

Article

GMRVGG: A Bearing Fault Diagnosis Method Based on Tri-Modal Image Feature Fusion

Ao Li , Yuantao Li, Xiaoli Wang  and Jiancheng Yin 

School of Mechanical Engineering, Shandong University of Technology, Zibo 255000, China; 24501040035@stumail.sdut.edu.cn (A.L.); 24401020028@stumail.sdut.edu.cn (Y.L.); 24501040032@stumail.sdut.edu.cn (X.W.)

* Correspondence: yinjc@sdut.edu.cn

Highlights

What are the main findings?

- A novel bearing fault diagnosis method (GMRVGG) is proposed, which fuses tri-modal features extracted by VGG16 from 1D vibration signals converted into 2D images via GADF, MTF, and RP techniques.
- The proposed methodology demonstrates exceptional robustness, achieving superior overall accuracies on the CWRU and Ottawa datasets, respectively, even under severe 6 dB noise interference.

What are the implications of the main findings?

- By integrating diverse signal-to-image conversion techniques, this approach effectively overcomes the limitations of insufficient information dimensionality and inadequate feature representation inherent in conventional single-modal methods.
- It establishes a highly reliable and highly generalizable solution for intelligent condition monitoring and fault diagnosis of rotating machinery operating in complex, noisy industrial environments.

Abstract

Bearings serve as vital components in rotating machinery. Fault diagnosis of bearings constitutes an essential area within mechanical health monitoring. However, most existing methods rely solely on single-modal data or employ a single signal-to-image conversion technique, leading to insufficient information dimensionality and inadequate feature representation, which ultimately limits diagnostic accuracy. To address these challenges, this paper proposes a bearing fault diagnosis method (GADF-MTF-RP-VGG16, GMRVGG) based on tri-modal image feature fusion. Specifically, three image conversion techniques—Gramian Angular Difference Field (GADF), Markov Transition Field (MTF), and Recurrence Plot (RP)—are utilized to first convert 1D vibration signals into 2D images. Subsequently, shallow to deep features are extracted and fused through the VGG16 backbone network. Finally, fault diagnosis is achieved by integrating a fully connected classifier layer. The proposed methodology was comprehensively validated on both the Case Western Reserve University (CWRU) and the University of Ottawa datasets, which were augmented with severe 6 dB Gaussian white noise and 6 dB pink noise to simulate complex industrial environments. Under these harsh conditions, the proposed method achieved superior overall accuracies (up to 96.9% on the CWRU dataset and consistently 95.8% on the Ottawa dataset), significantly surpassing conventional single-modal approaches. This effectively addresses the limitations of insufficient feature dimensionality and inadequate representation, establishing a highly reliable and robust solution for intelligent bearing fault diagnosis.



Academic Editors: Jong-Myon Kim and Zahoor Ahmad

Received: 12 March 2026

Revised: 9 April 2026

Accepted: 13 April 2026

Published: 15 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

Keywords: bearing fault diagnosis; tri-modal image feature fusion; signal-to-image conversion; Gramian Angular Difference Field; Markov Transition Field; Recurrence Plot; VGG16

1. Introduction

Modern industry relies heavily on diverse mechanical equipment, wherein bearings invariably play a key role. Functionally, bearings serve as one of the core components in rotating machinery [1], finding critical applications across mechanical manufacturing [2], aerospace [3], wind power generation [4], agriculture [5], and beyond. However, rolling bearings are simultaneously among the most failure-prone parts in modern mechanical systems. For instance, statistics indicate that approximately 40% of motor failures are attributed to bearing failures [6]. Thus, bearing fault diagnosis constitutes an essential aspect of mechanical health monitoring.

In the field of bearing fault diagnosis, while traditional signal processing methods are effective for analyzing individual samples, they often rely on expert knowledge for manual interpretation. Consequently, these methods face challenges in scaling to the massive datasets typical of modern industrial monitoring. Deep learning can uncover latent features from abstract data. With the advancement of artificial intelligence technologies, particularly deep learning techniques, diverse intelligent algorithms have proliferated, thereby furnishing a robust theoretical foundation for intelligent fault diagnosis in bearings [7]. Currently, deep learning has become an essential tool for intelligent bearing fault diagnosis [8]. Many scholars are currently focusing on utilizing deep learning technologies, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Residual Networks (ResNet), to process vibration signals for bearing fault diagnosis. Cui et al. [9] proposed a bearing fault diagnosis method integrating an attention mechanism with a bidirectional temporal convolutional network (BiTCN). This approach employs SENet to screen fault-related features, utilizes BiTCN to bidirectionally extract temporal dependencies, dynamically weights key features through multi-head attention, and ultimately performs classification via a fully connected layer. On the Jiangnan University and Case Western Reserve University datasets, the method achieved accuracies of 99.49% and 99%, respectively, with an F1-score of 0.98, significantly outperforming comparative models and demonstrating strong generalization capability. Wang et al. [10] designed a 1D Vision ConvNet by parallelizing multi-layer small convolutional kernels and a single-layer large convolutional kernel, validating its effectiveness on both the Case Western Reserve University dataset and the Paderborn University dataset. Si et al. [11] proposed a Multi-Scale Dilated Convolution and Dense Temporal Convolutional Network (MSDC-DenseTCN) model for bearing fault diagnosis under heavy noise conditions. The model first employs a Multi-Scale Dilated Convolution (MSDC) module with three branches of different dilation rates to filter noise and extract multi-scale features, thereby enhancing the receptive field. Subsequently, it utilizes a Dense Temporal Convolutional Network (DenseTCN) module, which combines the feature reuse mechanism of DenseNet and the temporal modeling capability of TCN, to capture long-range temporal dependencies and optimize feature propagation. The model was validated on the Paderborn University and Case Western Reserve University bearing datasets, demonstrating excellent diagnostic performance under various noise conditions. Duan et al. [12] proposed LSBT-Net, a lightweight and interpretable spatiotemporal model, for insulated bearing fault diagnosis. The model first replaces the standard linear projection layer with a multi-scale convolutional architecture to extract both local fine-grained and global coarse-grained features from vibration signals in parallel. It then integrates high-dimensional fault features through a spatiotemporal information fusion module and

employs adaptive average pooling to compress spatial dimensions, significantly reducing parameter count and computational resource requirements. Evaluated on wind turbine and rail transit motor insulated bearing datasets, the framework achieved an accuracy of 99.6%. However, despite the demonstrated effectiveness of these deep learning models in processing 1D vibration signals, their performance remains constrained by the inherent limitations of the original time-series data. Specifically, 1D vibration signals often contain substantial non-stationary components and background noise, thereby making it difficult for the models to comprehensively capture the subtle differences and dynamic evolution patterns of fault characteristics. This incompleteness in feature extraction frequently leads to a degradation in model performance.

Image data exhibits superior representational capacity. By employing signal-to-image transformation techniques to convert 1D signals into 2D images, the spatial structures and textural features of the resulting images can be leveraged to clearly reveal the information embedded within the signals. Wang et al. [13] proposed two methods for converting time-series signals into images: the Gramian Angular Field (GAF) based on polar coordinate mapping, and the Markov Transition Field (MTF) based on bin discretization and state transition probability matrices. By employing a tiled convolutional neural network integrated with a linear support vector machine for classifying images generated by these two methods, competitive classification accuracy was achieved on 12 UCR datasets. Daubechies et al. [14] proposed the Synchrosqueezed Wavelet Transform (SST), which generates an initial time-frequency representation through wavelet transform and then compresses it into a high-resolution time-frequency image via reassignment techniques. Its high-precision capability in extracting instantaneous frequencies under strong noise conditions was validated using medical ECG data. Owing to the advantages of image data, numerous scholars in the field of bearing fault diagnosis opt to first convert time-series vibration signals into images, and subsequently extract features contained within these images to accomplish fault diagnosis. Li et al. [15] employed a modified successive variational mode decomposition (SVMD) to extract key signal components. The reconstructed signal was then converted into images via the Gramian Angular Difference Field (GADF) technique and fed into a convolutional neural network for fault classification, achieving an accuracy of 95.2% on the Case Western Reserve University dataset. Wang et al. [16] converted time-series vibration signals into images using the Markov Transition Field (MTF) and input them into a convolutional neural network for bearing fault diagnosis, achieving an accuracy of nearly 100% on the Case Western Reserve University dataset.

However, the aforementioned methods generally suffer from the following limitations: Relying solely on single-modal data for bearing fault diagnosis results in insufficient information dimensionality. Similarly, employing a single signal-to-image conversion technique to process temporal signals leads to inadequate feature representation, as any individual conversion method may fail to comprehensively capture the full characteristics of faults.

To address these limitations, a bearing fault diagnosis method named GADF-MTF-RP-VGG16 (GMRVGG), based on tri-modal image feature fusion, is proposed: The proposed methodology first converts 1D vibration signals into images via three distinct signal-to-image encoding techniques—Gramian Angular Difference Field (GADF), Markov Transition Field (MTF), and Recurrence Plot (RP)—to extract comprehensive signal information. Subsequently, VGG16 is employed to extract features spanning shallow to deep layers from these images. Finally, the feature vectors from different modalities are fused through concatenation and fed into a fully connected classifier for fault diagnosis.

The remainder of this paper is structured as follows: Section 2 provides a detailed introduction to the background methods employed in this study; Section 3 elaborates on

the training pipeline and architecture of the GMRVGG framework; Section 4 describes the dataset, experimental methodology, and results; and Section 5 concludes with a summary of the work presented in this paper.

2. Background Methods

2.1. Gramian Angular Difference Field

The Gramian Angular Difference Field (GADF) is a technique that converts 1D vibration signals into 2D images by characterizing the dynamic evolution of time series through angular differences between time points [17]. This method has been widely applied in fields such as rotating machinery fault diagnosis [18], transmission line fault diagnosis [19], and spectroscopic analysis [20]. During the conversion of 1D vibration signals into images, the GADF preserves the critical information within the time series [21]. Fluctuations in 1D vibration signals are mapped into cross-shaped patterns within their corresponding GADF images. The greater the amplitude of these fluctuations, the more pronounced the cross-shaped patterns become in the imagery. Therefore, GADF images can effectively capture fault impulses within bearing vibration signals and characterize the fault features present during the vibration process [22]. For a given time series $X = \{x_1, x_2, \dots, x_n\}$, where n denotes the length of the time series sequence and x_i represents the vibration signal amplitude at the i -th time step, the GADF transformation is defined as follows:

- (1) Normalization: For a given time series $X = \{x_1, x_2, \dots, x_n\}$, it is normalized to the range $[-1, 1]$. Let $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n\}$ denote the normalized time series.

$$\tilde{x}_i = \frac{(x_i - \max(X)) + (x_i - \min(X))}{\max(X) - \min(X)}, x_i \in X, \quad (1)$$

where x_i is the raw signal and $\tilde{x}_i$ denotes the normalized signal.

- (2) Convert to polar coordinates.

$$\begin{cases} \phi = \arccos(\tilde{x}_i), -1 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \tilde{X} \\ r = \frac{t_i}{N}, t_i \in N \end{cases}, \quad (2)$$

where t_i denotes the time index, and N represents a constant scaling factor controlling the range of polar coordinates.

- (3) Constructing the Gramian Angular Difference Field.

$$GADF = \begin{bmatrix} \sin(\phi_1 - \phi_1) & \cdots & \sin(\phi_1 - \phi_n) \\ \sin(\phi_2 - \phi_1) & \cdots & \sin(\phi_2 - \phi_n) \\ \vdots & \ddots & \vdots \\ \sin(\phi_n - \phi_1) & \cdots & \sin(\phi_n - \phi_n) \end{bmatrix}, \quad (3)$$

Taking a segment of signal under outer race fault condition as an example, it is converted into a 2D image using the GADF and “jet” colormap. The generated GADF image is shown in Figure 1.

As shown in Figure 1, the 1D vibration signal is mapped into a two-dimensional matrix based on the trigonometric difference between angular points, forming a cross-like texture pattern.

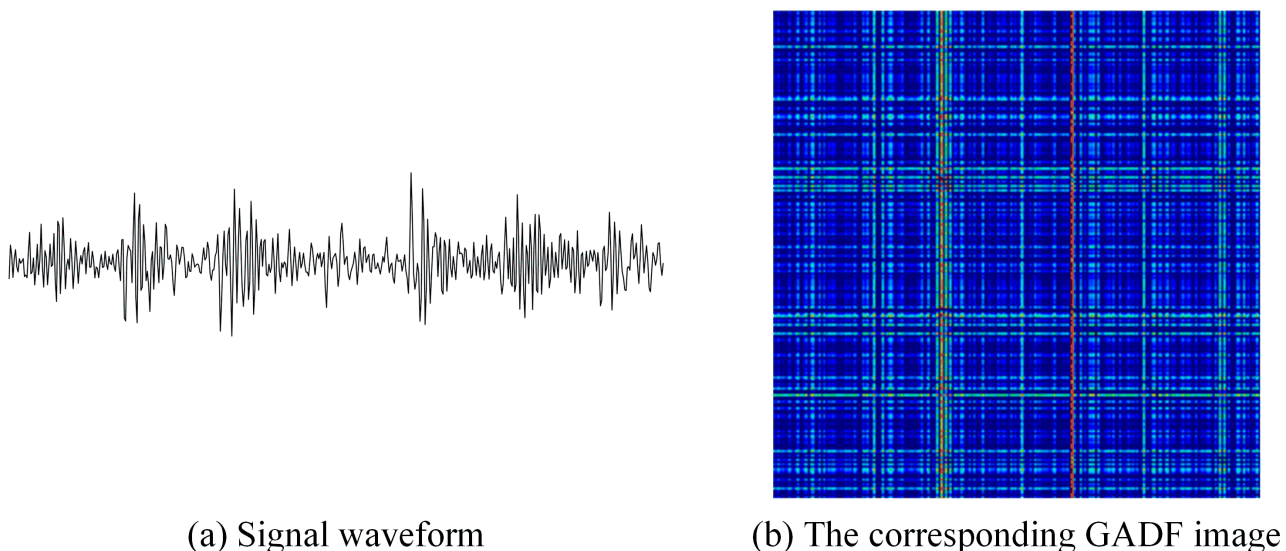


Figure 1. Illustration of the GADF transformation process. **(a)** Raw vibration signal waveform under an outer race fault condition. **(b)** The corresponding converted GADF image. The colors in the image correspond to the scalar values mapped by the “jet” colormap, indicating the trigonometric difference between angular points.

