

## Article

# A Novel Method for Rolling Bearing Fault Diagnosis Based on Gramian Angular Field and CNN-ViT

Zijun Zhou <sup>1</sup>, Qingsong Ai <sup>1</sup>, Ping Lou <sup>1</sup>, Jianmin Hu <sup>2</sup> and Junwei Yan <sup>1,\*</sup>

<sup>1</sup> School of Information, Wuhan University of Technology, Wuhan 430070, China; zijunzhou@whut.edu.cn (Z.Z.); qingsongai@whut.edu.cn (Q.A.); louping@whut.edu.cn (P.L.)

<sup>2</sup> School of Information Engineering, Hubei University of Economics, Wuhan 430205, China; hujianmin@hbue.edu.cn

\* Correspondence: junweiyan@whut.edu.cn

**Abstract:** Fault diagnosis is one of the important applications of edge computing in the Industrial Internet of Things (IIoT). To address the issue that traditional fault diagnosis methods often struggle to effectively extract fault features, this paper proposes a novel rolling bearing fault diagnosis method that integrates Gramian Angular Field (GAF), Convolutional Neural Network (CNN), and Vision Transformer (ViT). First, GAF is used to convert one-dimensional vibration signals from sensors into two-dimensional images, effectively retaining the fault features of the vibration signal. Then, the CNN branch is used to extract the local features of the image, which are combined with the global features extracted by the ViT branch to diagnose the bearing fault. The effectiveness of this method is validated with two datasets. Experimental results show that the proposed method achieves average accuracies of 99.79% and 99.63% on the CWRU and XJTU-SY rolling bearing fault datasets, respectively. Compared with several widely used fault diagnosis methods, the proposed method achieves higher accuracy for different fault classifications, providing reliable technical support for performing complex fault diagnosis on edge devices.

**Keywords:** fault diagnosis; gramian angular field; convolutional neural network; vision transformer; sensors



**Citation:** Zhou, Z.; Ai, Q.; Lou, P.; Hu, J.; Yan, J. A Novel Method for Rolling Bearing Fault Diagnosis Based on Gramian Angular Field and CNN-ViT. *Sensors* **2024**, *24*, 3967. <https://doi.org/10.3390/s24123967>

Academic Editor: Alessandro Casavola

Received: 22 May 2024

Revised: 11 June 2024

Accepted: 17 June 2024

Published: 19 June 2024



**Copyright:** © 2024 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

With the advancement of Industry 4.0 and intelligent manufacturing, the Industrial Internet of Things (IIoT) is rapidly becoming a core component of modern industrial systems. Edge computing, by placing computing resources and data storage near the data generation source at edge nodes, greatly reduces data transmission latency and bandwidth requirements, providing a more efficient solution for data processing and analysis. Fault diagnosis is one of the important applications of edge computing in the industrial IoT. In the industrial IoT environment, devices and sensors are interconnected via networks, and data analysis and fault diagnosis are performed at the device end, enabling the system to respond to anomalies more quickly. However, fault diagnosis methods face greater challenges in this context. First, there is the issue of high latency and bandwidth limitations in data transmission. In traditional centralized data processing systems, all data must be transmitted to a central server for processing and analysis. This approach is impractical in an IIoT environment due to the enormous volume of data generated by industrial devices and sensors, which can cause network congestion, increase latency, and reduce the system's response speed. Secondly, edge nodes have limited resources. Edge computing devices usually have lower computing power and storage capacity compared to cloud servers, making resources relatively scarce. Furthermore, the heterogeneity of IIoT devices and the dynamic nature of the environment add to the difficulty of fault diagnosis. IIoT systems encompass various types of devices and sensors, each with different operating conditions, working modes, and fault types. Therefore, researching efficient fault diagnosis methods for

mechanical equipment is of significant importance for improving the operational efficiency and lifespan of individual devices and ensuring the stable operation and continuous development of industrial systems.

Rolling bearings are critical components in rotating machinery, and their health, such as cracks or faults in different parts, directly affects the performance, efficiency, stability, and lifespan of the equipment [1]. As industry develops, most rotating machinery needs to withstand enormous workloads. When rolling bearings operate under high loads, strong impacts, heavy workloads, and complex environments, faults often occur in the inner ring, outer ring, balls, and cage. These faults cause the characteristics of the vibration sensor signals to differ from the normal vibration patterns, and fault information is embedded in these vibration signals. If these faults are not detected and addressed promptly, the equipment often stops working, leading to significant economic losses and potentially causing safety accidents. Fault diagnosis methods for bearings can be divided into two categories: mechanistic model-based methods and data-driven methods. Mechanistic model-based approaches mainly build models based on mathematical and physical equations, which require a large amount of a priori knowledge and in-depth understanding, making it extremely difficult to establish an accurate mechanistic model. Data-driven methods include signal processing techniques, shallow machine learning methods, and deep learning methods. Signal processing methods mainly include wavelet transform (WT) [2,3], local mean decomposition (LMD) [4], empirical modal decomposition (EMD) [5], and short-time Fourier transform (STFT) [6,7]. The main shallow machine learning methods are support vector machine (SVM) [8,9], k-nearest neighbor (KNN) [10], and random forest (RF) [11]. Zheng et al. [12] implemented fault classification by calculating the composite multiscale fuzzy entropy (CMFE) of vibration signals based on ensemble support vector machine (ESVM). Huang et al. [13] analyzed a feature extraction method based on improved EMD energy entropy and used a genetic algorithm-based SVM as a classifier to classify faults. Yan et al. [14] proposed a new method to obtain feature vectors, instantaneous energy distribution-permutation entropy (IED-PE), by decomposing the vibration signal with improved VMD (IVMD) to obtain the components and then using KNN to achieve rolling bearing fault diagnosis.

However, traditional signal processing methods based on shallow machine learning struggle to meet the requirements of scalability, adaptability, and real-time performance in edge computing environments of the IIoT. These methods heavily rely on manual feature extraction, lack the ability to diagnose faults in complex high-dimensional signals, and are directly influenced by the expertise of personnel, making them unsuitable for different types of devices or operating conditions.

