat mechanisms exhibit comparable performance across both datasets, while the Add mechanism demonstrates competitive results but generally falls short of CAM's superior accuracy.

Overall, the experimental results identify CAM as the most effective mechanism for fault diagnosis in noisy environments, particularly under conditions where noise is a significant factor. These findings suggest that CAM holds strong potential for real-world applications where noise interference is prevalent. In contrast, the Dot, Cat, and Add mechanisms may be more suitable for scenarios characterized by cleaner data and higher signal quality.

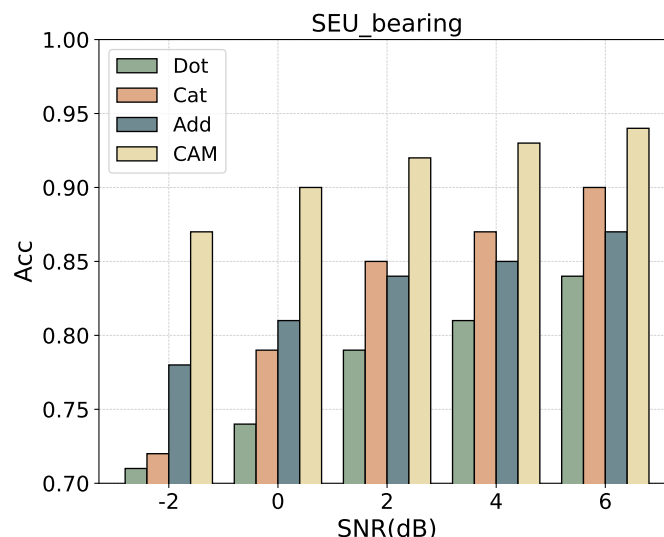


Figure 5. Accuracy of SEU_bearing under different SNR levels.

6. Conclusions

In this paper, we proposed a novel multi-view framework for bearing fault diagnosis that leverages complementary information from the time, frequency, and time–frequency domains through the CAM. By integrating self-attention and cross-attention mechanisms, the proposed approach effectively captures both intra-view and inter-view dependencies, enabling a comprehensive fusion of multi-domain features. This design facilitates improved diagnostic accuracy and robustness, particularly in noisy and variable industrial environments. Experimental evaluations on benchmark bearing datasets demonstrate that the proposed framework consistently outperforms state-of-the-art fault diagnosis methods, achieving superior accuracy and resilience to noise. These results underscore the significance of multi-view learning and cross-attention in extracting and integrating complementary information, which is essential for accurate and reliable fault diagnosis in practical industrial applications.

Future work will focus on two key directions: (1) extending the framework to incorporate additional views or sensor modalities to further enhance diagnostic capabilities; (2) investigating the interpretability of the cross-attention mechanism to gain deeper insights into the contributions of different domains to fault diagnosis. Specifically, we plan to develop visualization techniques and quantitative metrics to analyze how the attention weights are distributed across different features and domains under various fault conditions. This interpretability research could not only improve model design but also enhance its reliability and applicability in industrial environments by providing transparent decision-making processes and enabling domain experts to validate the model’s reasoning.

Author Contributions: Conceptualization, X.L. and D.H.; methodology, X.L.; software, Z.C.; writing—original draft preparation, X.L.; writing—review and editing, X.L.; visualization, X.L.; supervision, L.Z.; project administration, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Dongfeng Hu was employed by the company Shudao Investment Group 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