2.2. Markov Transition Field

The Markov Transition Field (MTF) is a technique that converts 1D vibration signals into 2D images by leveraging the dynamic evolution process of temporal signals [23]. It has been widely applied in fields such as fault diagnosis in rotating machinery [24] and condition monitoring of industrial equipment [25]. The MTF constructs a Markov transition matrix through the state transition probabilities of adjacent temporal nodes, which is then expanded into a transition field and ultimately mapped to a grayscale image via downsampling. When processing raw time-series vibration signals directly with CNNs, potential information loss may occur. In contrast, the MTF effectively preserves the temporal characteristics of the signals, thereby enhancing the capability of CNNs in handling time-series data [26]. For a given time series $X = \{x_1, x_2, \dots, x_n\}$, where n denotes the length of the time series sequence and x_i represents the vibration signal amplitude at the i -th time step, the MTF transformation is defined as follows:

- (1) Discretization: For a given time series $X = \{x_1, x_2, \dots, x_n\}$, partition the time series values into Q quantile bins, with each x_i mapped to its corresponding quantile $q_j (j \in [1, Q])$.
- (2) Constructing the Markov transition matrix:

The transition probability w_{ij} from quantile bin q_i to q_j is defined as:

$$w_{ij} = p\{x_{t+1} \in q_j | x_t \in q_i\}, \quad (4)$$

where x_t represents the signal amplitude at time t , and q_i indicates the quantile region corresponding to x_i .

Specifically, w_{ij} is computed by counting the transition frequency N_{ij} from bin q_i to bin q_j and performing row normalization:

$$w_{ij} = \frac{N_{ij}}{\sum_{k=1}^Q N_{ik}}, \quad (5)$$

where N_{ij} denotes the number of times a point in bin q_i is immediately followed by a point in bin q_j .

This results in the Markov Transition Matrix W .

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1Q} \\ w_{21} & w_{22} & \cdots & w_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ w_{Q1} & w_{Q2} & \cdots & w_{QQ} \end{bmatrix}, \quad (6)$$

(3) Expanded into the Markov Transition Field.

$$MTF = \begin{bmatrix} w_{ij}|x_1 \in q_i, x_1 \in q_j & \cdots & w_{ij}|x_1 \in q_i, x_n \in q_j \\ w_{ij}|x_2 \in q_i, x_1 \in q_j & \cdots & w_{ij}|x_2 \in q_i, x_n \in q_j \\ \vdots & \ddots & \vdots \\ w_{ij}|x_n \in q_i, x_1 \in q_j & \cdots & w_{ij}|x_n \in q_i, x_n \in q_j \end{bmatrix}, \quad (7)$$

Taking a segment of signal under outer race fault condition as an example, it is converted into a 2D image using the MTF. The generated MTF image is shown in Figure 2.

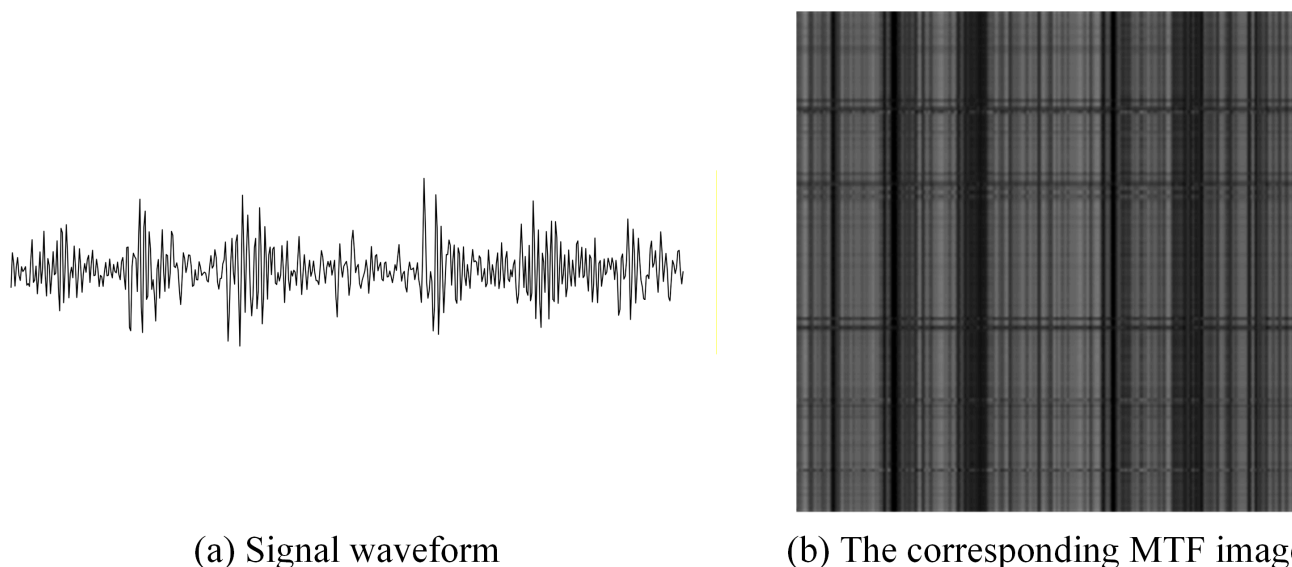


Figure 2. Illustration of the MTF transformation process. (a) Raw vibration signal waveform under an outer race fault condition. (b) The corresponding converted MTF image. The grayscale intensities in the image represent the transition probabilities between different quantile bins.

As shown in Figure 2, the pixel intensities in the image represent the Markov transition probabilities between different quantile bins of the time-series data.

2.3. Recurrence Plot

The Recurrence Plot (RP) employs phase space reconstruction technology to generate 2D images by quantifying the recurrence characteristics between state points in dynamic systems [27]. This technique enables the visual conversion of 1D signals into high-dimensional dynamic features, and it has been widely applied in fields such as fault diagnosis for rotating machinery [28] and fluid dynamics research [29]. The RP algorithm is capable of effectively detecting characteristic patterns hidden within short-term non-smooth time series, making it well-suited for bearing fault diagnosis scenarios [30]. For a given time series $X = \{x_1, x_2, \dots, x_n\}$, where n denotes the length of the time series

sequence and x_i represents the vibration signal amplitude at the i -th time step, the RP transformation is defined as follows:

- (1) Phase Space Reconstruction: For a given time series $X = \{x_1, x_2, \dots, x_n\}$, construct the phase space vectors.

$$\vec{v}_i = (x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(m-1)\tau}), i = 1, 2, 3 \dots N, N = n - (m-1)\tau, \quad (8)$$

where m denotes the embedding dimension and τ represents the time delay.

- (2) Computing the Distance Matrix (Using Euclidean Distance as an Example).

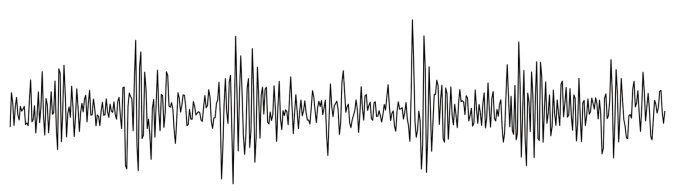
$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{bmatrix}, d_{ij} = \|\vec{v}_i - \vec{v}_j\| = \sqrt{\sum_{k=0}^{m-1} (x_{i+k\tau} - x_{j+k\tau})^2}, \quad (9)$$

- (3) Transformed into a continuous recurrence matrix via the Gaussian kernel function.

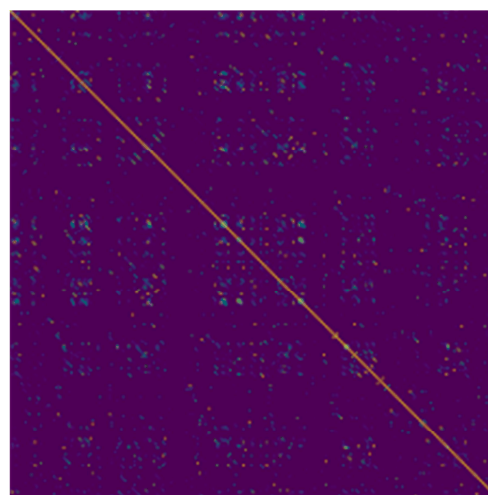
$$RP = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1N} \\ r_{21} & r_{22} & \dots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \dots & r_{NN} \end{bmatrix}, r_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \quad (10)$$

where σ is the Gaussian kernel parameter that controls the sensitivity of recurrence.

Taking a segment of signal under outer race fault condition as an example, it is converted into a 2D image using the RP and “viridis” colormap. The generated RP image is shown in Figure 3.



(a) Signal waveform



(b) The corresponding RP image

Figure 3. Illustration of the RP transformation process. (a) Raw vibration signal waveform under an outer race fault condition. (b) The corresponding converted RP image. The colors in the image correspond to the distance values calculated by the Gaussian kernel, visualized using the “viridis” colormap.

As shown in Figure 3, the image visualizes the recurrence states of the phase space trajectories, where the colors correspond to the distance values calculated by the Gaussian kernel.

convolutional filters are employed throughout the network to capture features. Finally, the feature maps are flattened and passed through the linear layers, followed by a Softmax layer to generate the final classification probabilities.

3. Proposed Method

3.1. Methodology Flow

To address the limitations of insufficient information dimensionality and inadequate feature representation in single-modal models, a novel GMRVGG-based bearing fault diagnosis method is proposed, which integrates three signal-to-image conversion techniques (GADF, MTF, RP) with VGG16 for tri-modal feature fusion. The training of the proposed method consists of two distinct stages: initially training the feature extractor, and subsequently training the fusion classifier after fixing the optimal parameters of the feature extractor. This modular design offers significant flexibility, allowing for straightforward addition, removal, or substitution of various modalities. Crucially, changes to the feature extractor of any single modality do not affect the extractors of other modalities. The training pipeline is illustrated in Figure 6.

As shown in Figure 6, the training pipeline of the GMRVGG Framework is as follows:

Step 1: Train the single-modality feature extractor.

Step 1.1: Preprocess the signals and partition the dataset into training and test sets proportionally.

Step 1.2: Convert the samples in both the training set and test set into their corresponding images using GADF, MTF, and RP, respectively.

GADF transformation: The 1D signals are converted into GADF images based on Equation (1) to (3) in Section 2.1. In this study, to match the input dimension of the VGG16 network, the value of N is determined as 224. To accommodate the input requirements of the VGG16 architecture, the scalar values of the GADF matrices were transformed into the RGB color space using the “jet” colormap.

MTF transformation: The 1D signals are converted into MTF images based on Equation (4) to (7) in Section 2.2.

RP transformation: The 1D signals are converted into RP images based on Equation (8) to (10) in Section 2.3. In this study, the parameters were set as embedding dimension $m = 2$ and time delay $\tau = 1$. Consequently, the reconstructed vector simplifies to $\vec{v}_i = (x_i, x_{i+1})$. This configuration preserves the complete dynamic evolution patterns information of the vibrational signals, avoiding potential information loss caused by large time delays, thereby providing rich input features for the subsequent VGG16 network. To accommodate the input requirements of the VGG16 architecture, the Gaussian-kernelized recurrence matrices were visualized using the “viridis” colormap.

Step 1.3: Input the training set images from the GADF branch into the VGG16 network for training. The network parameters are iteratively updated via the backpropagation algorithm using the Adam optimizer to minimize the Cross-Entropy Loss. The weights corresponding to the highest classification accuracy on the validation set are preserved as the optimal weights for the branch.