In recent years, deep learning models have gained increasing popularity in the field of fault diagnosis due to their efficient and reliable feature extraction capabilities. The convolutional neural network (CNN), as one of the most popular deep learning methods, can effectively extract deep features from data, making it highly applicable to fault diagnosis. Sun et al. [15] proposed a hybrid model based on CNN and long short-term memory (LSTM) networks, achieving higher fault diagnosis accuracy and robustness under different operating conditions. Zhang et al. [16] developed a CNN model incorporating an attention mechanism and dynamic learning rate for the intelligent diagnosis and identification of early faults in rolling bearings.

CNN has significant advantages over other deep learning methods for image processing and is specifically designed for processing 2D or 3D data. In some studies, raw data are converted into images to serve as input for CNNs. Ayas et al. [17] converted fault signals into grayscale images and used a deep residual network to learn the end-to-end mapping relationship between the images and the health condition of the moving bearings, enabling fault diagnosis. Xu et al. [18] transformed bearing vibration signals into time-frequency images using continuous wavelet transform and employed CNNs for fault feature extraction, which were then input into a classifier to achieve fault classification. Amarouayache et al. [19] utilized ensemble empirical modal decomposition (EEMD) to

reconstruct the original signal and then converted it into a two-dimensional image for input into a CNN for fault classification. He et al. [20] used Markov transform fields to convert bearing vibration signals into images, combined with deep residual networks to extract deep information from the images for fault diagnosis. Wen et al. [21] first converted bearing vibration signals into grayscale maps and then used a modified LeNet-5 model for fault diagnosis. Yuan et al. [22] proposed a method based on the Hilbert–Huang transform combined with CNN and successfully applied it to the fault diagnosis of rolling bearings. Additionally, many CNN-based architectures, such as VGG-net [23] and Res-net [24], have been proposed as image feature extractors and classifiers and have also been applied to the fault diagnosis of image-type inputs.

Although the above methods achieve the function of fault diagnosis, there are still the following problems: (1). When preprocessing the sensor signal, the reconstruction method is not perfect, and it is difficult to effectively retain the time-dependent relationship of the original signal, which is easy to cause the loss of information. (2). Although CNN networks have excellent local feature capture capability, they do not capture enough global features. Using CNN to extract global features requires multi-layer stacking, and the amount of information decays as the number of layers increases.

Based on the above analysis, this paper proposes a novel method for rolling bearing fault diagnosis that combines Gramian Angular Field (GAF), Convolutional Neural Network (CNN), and Vision Transformer (ViT). The main contributions of this paper are summarized as follows.

1. Using GAF to Convert Signals: The GAF converts one-dimensional vibration signals into two-dimensional images, preserving the temporal correlation of the original vibration signals and retaining sufficient fault features. This overcomes the limitations in representing fault features of complex signals;
2. Parallel Architecture of CNN and ViT: The proposed model employs a parallel architecture of CNN and ViT for fault diagnosis. CNN collects local features hierarchically by cascading convolution operations and retains local clues as feature maps. Meanwhile, ViT aggregates global representations by cascading self-attention modules. This effectively combines the local feature extraction capability of CNN with the global feature extraction capability of ViT.

The integration of these techniques ensures high accuracy and robustness of the model when handling complex high-dimensional signals, thereby enhancing the abstraction and representation capabilities for complex signals. Ultimately, our method significantly improves the accuracy of bearing fault diagnosis by effectively representing the features of the original vibration signals and their potential relationships, and this has been validated on popular bearing datasets.

The rest of this paper is organized as follows. Section 2 introduces the related basic theory and our bearing fault diagnosis method. Section 3 describes experiments and results validation. In the end, the main conclusions are given in Section 4.

## 2. Methods

### 2.1. Gramian Angular Field

The Gramian Angular Field is a method for encoding a time series as a two-dimensional image. In the conventional Cartesian coordinate system, the time series is a one-dimensional signal where the x-axis represents the sampling points in time order, and the y-axis represents the amplitude of the sampled signal. The encoding process is as follows:

First, the time domain signal  $X = \{x_1, x_2, x_3, \dots, x_n\}$  in the Cartesian coordinate system is normalized to be between  $-1$  and  $1$ . Here,  $x_i$  is the sampled points on the original time series, and  $\tilde{x}_i$  represents the sampled points after normalization. The normalization process is shown in Equation (1).

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

Next, the normalized values are transformed into polar coordinates by encoding the values  $\tilde{x}_i$  in  $\tilde{X}_i$  as angular cosines  $\phi$  and the timestamps as radii  $r$ . Thus, the time-domain signal  $X$  is transformed into a time series in polar coordinates, as shown in Equation (2).

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

where  $\tilde{X}_i$  is the normalized time series,  $t_i$  is the timestamp, and  $N$  is the constant factor of the generating space of the regularized polar coordinate system.

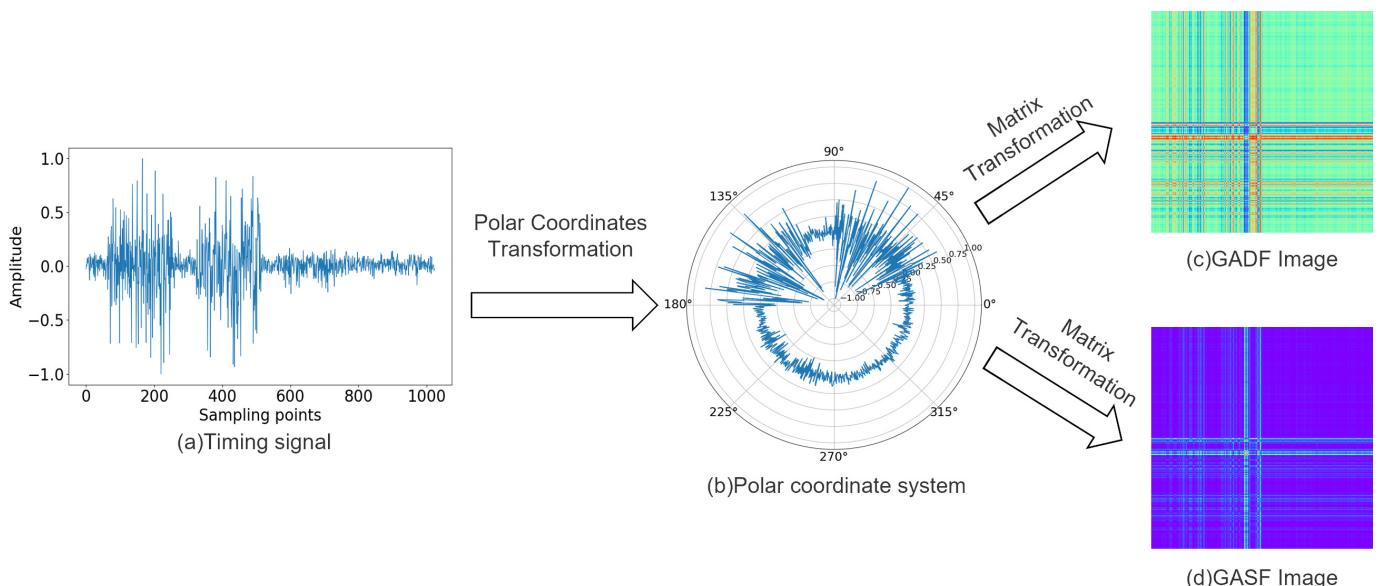
After converting the normalized time series into polar coordinates, the cosine value of the polar angle sum or polar angle difference between each time series point can be considered separately to identify the time correlation in different time intervals [25]. Therefore, it can be divided into the Gramian Angular Difference Field (GADF) based on the sine function and the Gramian Angular Sum Field (GASF) based on the cosine function, as defined by Equations (3) and (4).