- Lu, Y.; Wang, Z.; Zhu, D.; Gao, Q.; Sun, D. Bearing Fault Diagnosis Based on Clustering and Sparse Representation in Frequency Domain. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–14. [\[CrossRef\]](#)
- He, C.; Shi, H.; Li, R.; Li, J.; Yu, Z. Interpretable modulated differentiable STFT and physics-informed balanced spectrum metric for freight train wheelset bearing cross-machine transfer fault diagnosis under speed fluctuations. *Adv. Eng. Inform.* **2024**, *62*, 102568. [\[CrossRef\]](#)
- Peng, B.; Bi, Y.; Xue, B.; Zhang, M.; Wan, S. A Survey on Fault Diagnosis of Rolling Bearings. *Algorithms* **2022**, *15*, 347. [\[CrossRef\]](#)
- He, C.; Shi, H.; Si, J.; Li, J. Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings. *J. Manuf. Syst.* **2023**, *70*, 579–592. [\[CrossRef\]](#)
- Zhao, Z.; Li, T.; Wu, J.; Sun, C.; Wang, S.; Yan, R.; Chen, X. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA Trans.* **2020**, *107*, 224–255. [\[CrossRef\]](#) [\[PubMed\]](#)
- 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. [\[CrossRef\]](#)
- Qian, D.; Liu, K.; Zhang, S.; Yang, X. Semi-supervised feature selection by minimum neighborhood redundancy and maximum neighborhood relevancy. *Appl. Intell.* **2024**, *54*, 7750–7764. [\[CrossRef\]](#)
- Qian, D.; Liu, K.; Wang, J.; Zhang, S.; Yang, X. Attribute reduction based on directional semi-neighborhood rough set. *Int. J. Mach. Learn. Cyber.* **2024**. [\[CrossRef\]](#)
- Satpathi, K.; Yeap, Y.M.; Ukil, A.; Geddada, N. Short-Time Fourier Transform Based Transient Analysis of VSC Interfaced Point-to-Point DC System. *IEEE Trans. Ind. Electron.* **2018**, *65*, 4080–4091. [\[CrossRef\]](#)
- Yan, R.; Gao, R.X.; Chen, X. Wavelets for fault diagnosis of rotary machines: A review with applications. *Signal Process.* **2014**, *96*, 1–15. [\[CrossRef\]](#)
- Dragomiretskiy, K.; Zosso, D. Variational Mode Decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [\[CrossRef\]](#)
- Azami, H.; Rostaghi, M.; Abásolo, D.E.; Escudero, J. Refined Composite Multiscale Dispersion Entropy and its Application to Biomedical Signals. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2872–2879.
- Ruan, D.; Wang, J.; Yan, J.; Gühmann, C. CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis. *Adv. Eng. Inform.* **2023**, *55*, 101877. [\[CrossRef\]](#)
- You, K.; Puzhou, W.; Gu, Y. Toward Efficient and Interpretative Rolling Bearing Fault Diagnosis via Quadratic Neural Network With Bi-LSTM. *IEEE Internet Things J.* **2024**, *11*, 23002–23019.
- Fu, G.; Wei, Q.; Yang, Y. Bearing fault diagnosis with parallel CNN and LSTM. *Math. Biosci. Eng. MBE* **2024**, *21*, 2385–2406. [\[CrossRef\]](#)
- Liu, X.; Wang, J.; Meng, S.; Qiu, X.; Zhao, G. Multi-view rotating machinery fault diagnosis with adaptive co-attention fusion network. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106138. [\[CrossRef\]](#)
- Wang, J.; Yang, Y.; Liu, K.; Zhu, Z.; Liu, X. M3S: Scene Graph Driven Multi-Granularity Multi-Task Learning for Multi-Modal NER. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 111–120. [\[CrossRef\]](#)
- Wang, J.; Yang, Y.; Jiang, Y.; Ma, M.; Xie, Z.; Li, T. Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. *Inf. Fusion* **2024**, *103*, 102132. [\[CrossRef\]](#)
- Yu, Y.; He, Y.; Karimi, H.R.; Gelman, L.; Çetin, A.E. A two-stage importance-aware subgraph convolutional network based on multi-source sensors for cross-domain fault diagnosis. *Neural Netw.* **2024**, *179*, 106518. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, J.; Yang, Y.; Liu, K.; Xie, Z.; Zhang, F.; Li, T. CiteNet: Cross-modal incongruity perception network for multimodal sentiment prediction. *Knowl. Based Syst.* **2024**, *295*, 111848. [\[CrossRef\]](#)
- Zhu, Y.; Peng, T.; Su, S.; Li, C.; Lu, S. Discriminant Sensitive Harmonic Correlation Fault Diagnosis for Rolling Bearing. In Proceedings of the 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 29–31 October 2021; pp. 52–55.
- Peng, B.; Bi, Y.; Xue, B.; Zhang, M.; Wan, S. Multi-View Feature Construction Using Genetic Programming for Rolling Bearing Fault Diagnosis [Application Notes]. *IEEE Comput. Intell. Mag.* **2021**, *16*, 79–94. [\[CrossRef\]](#)
- Lu, Y.; Zhu, D.; Zhao, D.; Zhu, D.; Sun, D.; Gao, Q. A Novel Bearing Fault Diagnosis Algorithm Based on Adaptive Weighted Multiview Learning. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–13. [\[CrossRef\]](#)
- Zhang, K.; Tang, B.; Deng, L.; Liu, X. A hybrid attention improved ResNet based fault diagnosis method of wind turbines gearbox. *Measurement* **2021**, *179*, 109491. [\[CrossRef\]](#)
- He, C.; Shi, H.; Liu, X.; Li, J. Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis. *Knowl. Based Syst.* **2024**, *288*, 111499. [\[CrossRef\]](#)
- Cui, L.; Tian, X.; Wei, Q.; Liu, Y. A self-attention based contrastive learning method for bearing fault diagnosis. *Expert Syst. Appl.* **2024**, *238*, 121645. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)

28. Hou, Y.; Wang, J.; Chen, Z.; Ma, J.; Li, T. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved Transformer. *Eng. Appl. Artif. Intell.* **2023**, *124*, 106507. [\[CrossRef\]](#)
29. Pei, X.; Zheng, X.; Wu, J. Rotating Machinery Fault Diagnosis Through a Transformer Convolution Network Subjected to Transfer Learning. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–11. [\[CrossRef\]](#)
30. Lu, N.; Yin, T. Transferable common feature space mining for fault diagnosis with imbalanced data. *Mech. Syst. Signal Process.* **2021**, *156*, 107645. [\[CrossRef\]](#)
31. Li, Y.; Yang, M.; Zhang, Z. A Survey of Multi-View Representation Learning. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 1863–1883. [\[CrossRef\]](#)
32. Chen, X.; Zhang, B.; Gao, D. Bearing fault diagnosis base on multi-scale CNN and LSTM model. *J. Intell. Manuf.* **2021**, *32*, 971–987. [\[CrossRef\]](#)
33. Jiang, G.; Jia, C.; Nie, S.; Wu, X.; He, Q.; Xie, P. Multiview enhanced fault diagnosis for wind turbine gearbox bearings with fusion of vibration and current signals. *Measurement* **2022**, *196*, 111159. [\[CrossRef\]](#)
34. Wang, D.; Li, Y.; Jia, L.; Song, Y.; Wen, T. Attention-Based Bilinear Feature Fusion Method for Bearing Fault Diagnosis. *IEEE/ASME Trans. Mechatron.* **2023**, *28*, 1695–1705. [\[CrossRef\]](#)

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.