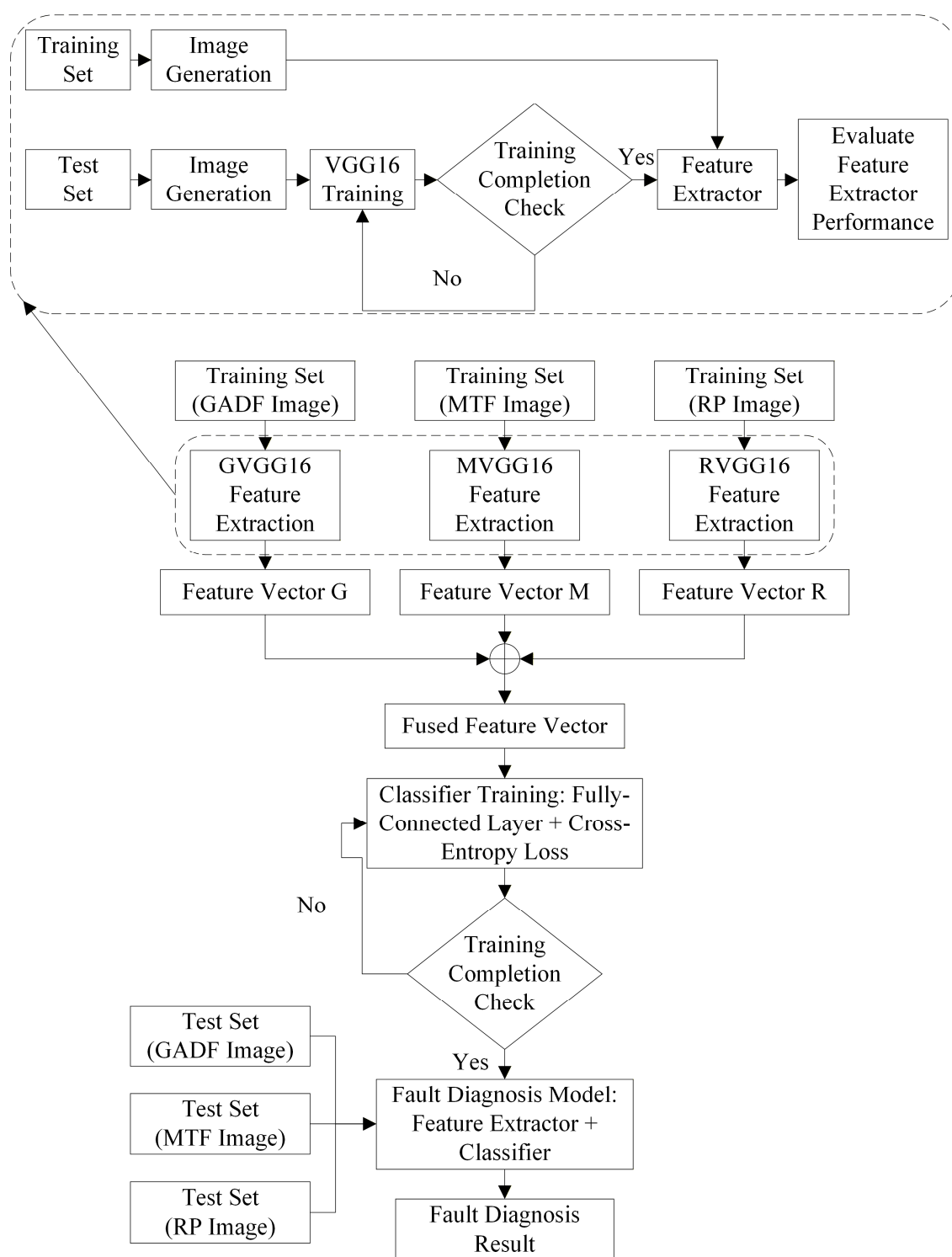


Figure 6. Training pipeline of the GMRVGG framework. The pipeline illustrates the two-stage training process, including the feature extractor training for each modality and the final classifier training.

The Cross-Entropy loss function is a highly suitable choice for multi-class classification tasks. It effectively quantifies the disparity between the predicted probability distribution and the true distribution of the labels. Compared to other loss functions, such as the Mean Squared Error, the Cross-Entropy loss generates stronger gradients for incorrect

predictions, thereby accelerating the convergence of gradient descent algorithms [33]. The Cross-Entropy Loss function is defined as follows:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_c^{(i)} \log(\hat{y}_c^{(i)}), \quad (11)$$

where N represents the number of samples in a batch; C denotes the total number of classes; $y_c^{(i)}$ is a binary indicator (0 or 1) specifying whether class c is the true classification for sample i ; and $\hat{y}_c^{(i)}$ is the predicted probability that sample i belongs to class c .

Select the Rectified Linear Unit (ReLU) activation function. The ReLU activation function is defined as follows:

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases} \quad (12)$$

where x is the input to the nonlinear activation function f across all input feature channels.

In this study, the variable m is used to represent the minibatch size. Considering the high memory consumption of the VGG16 architecture, the minibatch size was set to $m = 4$ for this feature extractor training phase.

Step 1.4: Load the optimal weights and input the test set images into the trained VGG16 for testing.

Step 1.5: Upon completion of training and testing, fix the weights of the feature extractor, referred to as GVGG16. Specifically, the output of the first fully connected layer is extracted as the feature vector for the subsequent fusion process.

Step 1.6: The same procedure is applied to train and test the feature extractors for the MTF and RP branches, resulting in the MVGG16 and RVGG16 feature extractors with fixed weights after training.

Step 2: Train the final fully connected classifier.

Step 2.1: Extract feature vectors from the training set images of GADF, MTF, and RP modalities using three branch feature extractors respectively. For each data sample, the proposed model extracts a 4096-dimensional feature vector.

Step 2.2: Concatenate the three distinct feature vectors to form a unified 12,288-dimensional vector and feed them into a fully connected classifier. Train the classifier using cross-entropy loss to optimize its parameters, and preserve the optimal weights. Although these features originate from the same signal and may share semantic overlaps, their fusion provides complementary visual perspectives, enabling the classifier to learn a more comprehensive representation. Similarly, the Adam optimizer is employed to update the weights iteratively. The model parameters that achieve the highest classification accuracy on the validation set are saved as the optimal weights for the final fault diagnosis model. For this lightweight fusion classifier, the minibatch size was increased to $m = 32$ to facilitate stable convergence.

Use Batch Normalization (BatchNorm) to accelerate the training process, enhance training stability, and confer a regularization effect. For a mini-batch of data $X = \{x_1, x_2, \dots, x_n\}$, the BatchNorm formulation is defined as follows:

(1) Compute the mean and variance of the current mini-batch data.

$$\mu_X = \frac{1}{m} \sum_{i=1}^m x_i, \quad (13)$$

$$\sigma_X^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_X)^2, \quad (14)$$

where X denotes the current mini-batch data; m represents the size of the mini-batch; x_i indicates the activation value of the i -th sample in the mini-batch; μ_X signifies the mean of the current mini-batch; and σ_X^2 denotes the variance of the current mini-batch.

(2) Standardization.

$$\hat{x}_i = \frac{x_i - \mu_X}{\sqrt{\sigma_X^2 + \epsilon}}, \quad (15)$$

where ϵ denotes a small constant to avoid division by zero.

(3) Scaling and Shifting.

$$y_i = \gamma \hat{x}_i + \beta, \quad (16)$$

where γ denotes the scale factor; β represents the shift factor. Both γ and β are learnable parameters.

Dropout is a regularization technique designed to prevent overfitting [34]. The formulation of Dropout is defined as follows:

$$\tilde{x} = \frac{1}{1-p} m \odot x, \quad (17)$$

where m denotes the binary mask, with 0 indicating dropout and 1 indicating retention; p represents the dropout probability.

Step 2.3: Load the optimal weights, process the test set images through the three branch feature extractors and the classifier for final testing and evaluation, and complete the multimodal bearing fault diagnosis task.

The architecture of the proposed model is illustrated in Figure 7.

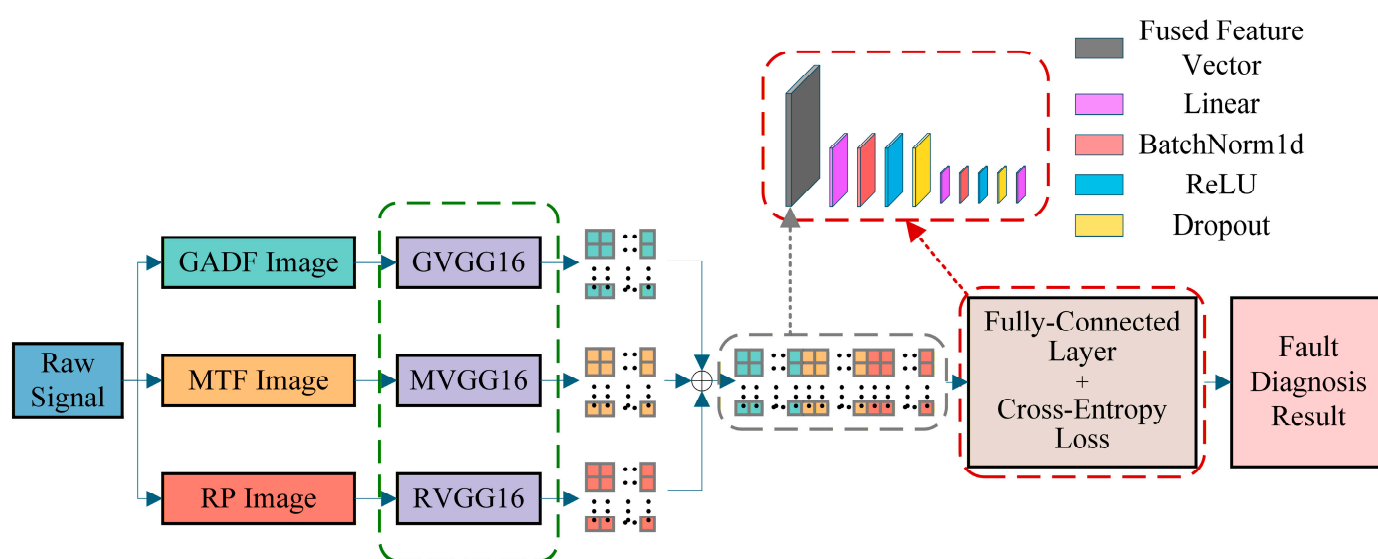


Figure 7. Architecture of the GMRVGG. The model integrates three branches (GADF, MTF, RP) with VGG16 backbones and fuses features via concatenation before the final classification. The arrays of colored squares interspersed with dots denote the extracted high-dimensional feature vectors corresponding to each modality (cyan for GADF, orange for MTF, and red for RP). The expanded dashed box illustrates the detailed internal layer sequence of the fully connected classifier.

As shown in Figure 7, the overall architecture of the GMRVGG model consists of three parallel VGG16 feature extraction branches, which extract high-dimensional features from GADF, MTF, and RP images, respectively. These multi-modal feature vectors are fused via concatenation and then fed into a fully connected classifier to output the final bearing fault diagnosis results.

3.2. Network Parameters

The network parameters of the VGG16 feature extractor are listed in Table 1.

Table 1. Network Parameters of the VGG16 Feature Extractor.

Layer	Output Size	Kernel Size	Stride	Repeat	Output Channels
Input	$224 \times 224 \times 3$	-	-	-	3
Conv1 + ReLU	224×224	3×3	1	2	64
MaxPool1	112×112	2×2	2	1	-
Conv2 + ReLU	112×112	3×3	1	2	128
MaxPool2	56×56	2×2	2	1	-
Conv3 + ReLU	56×56	3×3	1	3	256
MaxPool3	28×28	2×2	2	1	-
Conv4 + ReLU	28×28	3×3	1	3	512
MaxPool4	14×14	2×2	2	1	-
Conv5 + ReLU	14×14	3×3	1	3	512
MaxPool5	7×7	2×2	2	1	-
AvgPool	7×7	-	-	1	512
Flatten	25,088	-	-	1	-
Linear	4096	-	-	1	4096

As shown in Table 1, the model accepts an input image of size $224 \times 224 \times 3$. Through a series of convolutional and pooling layers, the spatial resolution is progressively reduced while the depth is increased. Finally, the feature map is flattened and processed by the linear layer to output a 4096-dimensional feature vector, which serves as the input for the subsequent fusion module.

The network parameters of the fully connected classifier are listed in Table 2.

Table 2. Network Parameters of the Fully Connected Classifier.

Layer	Output Size	Output Channels
Input	12,288	12,288
Linear1 + BN1 + ReLU1 + Dropout1	1024	1024
Linear2 + BN2 + ReLU2 + Dropout2	512	512
Output (Linear3)	4	4

As shown in Table 2, the classifier receives a fused feature vector with a dimensionality of 12,288, formed by concatenating three 4096-dimensional vectors. Its architecture comprises three linear layers. To ensure training stability and prevent overfitting, BatchNorm and Dropout layers are incorporated between the linear layers. The final layer maps the features to four output nodes, corresponding to the number of fault categories.

3.3. Implementation Details

To ensure the reproducibility of the proposed method, the hyperparameters used for training are listed in Table 3.

As shown in Table 3, the model was implemented using the PyTorch framework (version 2.6.0). The Adam optimizer was employed with an initial learning rate of 1×10^{-4} for all phases. To ensure stable convergence, a StepLR scheduler was used to decay the learning rate.

Table 3. Hyperparameters for Model.

Parameter	Feature Extractor	Fully Connected Classifier
Batch Size	2	108/24
Optimizer	Adam	Adam
Initial Learning Rate	1×10^{-4}	1×10^{-4}
LR Scheduler	StepLR (step = 40, gamma = 0.97)	StepLR (step = 40, gamma = 0.95)
Weight Decay	-	1×10^{-4}
Epochs	50	50
Loss Function	Cross-Entropy Loss	Cross-Entropy Loss

Regarding data preprocessing, all generated GADF, MTF, and RP images were resized to 224×224 pixels. Subsequently, standard ImageNet normalization (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) was applied to facilitate the convergence of the pre-trained VGG16 backbone.

4. Results and Discussion

4.1. Experiment I: Diagnostic Performance Analysis on CWRU Dataset

4.1.1. Dataset Introduction

In this section, the bearing fault diagnosis dataset from Case Western Reserve University (CWRU) is employed for experimental validation. The CWRU bearing dataset was developed by the Mechanical Engineering Laboratory at Case Western Reserve University in the United States. The experimental setup utilized a 2 hp Reliance Electric motor (Reliance Electric, Greenville, SC, USA), with acceleration data measured both near and away from the motor bearings [35]. The experiments utilized SKF 6205-2RS JEM deep groove ball bearings (SKF, Gothenburg, Sweden), which are standard industrial metallic bearings.

Bearing faults are induced by Electric Discharge Machining (EDM). The fault signals are categorized into two types: drive-end bearing signals and fan-end bearing signals. The motor operates under four load conditions: 0 hp, 1 hp, 2 hp, and 3 hp. The fault types include ball faults, inner race faults, and outer race faults. The fault diameters are artificially seeded at 0.007 inch, 0.014 inch, 0.021 inch, and 0.028 inch. The bearing fault signals are sampled at two frequencies: 12,000 Hz and 48,000 Hz.