$$GADF = \begin{bmatrix} \cos(\phi_1 - \phi_1) & \dots & \cos(\phi_1 - \phi_n) \\ \cos(\phi_2 - \phi_1) & \dots & \cos(\phi_2 - \phi_n) \\ \vdots & \cos(\phi_i - \phi_i) & \vdots \\ \cos(\phi_n - \phi_1) & \dots & \cos(\phi_n - \phi_n) \end{bmatrix} \quad (3)$$

$$GASF = \begin{bmatrix} \cos(\phi_1 + \phi_1) & \dots & \cos(\phi_1 + \phi_n) \\ \cos(\phi_2 + \phi_1) & \dots & \cos(\phi_2 + \phi_n) \\ \vdots & \cos(\phi_i + \phi_i) & \vdots \\ \cos(\phi_n + \phi_1) & \dots & \cos(\phi_n + \phi_n) \end{bmatrix} \quad (4)$$

After encoding, the temporal correlation of the 2D image from the top left to the bottom right is preserved, with the main diagonal containing the original information of the time series. Consequently, a time series of length  $n$  is encoded as a matrix of size  $n \times n$ .

The process of converting a one-dimensional signal into a two-dimensional image using GAF is illustrated in Figure 1. The GADF image exhibits richer colors compared to the GASF image, containing more detailed feature information.



**Figure 1.** GAF encoding process.

The advantages of using the GAF method to encode time signals are as follows:

1. The 1D signal and 2D image are bi-mapped, and the polar coordinate transformation encoding preserves both the values and temporal positions of the points without any information loss. Thus, the transformation process retains the complete information of the 1D signal;
2. The time order of the signal is represented from the upper left corner to the lower right corner of the matrix, ensuring that the converted matrix retains the time dependence of the original time signal.

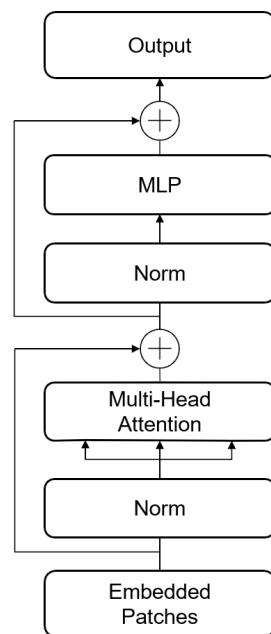
## 2.2. Vision Transformer

### 2.2.1. Patch Embedding

For an image with an input size of  $H \times W \times C$ , the patch embedding operation is first to partition it into feature blocks of size  $P \times P$ . The image is split into  $(H/P) \times (W/P)$  feature blocks, each with the same number of channels. Each feature block is then linearly mapped to a one-dimensional vector of length  $P \times P \times C$ . Finally, these one-dimensional vectors are combined to produce the output of the patch embedding process.

### 2.2.2. Transformer Encoder

The Transformer encoder, shown in Figure 2, includes a multi-headed self-attention (MSA) layer and a multi-layer perceptron (MLP) block. The computation process for the multi-headed self-attention layer used in this paper is as follows.



**Figure 2.** Transformer encoder.

$$Q = \text{Linear}(X) = XW^Q, Q_i = QW_i^Q \quad (5)$$

$$K = \text{Linear}(X) = XW^K, K_i = KW_i^K \quad (6)$$

$$V = \text{Linear}(X) = XW^V, V_i = VW_i^V \quad (7)$$

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax} \left( \frac{Q_i K_i^T}{\sqrt{d_{K_i}}} \right) \quad (8)$$

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) \times M + B \quad (9)$$

$$\text{MultiHead}(Q, K, V) =$$

$$\text{Concat}(\text{head}_1, \dots, \text{head}_h) \times W^o \quad (10)$$

$Q$ ,  $K$ , and  $V$  are three matrices generated by the linear mapping of input  $X$  with different weights assigned. Attention( $Q_i, K_i, V_i$ ) is the computed inter-sequence association matrix of feature blocks. Based on the multi-headed self-attention layer of Vision Transformer, this paper proposes defining a set of learnable parameters,  $M$  and  $B$ .  $M$  is used to adaptively adjust the scale of the association matrix, while  $B$  adjusts the offset of the association matrix to enhance the model's learning ability.

The MLP block comprises a fully connected layer, GELU activation function, and a Dropout layer with an inverted bottleneck structure. In this structure, the input feature layer is fully connected once, expanding the channel to four times its original size, and then the subsequent fully connected layer restores it to its original number. Layer normalization (LN) is applied before the residual connection in both the MSA and MLP blocks, which can be expressed by the following equation:

$$X_{\text{attention}} = X + \text{MSA}(\text{LN}(X)) \quad (11)$$

$$\tilde{X} = X_{\text{attention}} + \text{MLP}(\text{LN}(X_{\text{attention}})) \quad (12)$$

where  $X$  is the input sequence of feature blocks and  $\tilde{X}$  is the output result.