For the experiments in this section, drive-end bearing signals under a 0 hp load condition with a sampling frequency of 12,000 Hz were selected. Fault diameters of 0.007 inch, 0.014 inch, and 0.021 inch were chosen, and combined with signals from the normal state, this resulted in a total of ten distinct categories of bearing fault signal data. Signals of the same fault type but with different fault diameters are grouped under the same label, resulting in four final categories: “Normal”, “Ball”, “IR”, and “OR”.

In practical industrial settings, vibration signals from bearings are inevitably contaminated by background noise. To rigorously evaluate the robustness of the proposed method against noise interference, artificial noise is introduced into the original signals prior to sample segmentation. Acknowledging that real-world industrial noise is often complex and machine-condition-dependent, this study employs two distinct types of noise for comprehensive validation: Gaussian white noise and pink noise, both at 6 dB. Gaussian white noise is employed to simulate ideal random measurement noise, whereas pink noise is utilized to characterize the low-frequency background noise prevalent in practical industrial environments [36]. A comparison of the signal waveforms before and after adding 6 dB Gaussian white noise is shown in Figure 8.

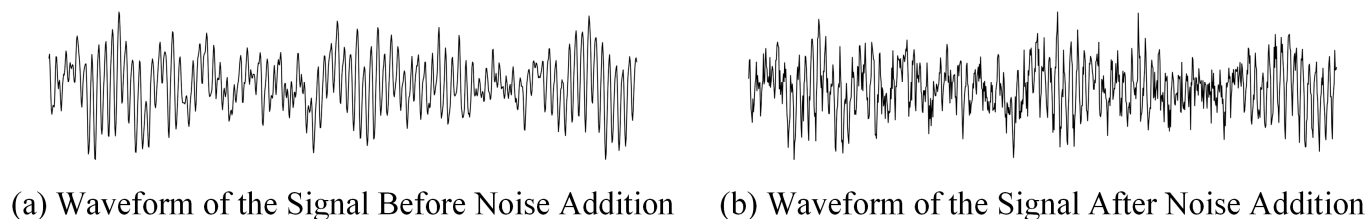


Figure 8. Signal waveform comparison under 6 dB Gaussian white noise. (a) Original signal before noise addition. (b) Signal after adding 6 dB Gaussian white noise.

As shown in Figure 8, the waveform of the signal undergoes a significant transformation, characterized by a marked increase in random irregularities and intense fluctuations in the amplitude of its peaks and troughs. This comparison visually illustrates the significant impact of the added noise. A comparison of the signal waveforms before and after adding 6 dB pink noise is shown in Figure 9.

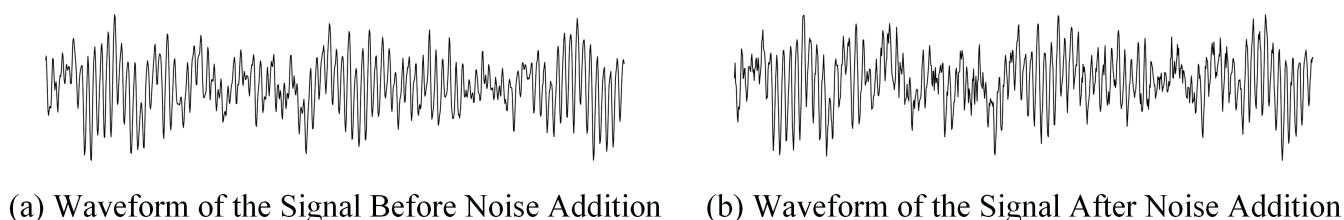


Figure 9. Signal waveform comparison under 6 dB pink noise. (a) Original signal before noise addition. (b) Signal after adding 6 dB pink noise.

As shown in Figure 9, the waveform also exhibits a significant transformation. In contrast to the sharp irregularities of white noise, this change is characterized by more pronounced low-frequency modulation and a blurring of the distinct peaks and troughs.

To prevent data leakage, for each noise scenario, the data points with added noise were partitioned into training and test sets in an 8:2 ratio. Subsequently, samples were obtained through sequential segmentation with a sample length set to 512 data points and an overlap ratio of 50%. The details of the processed CWRU dataset partition are listed in Table 4.

Table 4. Processed CWRU Dataset.

Fault	Label	Dataset (Train/Test)
Ball	1	108/24
IR	2	108/24
OR	3	108/24
Normal	0	108/24

As shown in Table 4, the dataset is categorized into four classes (Normal, Ball, IR, OR) labeled 0 to 3. To prevent the model from becoming biased toward majority classes and to ensure unbiased evaluation across all fault types, a strictly balanced dataset was constructed. For each noise scenario, each category contains 108 training samples and 24 test samples, ensuring a balanced distribution for model training and evaluation.

4.1.2. Experimental Results

(1) Performance under 6 dB Gaussian white noise

For the method proposed in this paper, the experimental results on the CWRU dataset under the 6 dB Gaussian white noise condition are listed in Table 5.

Table 5. Method comparison under 6 dB Gaussian white noise.

Method	Ball	IR	OR	Normal	Overall Accuracy
GADF + VGG16	58.3%	100.0%	70.8%	100.0%	82.3%
MTF + VGG16	79.2%	79.2%	62.5%	100.0%	80.2%
RP + VGG16	87.5%	87.5%	62.5%	100.0%	84.4%
Proposed method	87.5%	95.8%	83.3%	100.0%	91.7%

As can be seen from Table 5, in this noise environment, the proposed method achieves significantly higher accuracy than other single-modal methods. Under the interference of 6 dB Gaussian white noise, single-modal methods can only achieve high accuracy in identifying certain specific fault types, whereas the tri-modal fusion approach is capable of recognizing each fault type with consistently high accuracy. The experimental results demonstrate that the proposed multimodal fault diagnosis model exhibits a significant advantage in accuracy compared to other single-modal diagnostic models. It is important to note that, as the dataset is strictly balanced with an equal number of samples for each class, the variations in classification accuracy are not caused by class imbalance. Instead, they are solely determined by the distinctiveness of the feature representations learned by the model for different fault types.

To more intuitively evaluate the performance of the fault diagnosis model, this paper also introduces the confusion matrix. The confusion matrix clearly reveals the relationship between the predicted values and the true values. By comparing the color intensity within the matrix, one can intuitively assess the performance of fault classification. Figure 10 shows the confusion matrices for the four methods under 6 dB Gaussian white noise. In these matrices, the horizontal axis represents the predicted labels, while the vertical axis represents the true labels, where 0 represents the Normal condition, 1 represents the Ball Fault, 2 represents the Inner Race Fault, and 3 represents the Outer Race Fault.

As illustrated in Figure 10a, the proposed GMRVGG16 model achieves an overall accuracy of 91.7% and demonstrates high diagnostic accuracy across all fault categories. Specifically, the classification accuracy for the Normal state reaches 100% with no misclassifications. For Ball Faults, the accuracy is 87.5%, with the primary error being misclassification as Outer Race Faults (12.5%). Inner Race Faults achieve a high accuracy of 95.8%, with minor misclassification as Outer Race Faults (4.2%). The accuracy for Outer Race Faults is 83.3%, where the main confusion occurs with Ball Faults (16.7%). Overall, the GMRVGG16 model maintains high diagnostic accuracy and stability across all fault types.

As illustrated in Figure 10b, the GADF + VGG16 model achieves an overall accuracy of 82.3%, yet exhibits significant fluctuations across different fault categories. Specifically, it attains 100% accuracy in classifying the Normal state, correctly identifying all samples. However, for Ball Faults, the accuracy drops sharply to 58.3%, with primary misclassifications occurring as Inner Race Faults (16.7%) and Outer Race Faults (25.0%). In contrast, Inner Race Faults are perfectly classified with 100.0% accuracy and no errors. For Outer Race Faults, the accuracy is 70.8%, dominated by misclassifications as Ball Faults (20.8%) and Inner Race Faults (8.3%). Overall, this method demonstrates a pronounced deficiency in diagnosing Ball Faults, indicating limited robustness for this specific fault type.

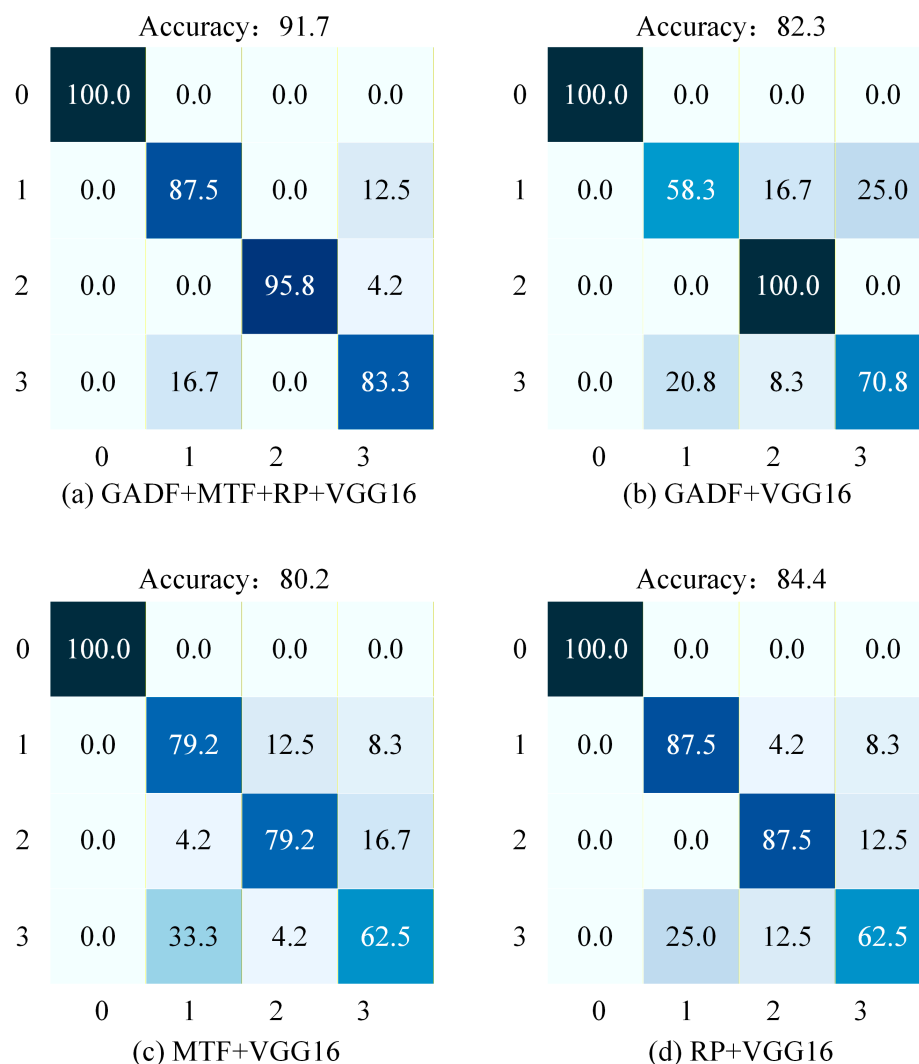


Figure 10. Confusion matrices of the four methods under 6 dB Gaussian white noise. (a) Proposed GMRVGG (Tri-modal). (b) GADF + VGG16. (c) MTF + VGG16. (d) RP + VGG16. The color intensity represents the percentage of classification accuracy, with darker shades indicating higher values.

As illustrated in Figure 10c, the MTF + VGG16 model achieves an overall accuracy of 80.2%. The classification accuracy for the Normal state reaches 100% with no misclassifications. The accuracy for Outer Race Faults is the lowest at 62.5%, with primary errors including misclassification as Ball Faults (33.3%) and Inner Race Faults (4.2%). For Ball Faults, the accuracy is 79.2%, with main errors being misclassified as Inner Race Faults (12.5%) and Outer Race Faults (8.3%). The accuracy for Inner Race Faults is also 79.2%, dominated by misclassifications as Ball Faults (4.2%) and Outer Race Faults (16.7%). Overall, this method proves to be particularly weak in diagnosing Outer Race Faults.

As illustrated in Figure 10d, the RP + VGG16 model achieves an overall accuracy of 84.4%, yet its diagnostic results exhibit certain instability. Specifically, it attains 100.0% accuracy in classifying the Normal state. For Ball Faults, the accuracy reaches 87.5%, with primary misclassifications occurring as Inner Race Faults (4.2%) and Outer Race Faults (8.3%). The accuracy for Inner Race Faults is also 87.5%, dominated by misclassification as Outer Race Faults (12.5%). However, the accuracy for Outer Race Faults is significantly lower at 62.5%, with errors including misclassification as Ball Faults (25.0%) and Inner Race Faults (12.5%). Overall, this method demonstrates substantial room for improvement in diagnosing Outer Race Faults.

In summary, under the 6 dB Gaussian white noise condition, the proposed method achieves significantly higher classification accuracy compared to the other three single-modal approaches, demonstrating its effectiveness on the CWRU dataset. It is noteworthy that for Inner Race Faults, the proposed method exhibits marginally lower accuracy compared to the GADF + VGG16 approach, which may be attributed to feature redundancy and feature conflict introduced during the multimodal fusion process.

(2) Performance under 6 dB pink noise

The experiments were repeated using the dataset added with 6 dB pink noise. The comparison results between the proposed method and other single-modal methods are listed in Table 6.

Table 6. Method comparison under 6 dB pink noise.

Method	Ball	IR	OR	Normal	Overall Accuracy
GADF + VGG16	79.2%	95.8%	87.5%	100.0%	90.6%
MTF + VGG16	91.7%	95.8%	83.3%	100.0%	92.7%
RP + VGG16	87.5%	95.8%	79.2%	100.0%	90.6%
Proposed method	95.8%	95.8%	95.8%	100.0%	96.9%

As can be seen from Table 6, in this noise environment, the proposed method continues to demonstrate outstanding performance, achieving the highest overall accuracy of 96.9%. In contrast, single-modal methods still exhibit instability in such noisy environments. For instance, GADF + VGG16 shows a significant deficiency in diagnosing Ball Faults (79.2%), while RP + VGG16 struggles to accurately identify Outer Race Faults (79.2%). In comparison, the proposed method effectively overcomes these limitations of individual modalities, maintaining an accuracy rate above 95% across all fault categories. These results further validate that the proposed model retains reliable diagnostic capability even when the spectral characteristics of the noise change. Consistent with the previous experiment, these variations in performance are solely attributed to the feature extraction capabilities of the different methods, as the dataset remains strictly balanced.