### 2.2.3. Position Embedding

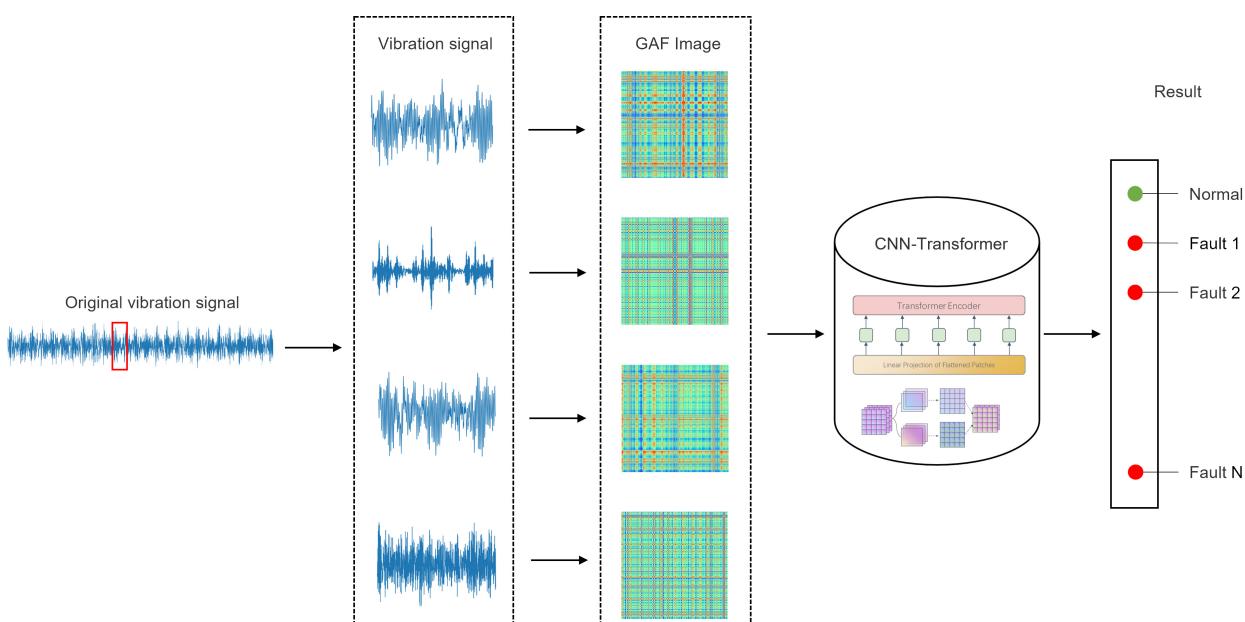
Position embedding involves integrating a set of learnable parameters that signify the positional relationships of individual feature block sequences. The embedding of the position encoding can be expressed as follows in Equation (13).

$$\tilde{x} = x + \theta \quad (13)$$

where  $x$  is the input feature block sequence and  $\theta$  is a set of parameters for the learnable position.

### 2.3. Proposed Method

Incorporating principles from GAF, CNN, and Vision Transformer theories, we present a novel fault diagnosis approach that integrates feature extraction and fault pattern recognition, facilitating accurate and efficient fault classification. The fault diagnosis procedure is illustrated in Figure 3, with detailed steps outlined below.



**Figure 3.** Fault diagnosis process based on GAF and CNN-ViT.

1. Intercept samples of the original vibration signal in the dataset according to the well-defined length;
2. Utilize GAF to encode one-dimensional vibration signal samples into two-dimensional GAF images, thereby generating the image dataset;
3. The image dataset are input to the CNN-ViT model for training and testing;
4. Obtain the fault classification results.

### 2.3.1. Signal-to-Image Conversion

The original vibration signal is initially captured as a series of samples. Subsequently, the one-dimensional vibration signal undergoes transformation into GADF and GASF matrices using the GAF method. These matrices serve as the basis for generating GADF and GASF images, wherein the elements within the matrices represent individual pixels. These images are then resized to a resolution of  $224 \times 224$  with a three-channel configuration, resulting in an image size of  $224 \times 224 \times 3$ . This resized image format is employed as the input for the CNN-ViT model.

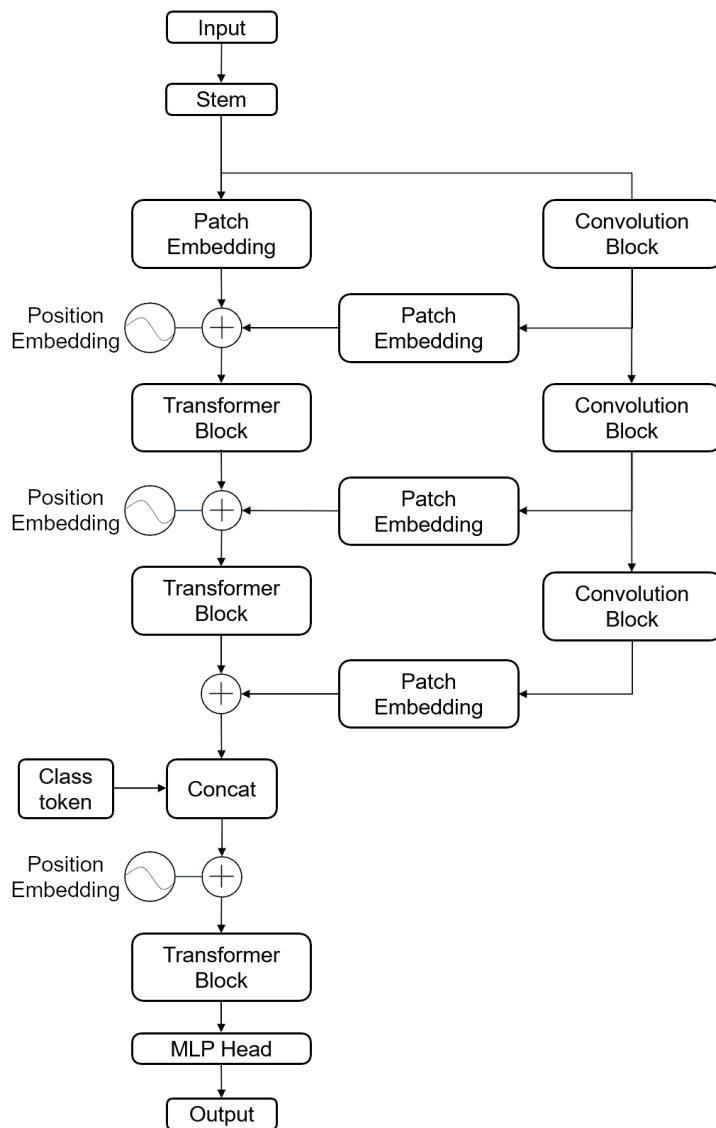
### 2.3.2. Construction of the Model

For GAF images, both global and local features are crucial. Traditional CNN models excel at extracting local information through convolution operations, capturing features at various levels, and producing feature maps via stacked convolutional layers. Conversely, ViT leverages self-attention modules to extract global image features starting from shallow layers [26]. To harness both global and local features effectively from GAF images, this paper introduces a model with a parallel architecture comprising both CNN and ViT components.

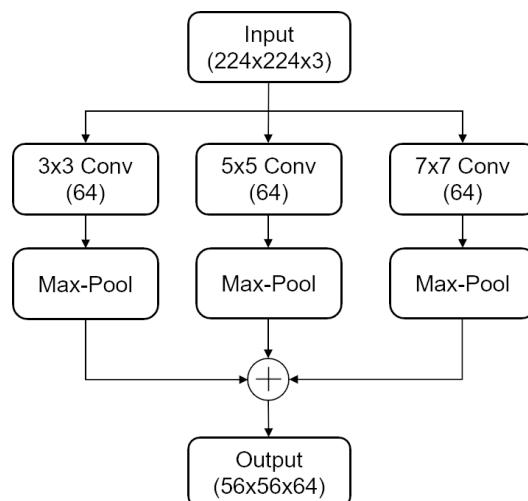
The architecture of the proposed CNN-ViT model, as shown in Figure 4, consists of three main parts: the stem block, the CNN branch, and the ViT branch. First, the GAF image is input into the stem block for initial feature extraction. This block captures multi-scale local features using three parallel paths with convolution and pooling operations of different kernel sizes ( $3 \times 3$ ,  $5 \times 5$ , and  $7 \times 7$ ). The initially processed feature maps are then directed to both the CNN branch and the ViT branch. The CNN branch and the ViT branch each consist of multiple convolutional modules and Transformer modules. This parallel structure allows the model to maximally retain both local features and global representations. The Patch Embedding acts as a bridging module, linearly mapping the complete feature map into a sequence of feature patches and gradually transmitting the local feature map to the Transformer branch. This integration enables the local representation feature map from the CNN branch to merge with the global feature representation map from the Transformer branch. Additionally, positional encoding is added to understand the global context and long-range dependencies. Finally, the output results are obtained through a fully connected layer.

The CNN-ViT model extracts the fault features of GAF images in the following process.

1. The stem block is used to extract initial local features, such as edge and texture information, from the GAF image using a three-branch parallel structure. In the first branch, a convolution operation of size 3 with a stride of 2 is applied, followed by a max pooling layer with a size of 3 and a stride of 2. The second branch involves a convolution operation of size 5 with a stride of 2, followed by a max pooling layer with a size of 3 and a stride of 2. Similarly, the third branch employs a convolution operation of size 7 with a stride of 2, followed by a max pooling layer with a size of 3 and a stride of 2. Local features at different scales of the original input image are extracted from each branch, and the resulting feature maps are aggregated by summation to serve as input for the CNN and ViT branches. For an input image of size  $224 \times 224 \times 3$ , the output after the stem module is a  $56 \times 56 \times 64$  feature map. The structure of the stem block is depicted in Figure 5.



**Figure 4.** Structure of CNN-ViT model.



**Figure 5.** Stem block.

2. In order to reduce the parameters and decrease the amount of operations, patch embedding selects a feature block size of  $4 \times 4$ . The first patch embedding in the Transformer branch of the model divides the  $56 \times 56 \times 64$  feature map into 196 feature blocks of size  $4 \times 4 \times 64$ , which are then transformed into an embedding of size  $196 \times 1024$ . The remaining three patch embedding transform the  $56 \times 56 \times 64$  feature maps output from the three convolutional blocks into embeddings of size  $196 \times 1024$ , which are dimensionally consistent with the sequence of feature blocks on the Transformer branch, so they can be directly passed to the Transformer branch for fusion in a summation manner.

3. The CNN branch is composed of three consecutive convolutional blocks, each applying convolutions with a kernel size of 3 and a stride of 1. This configuration preserves the spatial dimensions between the input and output feature maps. As the network progresses through these layers—gradually expanding its receptive field—it extracts local features across diverse scales [27], prior to forwarding them to the ViT branch. This integration strategy counterbalances the potential lack of local feature extraction within the ViT branch, which directly transforms the feature map into a sequence of feature blocks via the Patch Embedding operation.

4. We add position encoding to the sequence of feature blocks. Following position encoding, the size, number, and the number of channels of the sequence of feature blocks remain unchanged.

5. After position encoding, the embedded data are fed into a Transformer block, which consists of multiple Transformer encoders stacked together. Specifically, there are 2, 2, and 6 of these encoders in each of the successive blocks, respectively. Prior to the final Transformer block, a category symbol is generated to represent the input feature block sequence, and this symbol is appended to the beginning of the original feature sequence. An MLP Head module follows, connecting a fully connected layer to this category symbol vector, with the ultimate aim of outputting the fault category.

6. The combination of local and global features is achieved through branch fusion. Each convolutional block in the CNN branch extracts local features from the image without changing the feature map size. After patch embedding, the dimensions are aligned with the global features in the ViT branch. Before being input to the subsequent processing layers, the feature vectors are summed, thereby aggregating the feature information of the entire image.

### 3. Experimental Validation

In this paper, we use CWRU and XJTU-SY bearing dataset to validate the effectiveness and practicality of the GAF combined with CNN-ViT fault diagnosis method. The experimental platform is Python3.7 + Pytorch1.7.1, and the hardware is Intel Xeon(R) Gold 6226R + NVIDIA GeForce RTX 3090.

#### 3.1. Data Description

The Case Western Reserve University (CWRU) dataset was provided by the Case Western Reserve University Bearing Data Center [28]. Vibration signals from the experimental platform, obtained from both normal and defect-bearing conditions, were recorded at frequencies of 12 kHz or 48 kHz under four distinct motor load scenarios. Under each operational condition, faults characterized by diameters of 0.007, 0.014, and 0.021 inches were artificially induced in the rolling elements, inner rings, and outer rings, respectively. This study employs data sampled at 12 kHz from the drive end, categorizing them into one healthy state and three failure modes: inner ring failure, rolling element failure, and outer ring failure. Further, these conditions are stratified into ten classes based on the extent of failure: comprising one healthy state and nine distinct failure states. Table 1 outlines the fault classifications alongside their respective labels for the CWRU dataset.