To further visualize the detailed classification performance under the impact of pink noise, the confusion matrix of the proposed method is presented in Figure 11. Consistent with the previous setup, the horizontal axis represents the predicted labels, and the vertical axis represents the true labels. The class mapping remains identical to the white noise experiment, where 0, 1, 2, and 3 correspond to Normal condition, Ball Fault, Inner Race Fault, and Outer Race Fault, respectively.

As illustrated in Figure 11a, under the pink noise condition, the proposed GMRVGG16 model achieves an outstanding overall accuracy of 96.9%, demonstrating exceptional robustness. Specifically, the classification accuracy for the Normal state reaches 100% with no misclassifications. For Ball, Inner Race, and Outer Race faults, the model achieves a consistent accuracy of 95.8% for each category. The errors are minimal and evenly distributed: Ball Faults are slightly misclassified as Outer Race Faults (4.2%), Inner Race Faults as Ball Faults (4.2%), and Outer Race Faults as Inner Race Faults (4.2%). Overall, the GMRVGG16 model maintains the highest diagnostic accuracy and stability among all methods.

As illustrated in Figure 11b, the GADF + VGG16 model achieves an overall accuracy of 90.6%. While identifying the Normal state (100.0%) and Inner Race Faults (95.8%) with high precision, it still exhibits a distinct bottleneck in diagnosing Ball Faults. The accuracy for Ball Faults is limited to 79.2%, with primary misclassifications occurring as Inner Race Faults (12.5%) and Outer Race Faults (8.3%). Additionally, Outer Race Faults achieve an

accuracy of 87.5%, with minor confusions. This indicates that while the overall performance improves under pink noise compared to white noise, the GADF modality remains less sensitive to Ball Fault features.

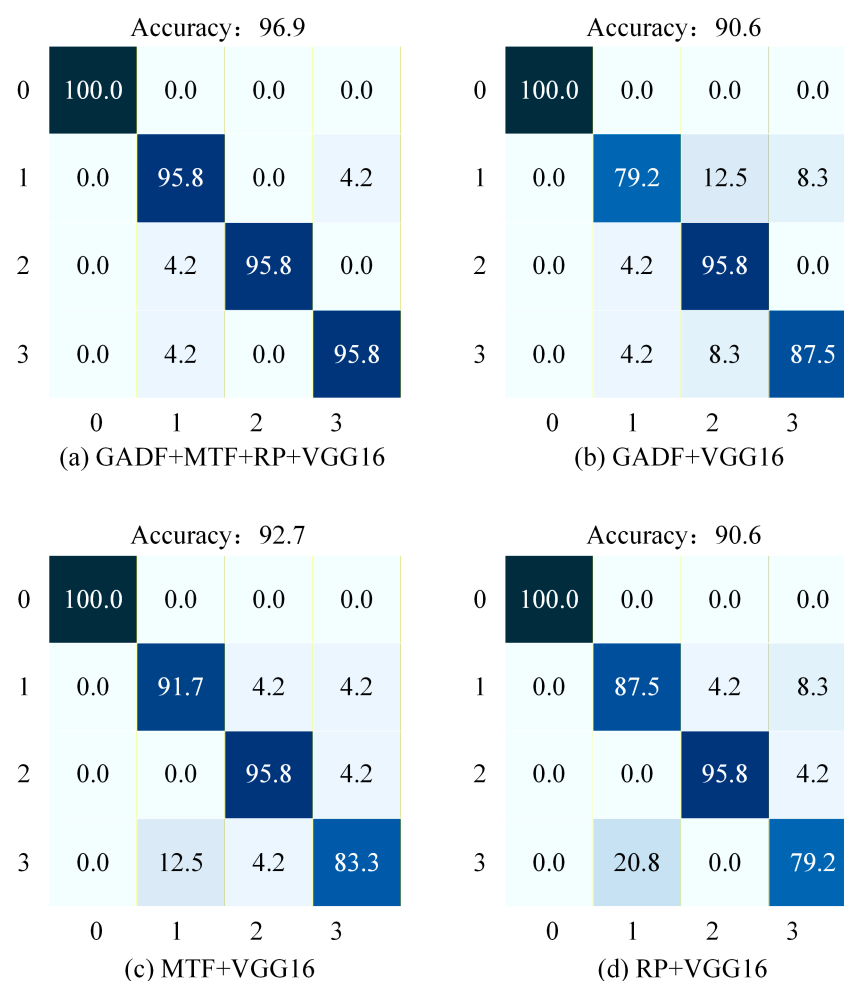


Figure 11. Confusion matrices of the four methods under 6 dB pink noise. (a) Proposed GMRVGG (Tri-modal). (b) GADF + VGG16. (c) MTF + VGG16. (d) RP + VGG16. The color intensity represents the percentage of classification accuracy, with darker shades indicating higher values.

As illustrated in Figure 11c, the MTF + VGG16 model achieves an overall accuracy of 92.7%. The classification accuracy for the Normal state reaches 100%. The model performs well on Inner Race Faults (95.8%) and Ball Faults (91.7%). However, the accuracy for Outer Race Faults is relatively lower at 83.3%, with the main error being misclassification as Ball Faults (12.5%). This suggests that under pink noise interference, the MTF modality faces certain challenges in distinguishing between Outer Race and Ball fault patterns.

As illustrated in Figure 11d, the RP + VGG16 model achieves an overall accuracy of 90.6%. Specifically, it attains 100.0% accuracy in classifying the Normal state and 95.8% for Inner Race Faults. However, the model exhibits significant instability in diagnosing Outer Race Faults, where the accuracy drops to 79.2%. The primary source of error is a substantial misclassification of Outer Race Faults as Ball Faults (20.8%). Meanwhile, Ball Faults are identified with 87.5% accuracy. Overall, this method demonstrates a specific deficiency in extracting discriminative features for Outer Race Faults in the presence of pink noise.

In summary, under the 6 dB pink noise condition, the proposed method achieves significantly higher classification accuracy compared to the other three single-modal approaches. By effectively fusing the features from three modalities, the proposed model compensates

for the specific shortcomings of individual modalities in identifying certain fault types, thereby demonstrating superior robustness against low-frequency noise interference.

4.2. Experiment II: Diagnostic Performance Analysis on Ottawa Dataset

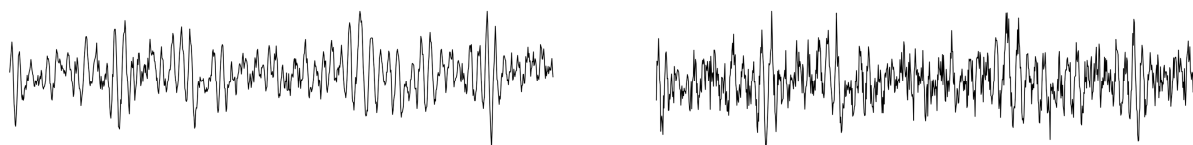
4.2.1. Dataset Introduction

To further validate the generalization capability of the proposed method, the bearing fault diagnosis dataset from the University of Ottawa is employed for experimental validation. This dataset was developed by the Department of Mechanical Engineering at the University of Ottawa. The experimental test rig is configured with a single-phase motor that drives the shaft and a cantilever beam mechanism for applying radial loads [37]. The experiments utilized FAFNIR 203KD ball bearings (Fafnir Bearing Company, New Britain, CT, USA), which are also standard metallic bearings.

Distinct from the CWRU dataset where faults are artificially induced by EDM, the faults in this dataset are generated through accelerated deterioration experiments. By removing the seals and degreasing the drive-end bearings, the bearings are run from a healthy state to failure, thereby simulating the natural evolution of bearing faults. The motor operates at a constant rotational speed of 1750 RPM. The load conditions are maintained at 400 N for inner race faults, outer race faults and cage faults, while ball faults are generated under a no-load condition to facilitate natural fault progression. The dataset covers different health conditions, categorized as healthy, developing faults, and faulty states. The bearing fault signals are sampled at 42,000 Hz.

For the experiments in this section, vibration signals acquired at a sampling frequency of 42,000 Hz were selected. Inner Race faults, Outer Race faults, and Cage faults were chosen. It is important to note that Ball faults were excluded from this experiment to ensure rigorous control of variables. According to the dataset specifications, Ball faults were generated under a no-load condition, whereas all other health states (Normal, Inner Race, Outer Race, and Cage) were recorded under a consistent nominal load of 400 N. Excluding Ball faults eliminates load variation as a confounding factor, ensuring that the classification is based solely on fault characteristics. For each selected fault type, signals corresponding to both ‘developing fault’ and ‘faulty’ states were selected and combined with signals from the normal state, resulting in a total of seven distinct categories of bearing fault signal data. Consistent with the previous categorization strategy, signals of the same fault type but representing different severity levels are grouped under the same label, resulting in four final categories: ‘Normal’, ‘IR’, ‘OR’, and ‘Cage’.

To ensure a consistent and rigorous evaluation of the model’s generalization ability, the noise injection protocol employed in Section 4.1 was strictly replicated for the Ottawa dataset. Consistent with the previous experiment, both 6 dB Gaussian white noise and 6 dB pink noise were individually introduced into the raw signals prior to sample segmentation. A comparison of the signal waveforms before and after adding 6 dB Gaussian white noise is shown in Figure 12.



(a) Waveform of the Signal Before Noise Addition (b) Waveform of the Signal After Noise Addition

Figure 12. Signal waveform comparison under 6 dB Gaussian white noise. (a) Original signal before noise addition. (b) Signal after adding 6 dB Gaussian white noise.

As shown in Figure 12, the signal waveform undergoes a significant transformation, similar to that observed in the CWRU dataset. The introduction of 6 dB Gaussian white noise leads to a pronounced increase in random irregularities and substantial fluctuations in amplitude, validating the successful simulation of a harsh noisy environment. A comparison of the signal waveforms before and after adding 6 dB pink noise is shown in Figure 13.

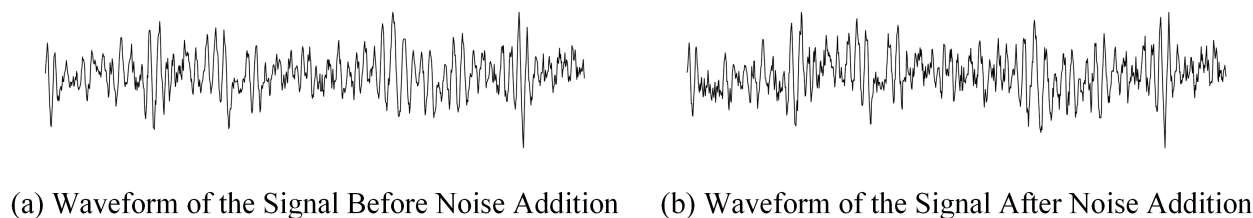


Figure 13. Signal waveform comparison under 6 dB pink noise. (a) Original signal before noise addition. (b) Signal after adding 6 dB pink noise.

As shown in Figure 13, the waveform of the signal also exhibits significant changes after the addition of 6 dB pink noise.

To prevent data leakage and ensure experimental consistency, the data partitioning strategy employed in Section 4.1 was strictly followed. For each noise scenario, the data points were partitioned into training and test sets in an 8:2 ratio. Subsequently, samples were obtained through sequential segmentation with a sample length of 512 data points and an overlap ratio of 50%. The details of the processed Ottawa dataset partition are listed in Table 7.

Table 7. Processed Ottawa Dataset.

Fault	Label	Dataset (Train/Test)
IR	1	108/24
OR	2	108/24
Cage	3	108/24
Normal	0	108/24

As shown in Table 7, the dataset is categorized into four classes (Normal, IR, OR, Cage) labeled 0 to 3. Consistent with the data balancing strategy employed in Section 4.1, a strictly balanced dataset was constructed for this experiment as well. For each noise scenario, every category comprises 108 training samples and 24 test samples, ensuring a uniform distribution for fair model evaluation.

4.2.2. Experimental Results

(1) Performance under 6 dB Gaussian white noise

For the method proposed in this paper, the experimental results on the Ottawa dataset under the 6 dB Gaussian white noise condition are listed in Table 8.

Table 8. Method comparison under 6 dB Gaussian white noise.

Method	IR	OR	Cage	Normal	Overall Accuracy
GADF + VGG16	79.2%	87.5%	100.0%	95.8%	90.6%
MTF + VGG16	58.3%	100.0%	79.2%	100.0%	84.4%
RP + VGG16	70.8%	95.8%	83.3%	100.0%	87.5%
Proposed method	83.3%	100.0%	100.0%	100.0%	95.8%

As can be seen from Table 8, in this noise environment, the proposed method achieves significantly higher accuracy than other single-modal methods, reaching an overall accuracy of 95.8%. Under the interference of 6 dB Gaussian white noise, single-modal methods exhibit distinct vulnerabilities in identifying specific fault types. For instance, the MTF + VGG16 model shows a severe deficiency in diagnosing Inner Race Faults, with an accuracy of only 58.3%. Similarly, RP + VGG16 struggles to distinguish Inner Race Faults (70.8%) and Cage Faults (83.3%). In contrast, the tri-modal fusion approach effectively compensates for these single-modal limitations, achieving 100% accuracy in recognizing Normal, Outer Race Faults, and Cage Faults, and significantly improving the recognition rate of Inner Race Faults to 83.3%. Consistent with the previous experiment, these variations in performance are solely attributed to the feature extraction capabilities of the different methods, as the dataset remains strictly balanced.

To intuitively assess the classification performance on the Ottawa dataset under 6 dB Gaussian white noise, the confusion matrices for the four methods are displayed in Figure 14. In these matrices, the horizontal axis represents the predicted labels, while the vertical axis represents the true labels, where 0 represents the Normal condition, 1 represents the Inner Race Fault, 2 represents the Outer Race Fault, and 3 represents the Cage Fault.

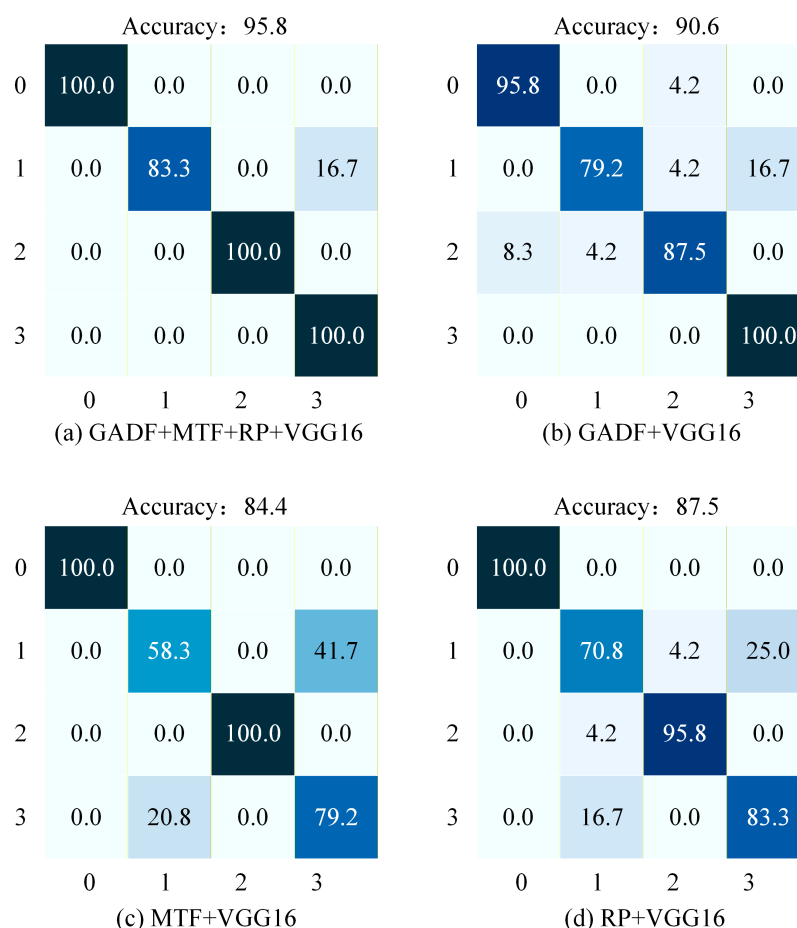


Figure 14. Confusion matrices of the four methods under 6 dB Gaussian white noise. (a) Proposed GMRVGG (Tri-modal). (b) GADF + VGG16. (c) MTF + VGG16. (d) RP + VGG16. The color intensity represents the percentage of classification accuracy, with darker shades indicating higher values.

As illustrated in Figure 14a, the proposed GMRVGG16 model achieves a superior overall accuracy of 95.8% and demonstrates exceptional stability. Specifically, the classification accuracy for the Normal state, Outer Race Faults, and Cage Faults all reach a perfect

100% with no misclassifications. The only minor confusion occurs in identifying Inner Race Faults, where the accuracy is 83.3%, with the remaining samples being misclassified as Cage Faults (16.7%). Overall, the proposed model effectively captures the discriminative features of different faults, minimizing confusion even in the presence of white noise.

As illustrated in Figure 14b, the GADF + VGG16 model achieves an overall accuracy of 90.6%. It attains 100% accuracy in classifying Cage Faults and 95.8% for the Normal state. However, its performance drops for Inner Race Faults (79.2%), with misclassifications occurring as Cage Faults (16.7%) and Outer Race Faults (4.2%). Similarly, for Outer Race Faults, the accuracy is 87.5%, with some confusion with Normal states (8.3%). Overall, while GADF performs well on Cage faults, it exhibits limitations in accurately distinguishing Inner Race faults.

As illustrated in Figure 14c, the MTF + VGG16 model achieves the lowest overall accuracy of 84.4%, revealing a severe deficiency in distinguishing specific fault types. While it perfectly identifies Normal and Outer Race conditions (100%), it fails significantly in diagnosing Inner Race Faults, achieving an accuracy of only 58.3%. A substantial portion of Inner Race samples (41.7%) are misclassified as Cage Faults. Conversely, Cage Faults (79.2%) are also frequently misclassified as Inner Race Faults (20.8%). This indicates that the MTF modality struggles to extract unique features to separate Inner Race and Cage Faults under white noise interference.

As illustrated in Figure 14d, the RP + VGG16 model achieves an overall accuracy of 87.5%, showing a similar pattern of confusion. Specifically, it attains 100.0% accuracy for the Normal state and 95.8% for Outer Race Faults. However, the accuracy for Inner Race Faults is low at 70.8%, dominated by misclassification as Cage Faults (25.0%). For Cage Faults, the accuracy is 83.3%, with primary errors being misclassified as Inner Race Faults (16.7%). Overall, this method demonstrates a specific weakness in distinguishing between Inner Race and Cage faults.

In summary, under the 6 dB Gaussian white noise condition, the proposed method achieves significantly higher classification accuracy compared to the other three single-modal approaches, demonstrating its effectiveness on the Ottawa dataset.

(2) Performance under 6 dB pink noise

The experiments were repeated using the dataset added with 6 dB pink noise. The comparison results between the proposed method and other single-modal methods are listed in Table 9.

Table 9. Method comparison under 6 dB pink noise.

Method	IR	OR	Cage	Normal	Overall Accuracy
GADF + VGG16	83.3%	87.5%	100.0%	95.8%	91.7%
MTF + VGG16	79.2%	91.7%	95.8%	95.8%	90.6%
RP + VGG16	75.0%	100.0%	91.7%	95.8%	90.6%
Proposed method	83.3%	100.0%	100.0%	100.0%	95.8%

As can be seen from Table 9, in this noise environment, the proposed method continues to demonstrate outstanding performance, achieving the highest overall accuracy of 95.8%. In contrast, single-modal methods still exhibit specific vulnerabilities when facing pink noise interference. For instance, the RP + VGG16 model struggles to accurately identify Inner Race Faults, with an accuracy of only 75.0%. Similarly, the MTF + VGG16 model shows limited capability in distinguishing Inner Race Faults (79.2%). While the GADF + VGG16 model achieves perfect recognition for Cage Faults, its performance on Outer Race Faults (87.5%) remains suboptimal. In comparison, the proposed tri-modal

fusion approach effectively integrates the strengths of individual features to overcome these limitations.

Consistent with previous experiments, these performance differences are solely attributed to the distinct feature representation capabilities of each method, as the dataset remains strictly balanced. The present experimental results further confirm that the proposed method possesses strong generalization capability, maintaining high reliability even under different noise patterns and mechanical systems.

To further visualize the detailed classification performance on the Ottawa dataset under the impact of pink noise, the confusion matrices for the four methods are displayed in Figure 15. Consistent with the previous setup, the horizontal axis represents the predicted labels, and the vertical axis represents the true labels. The class mapping remains identical to the white noise experiment, where 0, 1, 2, and 3 correspond to Normal condition, Inner Race Fault, Outer Race Fault, and Cage Fault, respectively.

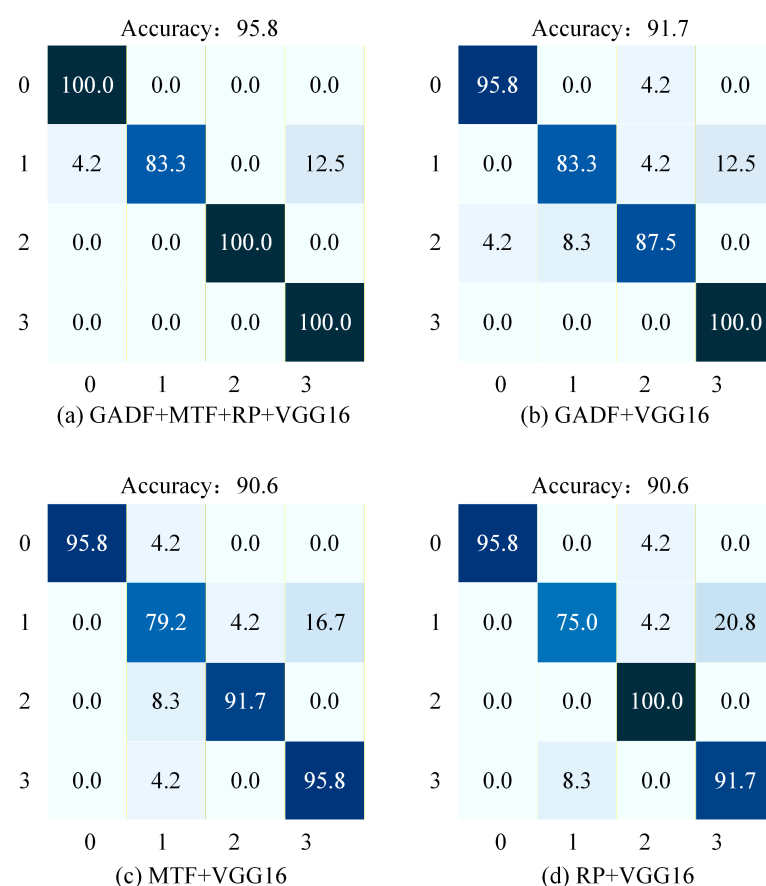


Figure 15. Confusion matrices of the four methods under 6 dB pink noise. (a) Proposed GMRVGG (Tri-modal). (b) GADF + VGG16. (c) MTF + VGG16. (d) RP + VGG16. The color intensity represents the percentage of classification accuracy, with darker shades indicating higher values.

As illustrated in Figure 15a, the proposed GMRVGG16 model achieves a superior overall accuracy of 95.8% and demonstrates exceptional stability under the impact of pink noise. Specifically, the classification accuracy for the Normal state, Outer Race Faults, and Cage Faults all reach a perfect 100% with no misclassifications. The primary source of error is confined to Inner Race Faults, where the accuracy is 83.3%. These misclassified samples are attributed to Normal states (4.2%) and Cage Faults (12.5%). Overall, despite the slight confusion in Inner Race patterns, the proposed model effectively integrates the strengths of multiple modalities to minimize errors, maintaining the highest diagnostic performance among all methods.

As illustrated in Figure 15b, the GADF + VGG16 model achieves an overall accuracy of 91.7%. It attains 100.0% accuracy in classifying Cage Faults, demonstrating high sensitivity to this fault type. However, its performance exhibits limitations in distinguishing other fault categories. The accuracy for Inner Race Faults is 83.3%, with misclassifications occurring as Outer Race Faults (4.2%) and Cage Faults (12.5%). Similarly, for Outer Race Faults, the accuracy is 87.5%, with confusion primarily involving Inner Race Faults (8.3%). This indicates that while the GADF modality is robust for Cage faults, it struggles to fully separate the features of Inner and Outer Race faults in a pink noise environment.

As illustrated in Figure 15c, the MTF + VGG16 model achieves an overall accuracy of 90.6%. While it maintains high precision for Normal states (95.8%) and Cage Faults (95.8%), it reveals a distinct deficiency in diagnosing Inner Race Faults. The accuracy for Inner Race Faults is limited to 79.2%, with a significant portion of samples being misclassified as Cage Faults (16.7%). Additionally, Outer Race Faults achieve an accuracy of 91.7%, with minor misclassification as Inner Race Faults (8.3%). This suggests that under pink noise interference, the MTF modality faces challenges in extracting unique features to distinguish Inner Race Faults from Cage Faults.

As illustrated in Figure 15d, the RP + VGG16 model achieves an overall accuracy of 90.6%, exhibiting a polarized performance distribution. It attains perfect classification (100.0%) for Outer Race Faults, outperforming the GADF and MTF models in this category. However, it demonstrates the lowest performance for Inner Race Faults, where the accuracy drops to 75.0%. The primary error is a substantial misclassification of Inner Race Faults as Cage Faults (20.8%). Overall, while the RP modality is highly effective for Outer Race features, it shows a specific weakness in capturing the discriminative patterns of Inner Race faults.

In summary, under the 6 dB pink noise condition, the proposed method achieves significantly higher classification accuracy compared to the other three single-modal approaches. These results further demonstrate the exceptional generalization capability and robustness of the proposed method on the Ottawa dataset.

4.3. Comprehensive Analysis and Evaluation

4.3.1. Ablation Study

To further investigate the specific contribution of each branch within the proposed tri-modal architecture and explicitly rule out the possibility of feature redundancy, an ablation study was conducted. Three dual-modal fusion fault diagnosis models were constructed by sequentially removing the RP, MTF, and GADF branches from the original three-branch model. To ensure consistency with the initial baseline performance analysis, these experiments were conducted on the CWRU dataset augmented with 6 dB Gaussian white noise. To maintain narrative conciseness and avoid redundant data presentation across all four distinct dataset and noise scenarios, these experiments were specifically conducted on the CWRU dataset augmented with 6 dB Gaussian white noise. This condition was selected as a representative benchmark to effectively demonstrate the structural necessity of the fusion strategy. The experimental results are listed in Table 10.

Table 10. Ablation Study Results.