**Table 1.** Description of CWRU sample categories.

Fault Location	Fault Diameter (Inch)	Label
NO	0	0
Inner ring	0.007	1
Inner ring	0.014	2
Inner ring	0.021	3
Rolling Element	0.007	4
Rolling Element	0.014	5
Rolling Element	0.021	6
Outer ring	0.007	7
Outer ring	0.014	8
Outer ring	0.021	9

XJTU-SY bearing dataset were provided by the Institute of Design Science and Basic Component at Xi'an Jiaotong University and the Changxing Sumyoung Technology Co., Ltd. (Huzhou, China) [29,30]. In this paper, experiments were conducted using Bearing1 from the dataset with a sampling frequency of 2.56 kHz. Five categories were classified according to the actual lifetime as well as the failure location, and each failure state was uniquely thermally coded. Table 2 describes the failure categories and corresponding labels of XJTU-SY dataset.

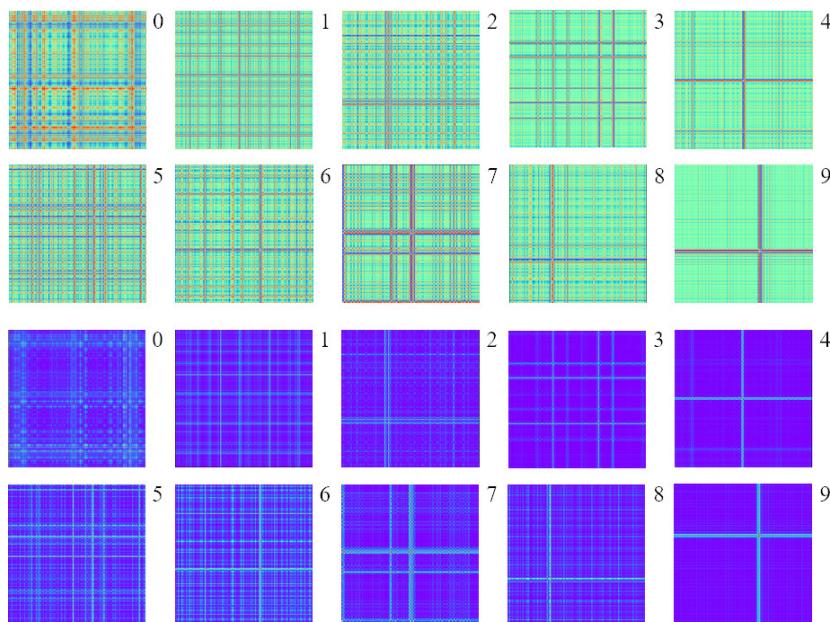
**Table 2.** Description of XJTU-SY sample categories.

File	Lifetime	Fault Element	Label
Bearing 1_1	2 h 3 min	Outer ring	0
Bearing 1_2	2 h 41 min	Outer ring	1
Bearing 1_3	2 h 38 min	Outer ring	2
Bearing 1_4	2 h 2 min	Cage	3
Bearing 1_5	52 min	Inner ring and Outer ring	4

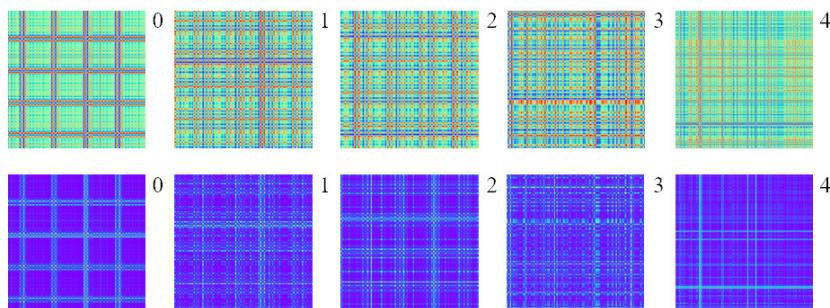
### 3.2. Data Processing

Given the limited length of the bearing vibration signal for each fault condition, we employ sliding window sampling to increase the sample size. The sliding window progresses with a step equivalent to its movement distance, set at a window length of 1024 samples and a step size of 512 samples, thereby sampling at a 50% overlap rate. This approach doubles the sample count while mitigating feature loss that would otherwise result from direct truncation of samples [31].

Each category of failure state signals of CWRU bearing data is divided into 2400 samples by sliding window sampling, i.e., 240 samples for each category. Then, the vibration signal samples are converted into 2400 images by GAF coding as the data set, and the GAF images of ten fault states are shown in Figure 6. Similarly, each type of fault state signal of XJTU-SY bearing data is divided into 4000 samples by sliding window sampling, i.e., 800 samples for each category, and then, the vibration signal samples are converted into 4000 images by GAF coding as the data set, and the GAF images of five fault states are shown in Figure 7.



**Figure 6.** Ten GAF images based on CWRU bearing dataset. The green ones are GADF images and the purple ones are GASF images.



**Figure 7.** Five GAF images based on XJTU-SY bearing dataset. The green ones are GADF images and the purple ones are GASF images.

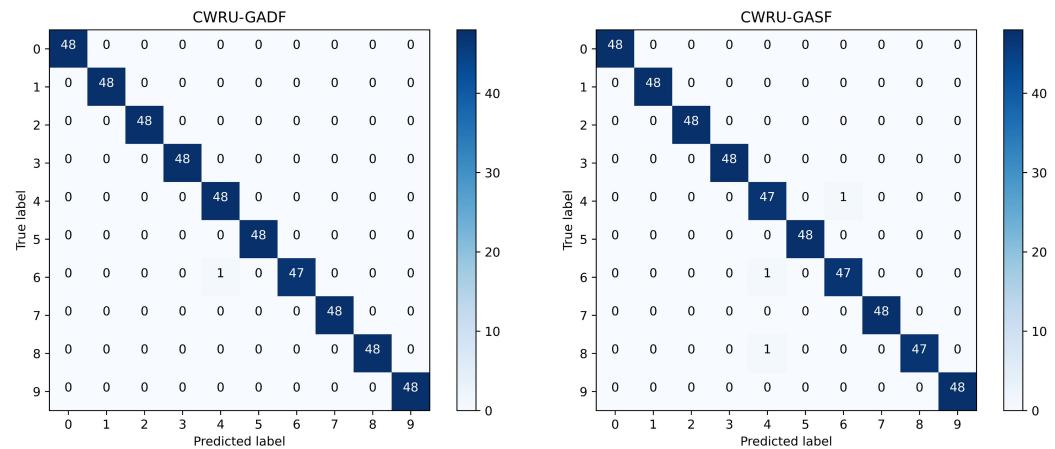
### 3.3. Experimental Results

In order to avoid classification bias caused by specific data splits, 80% of the dataset was randomly selected as the training set and the remaining 20% as the test set, for a total of 10 training rounds. The model parameters are as follows: learning rate is 0.0001, batch size is 64, number of epochs is 200, optimizer is Adam, and the loss function is cross-entropy.