Method	Ball	IR	OR	Normal	Overall Accuracy
GADF + MTF + VGG16	70.8%	95.8%	75.0%	100.0%	85.4%
GADF + RP + VGG16	83.3%	95.8%	79.2%	100.0%	89.6%
MTF + RP + VGG16	100.0%	83.3%	75.0%	100.0%	89.6%
Proposed method	87.5%	95.8%	83.3%	100.0%	91.7%

As shown in Table 10, the proposed method achieves higher accuracy than the three dual-modal fusion methods. Among the three dual-modal fusion methods, both GADF + RP + VGG16 and MTF + RP + VGG16 achieved the highest accuracy of 89.6%. However, the proposed tri-modal fusion method still outperforms these dual-modal approaches. Similarly, to more intuitively evaluate the model's performance, a confusion matrix was introduced to observe the effectiveness of fault classification. Figure 16 shows the confusion matrix from the ablation study. In these matrices, the horizontal axis represents the predicted labels, while the vertical axis represents the true labels, where 0 represents the Normal condition, 1 represents the Ball Fault, 2 represents the Inner Race Fault, and 3 represents the Outer Race Fault.

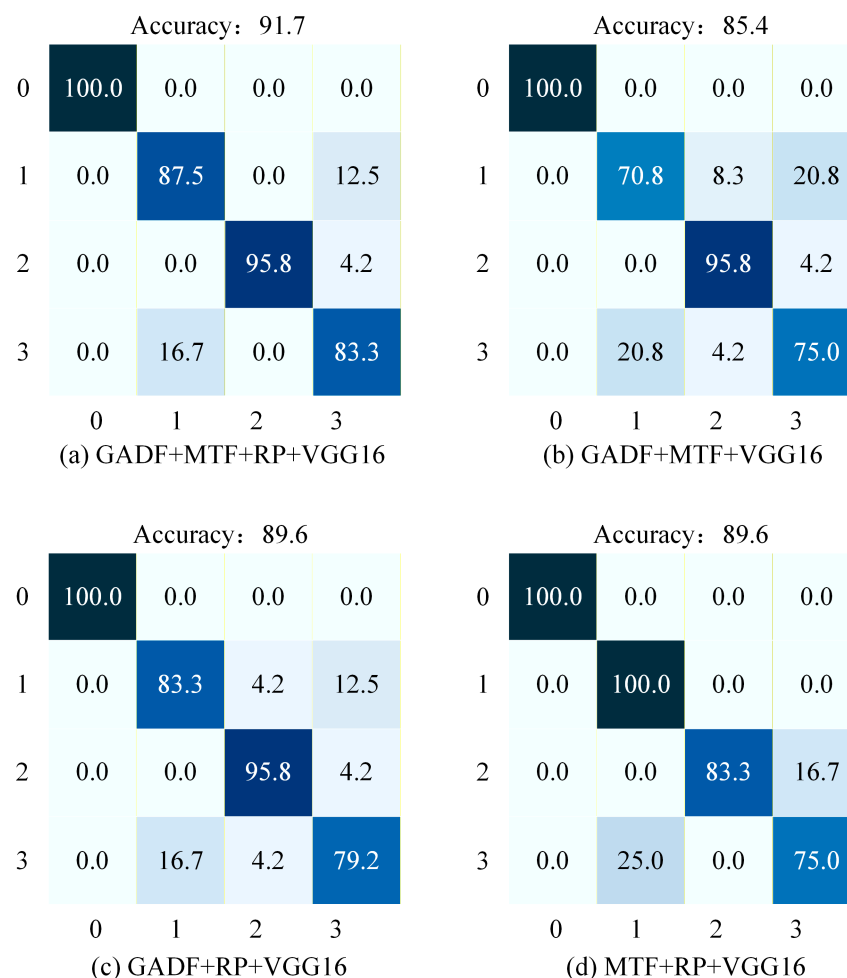


Figure 16. Confusion matrix for the ablation study. (a) Proposed GMRVGG (Tri-modal). (b) GADF + MTF + VGG16. (c) GADF + RP + VGG16. (d) MTF + RP + VGG16. The color intensity represents the percentage of classification accuracy, with darker shades indicating higher values.

As illustrated in Figure 16b, the GADF + MTF + VGG16 model achieves an overall accuracy of 85.4%. Specifically, it attains 100.0% accuracy in classifying the Normal state. However, the accuracy for Ball Faults is relatively low at 70.8%, with primary misclassifications occurring as Inner Race Faults (8.3%) and Outer Race Faults (20.8%). For Inner Race Faults, the accuracy reaches 95.8%, dominated by misclassification as Outer Race Faults (4.2%). The accuracy for Outer Race Faults is 75.0%, with errors including misclassification as Ball Faults (20.8%) and Inner Race Faults (4.2%).

As illustrated in Figure 16c, the GADF + RP + VGG16 model achieves an overall accuracy of 89.6%. It demonstrates perfect classification (100.0% accuracy) for the Normal

state. For Ball Faults, the accuracy is 83.3%, with primary misclassifications occurring as Inner Race Faults (4.2%) and Outer Race Faults (12.5%). The accuracy for Inner Race Faults reaches 95.8%, dominated by misclassification as Outer Race Faults (4.2%). However, the accuracy for Outer Race Faults is 79.2%, with main errors including misclassification as Ball Faults (16.7%) and Inner Race Faults (4.2%).

As illustrated in Figure 16d, the MTF + RP + VGG16 model also achieves an overall accuracy of 89.6%. It demonstrates perfect classification (100.0% accuracy) for both the Normal state and Ball Faults. However, the accuracy for Outer Race Faults is the lowest at 75.0%, with the primary error being misclassification as Ball Faults (25.0%). For Inner Race Faults, the accuracy is 83.3%, dominated by misclassification as Outer Race Faults (16.7%).

Overall, both the GADF + RP + VGG16 and MTF + RP + VGG16 models demonstrate relatively good performance; however, the overall accuracy of the proposed method remains higher than these two approaches. It is noteworthy that for Ball Faults, the proposed method exhibits lower accuracy compared to the MTF + RP + VGG16 model. This observation indicates that the feature vectors extracted from GADF images, when fused, exert a certain adverse impact on the model's diagnostic performance for this specific fault type. To deeply investigate the cause of this specific inter-modality conflict and to mathematically prove the overall necessity of fusing these three distinct modalities, a quantitative feature similarity analysis is conducted in the following Section 4.3.2.

In summary, this ablation study systematically validates the contribution of each branch within the multimodal fusion architecture by comparing the performance of the proposed complete model against three dual-modal simplified models (constructed by sequentially removing the RP, MTF, or GADF branches). The experimental data confirms that the tri-modal fusion strategy significantly enhances the comprehensive performance of fault diagnosis. With the complete model achieving an overall accuracy of 91.7%, it outperforms all dual-modal combinations, justifying the necessity of the proposed tri-modal framework.

Additionally, a brief discussion on interpretability is provided here. It is acknowledged that the feature vectors extracted by the CNN backbone are high-level abstract representations generated through complex nonlinear mappings. However, their physical interpretability can be inferred from the encoding mechanisms of the input modalities and is further validated by the results of ablation studies. Specifically, since the GADF mathematically maps the 1D signal to a 2D matrix based on the trigonometric difference between each pair of time points, the features extracted from this branch inherit information about the temporal correlations and amplitude fluctuations of the signal. Similarly, as the MTF images are constructed from transition probabilities, the corresponding features effectively encode the dynamic statistical properties of the system, reflecting the probabilities of transitioning between different energy states. Furthermore, given that the RP visualizes the phase space trajectories, the extracted features primarily characterize the periodicity and non-stationarity of the fault impulses. This interpretation is further supported by the results in Table 10, where removing any single branch leads to a noticeable decline in accuracy. This implies that the specific physical information provided by each modality has been successfully captured by the network and plays a unique, non-redundant role in the final fault diagnosis.

4.3.2. Feature Redundancy and Modality Conflict Analysis

To quantify the inter-modality relationships and investigate the performance anomaly observed during the Ball Fault diagnosis (Section 4.3.1), we calculated the cosine similarity among the 4096-dimensional feature vectors extracted by the GADF, MTF, and RP branches.

Consistent with the ablation study, this analysis was conducted using the CWRU dataset augmented with 6 dB Gaussian white noise. The results are illustrated in Figure 17.

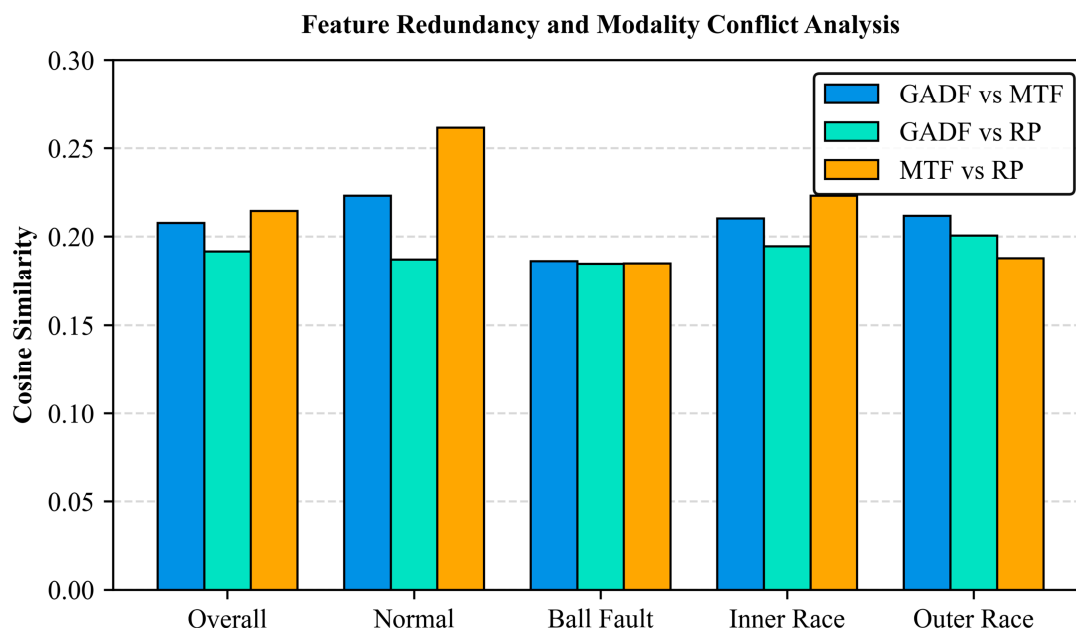


Figure 17. Quantitative analysis of feature similarity among the GADF, MTF, and RP branches across different health conditions. The consistently low cosine similarity indicates strong complementarity.

As depicted in Figure 17, the average cosine similarities among the three modalities are consistently low, hovering around 0.20. In high-dimensional spaces, such near-zero values indicate that the extracted features are largely orthogonal. This suggests that the GADF, MTF, and RP branches successfully capture distinct, non-overlapping physical properties of the signals. This lack of feature redundancy confirms their strong complementarity, providing a clear rationale for the tri-modal fusion strategy.

Notably, the similarity drops to its lowest level under the Ball Fault condition. This phenomenon helps explain the slight performance degradation of the GADF-related combinations observed in the ablation study. Since Ball Fault impulses are inherently weak and easily masked by the severe 6 dB background noise, the GADF encoding—which computes trigonometric differences between temporal points—becomes highly sensitive to high-frequency random interference. Consequently, the GADF feature space misaligns with those of MTF and RP, leading to a geometric deviation or “modality conflict.” While this conflict introduces minor interference during fusion, the overall tri-modal framework effectively neutralizes it. By integrating the robust topological dynamics from RP and the stable statistical transitions from MTF, the network successfully compensates for GADF’s vulnerability under extreme noise.

4.3.3. Comparison with State-of-the-Art Methods

To further evaluate the comprehensive performance and environmental robustness of the proposed GMRVGG, it was compared against three recent advanced 1D deep learning methods: 1D-ResNet, Attention-based BiTCN (Attn-BiTCN), and Multi-Scale Dilated Convolution and Dense Temporal Convolutional Network (MSDC-DenseTCN). To ensure a fair comparison, all baseline models were trained and tested under the exact same configurations across four distinct diagnostic scenarios, encompassing two datasets (CWRU and Ottawa) and two severe noise conditions (6 dB White noise and Pink noise). The diagnostic accuracies are summarized in Table 11.

Table 11. Comparative Experimental Results.

Method	CWRU	CWRU	Ottawa	Ottawa
	6 dB	6 dB	6 dB	6 dB
Noisy Environment	White Noise	Pink Noise	White Noise	Pink Noise
1D-ResNet	47.9%	39.6%	47.9%	50.0%
Attn-BiTCN	88.5%	68.8%	100.0%	94.8%
MSDC-DenseTCN	65.6%	75.6%	100.0%	85.42%
Proposed method	91.7%	96.9%	95.8%	95.8%

As shown in Table 11, the 1D baseline models exhibit significant sensitivity to the type of background noise, revealing a critical cross-domain vulnerability. For instance, while Attn-BiTCN achieves a 100% accuracy on the Ottawa dataset under white noise, its performance drastically drops to 68.75% when applied to the CWRU dataset under low-frequency pink noise. This severe fluctuation indicates that traditional 1D temporal models tend to overfit to specific frequency distributions and struggle to maintain robustness when the spectral characteristics of the interference change.

In contrast, GMRVGG maintains a more consistent accuracy range of 91.7% to 96.9% across the four evaluated scenarios. The results indicate that mapping 1D signals into multiple 2D feature spaces (GADF, MTF, and RP) can reduce the impact of spectral variations in background noise. While the 1D state-of-the-art models also achieve competitive performance, the proposed method demonstrates more consistent stability under the evaluated experimental conditions. These findings suggest that the proposed framework offers a viable alternative for fault diagnosis in environments with varying interference.