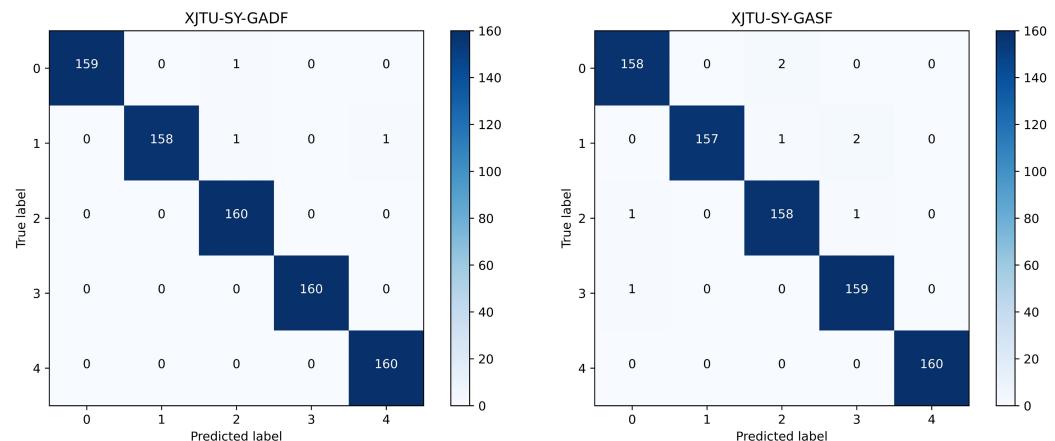
To better show the experimental results of the model for different fault types, a confusion matrix was introduced to visualize the prediction results, where the horizontal axis is the prediction label, and the vertical axis is the true label. The confusion matrix of the classification results for ten bearing states of the CWRU dataset is shown in Figure 8, and the confusion matrix of the classification results for five bearing states of the XJTU-SY dataset is shown in Figure 9.

We conducted several experiments, and the average recognition rates of GADF and GASF images in the test set were 99.79% and 99.38%, respectively, in the CWRU dataset, and 99.63% and 99.00%, respectively, in the XJTU-SY dataset. In the experiments conducted using the CWRU dataset, we found that the method proposed in this paper accurately and clearly predicts faults or non-faults, with very few fault-free images being predicted as faulty. Additionally, nearly all misclassifications occurred in ball faults, with undetected inner or outer ring faults possibly caused by Brinell hardness. In the experiments using the XJTU-SY dataset, we found that the proposed method generally accurately identifies the

specific location of rolling bearing faults, but it tends to misclassify different fault states at the same location.

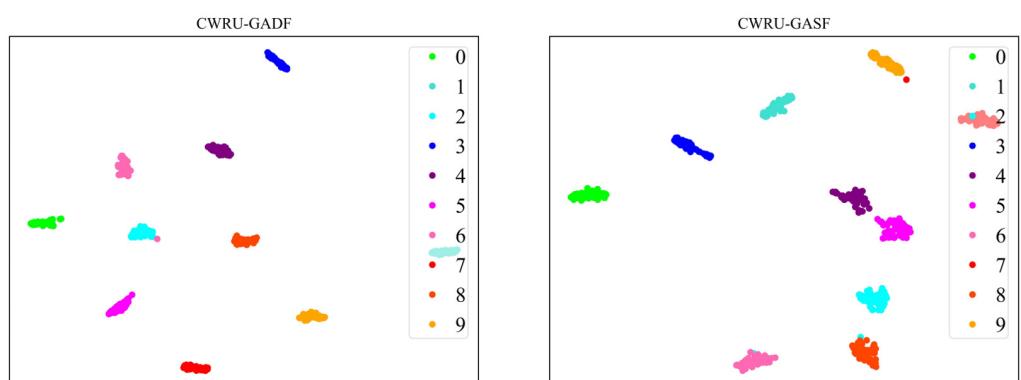


**Figure 8.** Confusion matrix of CWRU classification results.

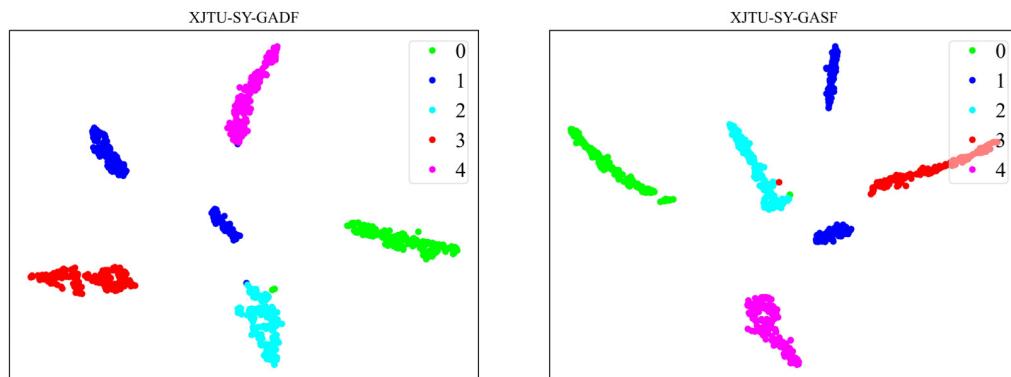


**Figure 9.** Confusion matrix of XJTU-SY classification results.

To better illustrate the performance of the model, we employ the nonlinear dimensionality reduction algorithm, t-SNE [32], to project and visualize the data in a two-dimensional space. Specifically, for the CWRU dataset, Figure 10 showcases the t-SNE visualization results of the CNN-ViT model when extracting fault features from both GADF and GASF images. Meanwhile, the t-SNE visualization outcomes for the XJTU-SY dataset, illustrating the CNN-ViT model's feature extraction from GADF and GASF images, are depicted in Figure 11.