However, compared to lightweight 1D methods, the proposed method requires higher computational overhead due to the parallel execution of three VGG16 backbones. Therefore, direct deployment on memory-limited portable devices is challenging. To enable edge deployment, exploring model compression techniques will be a necessary direction for future research.

5. Conclusions

To address the limitations of existing bearing fault diagnosis methods, which often rely on single-modal data or a solitary signal-to-image conversion technique—leading to insufficient information dimensionality and inadequate feature representation—this paper proposes a bearing fault diagnosis method (GMRVGG) based on tri-modal image feature fusion. First, in order to rigorously simulate complex real-world industrial environments, severe 6 dB Gaussian white noise and 6 dB pink noise were injected into the raw signals prior to sample partitioning. Subsequently, the 1D vibration signals are converted into 2D images by employing three distinct signal-to-image transformation methods: GADF, MTF, and RP. Shallow to deep features are extracted from these images using the VGG16 architecture, and the resulting feature vectors are fused via concatenation. Finally, fault diagnosis is accomplished by a fully connected classifier. Extensive experiments conducted on both the Case Western Reserve University (CWRU) and the University of Ottawa bearing datasets demonstrate the exceptional robustness of the proposed method. It achieved overall accuracies of up to 96.9% and 95.8% on the two datasets respectively, consistently outperforming traditional single-modal approaches across different noise interferences. This effectively addresses the issues of insufficient information dimensionality and inadequate feature representation, providing a highly reliable and generalizable solution for intelligent bearing condition monitoring.

Nevertheless, the current evaluation is based on constant load and speed conditions, which still differs from actual industrial datasets. To further bridge the gap between

laboratory research and complex industrial applications, our future work will focus on addressing non-stationary operating conditions and cross-sensor variability. Specifically, inspired by the advanced digital twin-guided physical-virtual denoising framework [38] and the innovative concept of positive-incentive noise [39], we plan to explore these directions by integrating domain adaptation techniques to further enhance diagnostic intelligence and reliability under extreme industrial conditions.

Furthermore, to ensure the long-term adaptability and computational efficiency of the model in continuous operations, we will also explore incremental fault diagnosis methodologies [40]. Integrating lifelong learning strategies with dynamic learnable pruning mechanisms [41] will be a crucial next step to evolve our current static network into a lightweight, continuously adaptive framework capable of learning emerging fault patterns without catastrophic forgetting.

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

Funding: This research was supported by the National Natural Science Foundation of China, [No. 52502531], the Natural Science Foundation of Shandong Province, China, [Grant No. ZR2023QE 214 and No. ZR2025MS765], Opening Foundation of Shandong Key Laboratory of Intelligent Manufacturing Technology for Advanced Power Equipment, Weifang University, China [No. SKLOIMT-FAPE26013].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in the Case Western Reserve University Bearing Data Center Website and Mendeley Data at <https://engineering.case.edu/bearingdatacenter> (accessed on 1 September 2025) and <https://data.mendeley.com/datasets/y2px5tg92h/5> (accessed on 1 September 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

1D	One Dimensional
2D	Two Dimensional
Attn-BiTCN	Attention-based Bidirectional Temporal Convolutional Network
BiTCN	Bidirectional Temporal Convolutional Network
BN/BatchNorm	Batch Normalization
CNN	Convolutional Neural Networks
CWRU	Case Western Reserve University
DenseTCN	Dense Temporal Convolutional Network
EDM	Electric Discharge Machining
GADF	Gramian Angular Difference Field
GAF	Gramian Angular Field
GMRVGG	GADF-MTF-RP-VGG16
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IR	Inner Race

MSDC	Multi-Scale Dilated Convolution
MTF	Markov Transition Field
OR	Outer Race
ReLU	Rectified Linear Unit
ResNet	Residual Networks
RNN	Recurrent Neural Networks
RP	Recurrence Plot
RPM	Revolutions Per Minute
SST	Synchrosqueezed Wavelet Transform
SVMD	Successive Variational Mode Decomposition
VGG	Visual Geometry Group

References

1. Zhang, X.; Zhao, B.; Lin, Y. Machine Learning Based Bearing Fault Diagnosis Using the Case Western Reserve University Data: A Review. *IEEE Access* **2021**, *9*, 155598–155608. [\[CrossRef\]](#)
2. Yu, X.; Gao, W.; Feng, Y.; Shi, G.; Li, S.; Chen, M.; Zhang, R.; Wang, J.; Jia, W.; Jiao, J.; et al. Research Progress of Hydrostatic Bearing and Hydrostatic-Hydrodynamic Hybrid Bearing in High-End Computer Numerical Control Machine Equipment. *Int. J. Precis. Eng. Manuf.* **2023**, *24*, 1053–1081. [\[CrossRef\]](#)
3. Xu, J.; Li, C.; Miao, X.; Zhang, C.; Yuan, X. An Overview of Bearing Candidates for the Next Generation of Reusable Liquid Rocket Turbopumps. *Chin. J. Mech. Eng.* **2020**, *33*, 26. [\[CrossRef\]](#)
4. Peng, H.; Zhang, H.; Shangguan, L.; Fan, Y. Review of Tribological Failure Analysis and Lubrication Technology Research of Wind Power Bearings. *Polymers* **2022**, *14*, 3041. [\[CrossRef\]](#)
5. Wang, Z.; Zhou, Y.; Chu, Y.; Han, J.; Zhang, L.; Wang, R. Fault diagnosis of agricultural machinery bearings based on multimodal feature fusion. *Appl. Eng. Agric.* **2025**, *41*, 279–288. [\[CrossRef\]](#)
6. Hakim, M.; Omran, A.A.B.; Ahmed, A.N.; Al-Waily, M.; Abdellatif, A. A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations. *Ain Shams Eng. J.* **2023**, *14*, 101945. [\[CrossRef\]](#)
7. Mishra, R.K.; Choudhary, A.; Fatima, S.; Mohanty, A.R.; Panigrahi, B.K. A systematic review on advancement and challenges in multi-fault diagnosis of rotating machines. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111306. [\[CrossRef\]](#)
8. Hoang, D.-T.; Kang, H.-J. A survey on Deep Learning based bearing fault diagnosis. *Neurocomputing* **2019**, *335*, 327–335. [\[CrossRef\]](#)
9. Cui, J.; Gao, J.; Xing, R.; Wu, W.; Li, M.; Yang, Y. Fault diagnosis method for rolling bearing based on attention mechanism and BiTCN model. *Digit. Signal Process.* **2025**, *167*, 105454. [\[CrossRef\]](#)
10. Wang, Y.; Ding, X.; Zeng, Q.; Wang, L.; Shao, Y. Intelligent Rolling Bearing Fault Diagnosis via Vision ConvNet. *IEEE Sens. J.* **2021**, *21*, 6600–6609. [\[CrossRef\]](#)
11. Si, X.; Yan, H.; Hu, Y.; Duan, J.; Shi, T. Bearing fault diagnosis under heavy noise: A multi-scale dilated convolution and dense temporal convolutional network. *Neurocomputing* **2025**, *652*, 131092. [\[CrossRef\]](#)
12. Duan, Y.; Yang, T.; Wang, C.; Zhang, Y.; Han, Q.; Guo, S. LSBT-Net: A lightweight framework for fault diagnosis of bearings based on an interpretable spatial-temporal model. *Expert Syst. Appl.* **2025**, *281*, 127718. [\[CrossRef\]](#)
13. Wang, Z.; Oates, T. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In Proceedings of the Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–26 January 2015; pp. 91–96.
14. Daubechies, I.; Lu, J.; Wu, H.-T. Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. *Appl. Comput. Harmon. Anal.* **2011**, *30*, 243–261. [\[CrossRef\]](#)
15. Li, Y.; Zhang, H.; Ma, S.; Cheng, G.; Yao, Q.; Zuo, C. A Novel Method Based on Stepwise Variational Modal Decomposition and Gramian Angular Difference Field for Bearing Health Monitoring. *Arab. J. Sci. Eng.* **2024**, *49*, 15773–15786. [\[CrossRef\]](#)
16. Wang, M.; Wang, W.; Zhang, X.; Iu, H.H.-C. A New Fault Diagnosis of Rolling Bearing Based on Markov Transition Field and CNN. *Entropy* **2022**, *24*, 751. [\[CrossRef\]](#)
17. Sun, Y.; Wang, W. Role of image feature enhancement in intelligent fault diagnosis for mechanical equipment: A review. *Eng. Fail. Anal.* **2024**, *156*, 107815. [\[CrossRef\]](#)
18. Liang, H.; Cao, J.; Zhao, X. Average Descent Rate Singular Value Decomposition and Two-Dimensional Residual Neural Network for Fault Diagnosis of Rotating Machinery. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 3512616. [\[CrossRef\]](#)
19. Kumar, A.; Wu, H.; Yang, Y.; Deng, S.; Wang, Q.; Song, H. GADF-VGG16 based fault diagnosis method for HVDC transmission lines. *PLoS ONE* **2022**, *17*, e0274613. [\[CrossRef\]](#)

20. Wang, Q.; Pian, F.; Wang, M.; Song, S.; Li, Z.; Shan, P.; Ma, Z. Quantitative analysis of Raman spectra for glucose concentration in human blood using Gramian angular field and convolutional neural network. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *275*, 121189. [CrossRef]
21. Bian, H.; Zhang, L. Gesture recognition based on Gramian angular difference field and multi-stream fusion methods. *Signal Image Video Process.* **2024**, *19*, 120. [CrossRef]
22. Wei, L.; Peng, X.; Cao, Y. Rolling bearing fault diagnosis based on Gramian angular difference field and improved channel attention model. *PeerJ Comput. Sci.* **2024**, *10*, e1807. [CrossRef]
23. Cai, C.; Xu, T.; Ren, J.; Xue, Y. Bearing Fault Diagnosis Based on the Markov Transition Field and SE-IShuffleNetV2 Model. *Struct. Durab. Health Monit.* **2025**, *19*, 125–144. [CrossRef]
24. Yan, J.; Kan, J.; Luo, H. Rolling Bearing Fault Diagnosis Based on Markov Transition Field and Residual Network. *Sensors* **2022**, *22*, 3936. [CrossRef]
25. Sun, W.; Zhou, J.; Sun, B.; Zhou, Y.; Jiang, Y. Markov Transition Field Enhanced Deep Domain Adaptation Network for Milling Tool Condition Monitoring. *Micromachines* **2022**, *13*, 873. [CrossRef]
26. Pu, Y.; Chen, J.; Sheng, X.; Wang, Z.; Li, J.; Tang, H. Data-driven fault diagnosis with Markov transition field and deep learning for water management of automotive PEMFC power system. *Int. J. Hydrogen Energy* **2025**, *148*, 150022. [CrossRef]
27. Marwan, N.; Kurths, J.; Foerster, S. Analysing spatially extended high-dimensional dynamics by recurrence plots. *Phys. Lett. A* **2015**, *379*, 894–900. [CrossRef]
28. Petrauskiene, V.; Pal, M.; Cao, M.; Wang, J.; Ragulskis, M. Color Recurrence Plots for Bearing Fault Diagnosis. *Sensors* **2022**, *22*, 8870. [CrossRef]
29. Sankararaman, S. Recurrence network and recurrence plot: A novel data analytic approach to molecular dynamics in thermal lensing. *J. Mol. Liq.* **2022**, *366*, 120353. [CrossRef]
30. Zhou, Y.; Wang, Z.; Zuo, X.; Zhao, H. Identification of wear mechanisms of main bearings of marine diesel engine using recurrence plot based on CNN model. *Wear* **2023**, *520–521*, 204656. [CrossRef]
31. Shaari, F.N.; Abdul Nasir, A.S.; Mustafa, W.A.; Wan Ahmed, W.A.; Abdull Sukor, A.S. Attention-enhanced hybrid CNN–LSTM network with self-adaptive CBAM for COVID-19 diagnosis. *Array* **2025**, *26*, 100424. [CrossRef]
32. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
33. Zhou, Y.; Wang, X.; Zhang, M.; Zhu, J.; Zheng, R.; Wu, Q. MPCE: A Maximum Probability Based Cross Entropy Loss Function for Neural Network Classification. *IEEE Access* **2019**, *7*, 146331–146341. [CrossRef]
34. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
35. Available online: <https://engineering.case.edu/bearingdatacenter/welcome> (accessed on 1 September 2025).
36. Huo, Z.; Zhang, Y.; Shu, L.; Gallimore, M. A New Bearing Fault Diagnosis Method Based on Fine-to-Coarse Multiscale Permutation Entropy, Laplacian Score and SVM. *IEEE Access* **2019**, *7*, 17050–17066. [CrossRef]
37. Sehri, M.; Dumond, P.; Bouchard, M. University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets. *Data Brief* **2023**, *49*, 109327. [CrossRef]
38. Qiao, Z.; Ning, S.; Gai, Y.; Xie, C. A digital twin guided physical-virtual denoising method for early fault detection of rolling element bearings. *Mech. Syst. Signal Process.* **2026**, *249*, 114108. [CrossRef]
39. Yang, C.; Qiao, Z.; Liu, L.; Kumar, A.; Zhu, R. Positive-incentive noise in artificial intelligence-enabled machine fault diagnosis. *Struct. Health Monit.* **2025**, 14759217251370358. [CrossRef]
40. Tang, H.-H.; Li, J.-W.; Feng, W.-W.; Chen, P.; Xue, H.-T. Towards machinery incremental fault diagnosis based on inverted transformer lifelong learning with learnable pruning mechanism. *Eng. Appl. Artif. Intell.* **2025**, *152*, 110763. [CrossRef]
41. Tang, H.; Zu, X.; Guoa, Y.; Jiang, X.; Wang, J.; Lin, R.; Xue, H.; Wang, H. A novel incremental method with dynamic learnable pruning mechanism for low-speed machinery fault diagnosis. *Eng. Appl. Artif. Intell.* **2026**, *166*, 113562. [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.