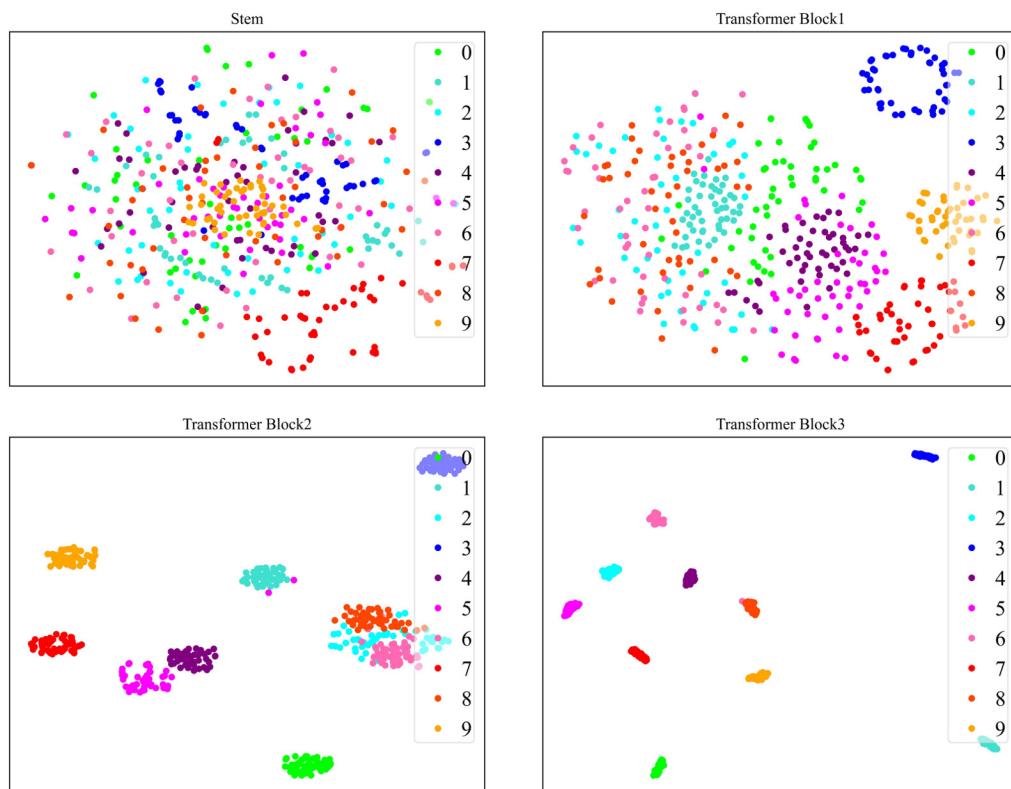


**Figure 10.** t-SNE visualization of GAF images fault feature extraction results for CWRU dataset.



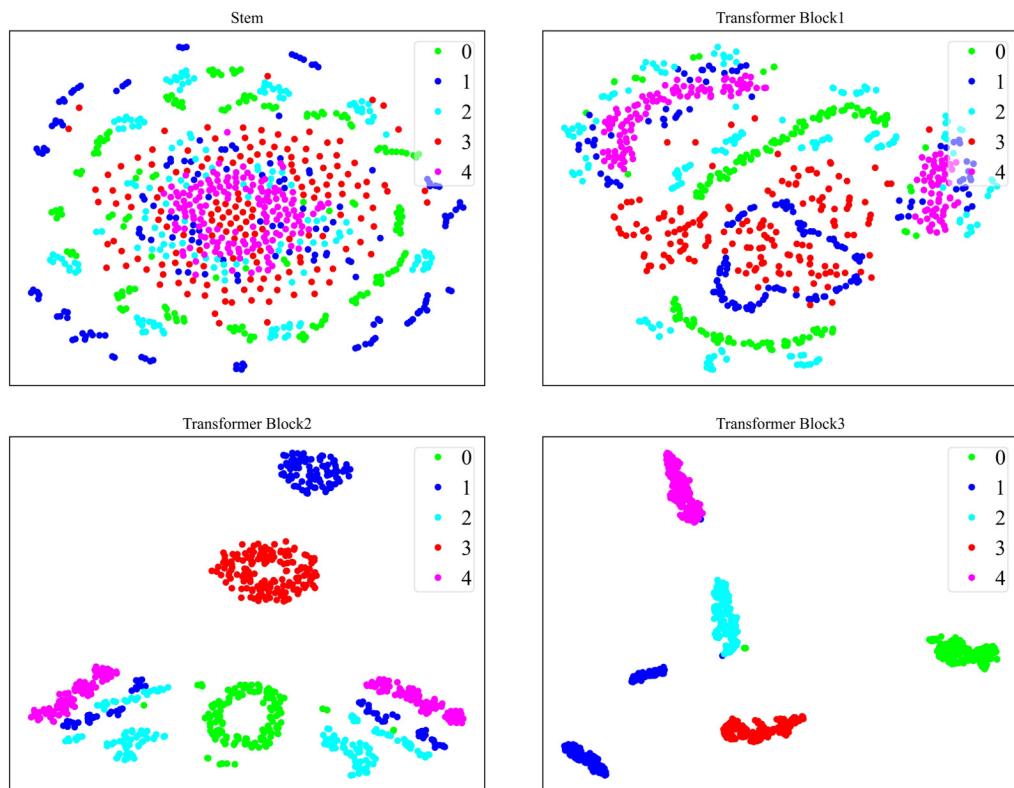
**Figure 11.** t-SNE visualization of GAF images fault feature extraction results for XJTU-SY dataset.

In order to understand the process of feature extraction by the model more clearly, we visualize the outputs from the Stem module and the three Transformer Blocks using t-SNE. Taking the GADF images from both the CWRU and XJTU-SY datasets as examples, Figures 12 and 13, respectively, illustrate this process.



**Figure 12.** t-SNE visualization of feature extraction process for GADF images of CWRU dataset.

Initially, the Stem module extracts initial local features from the raw GADF image; however, at this stage, distinguishing fault samples belonging to different categories is difficult. As the process progresses through the CNN branch and ViT branch, which incrementally extract deeper image features, the sample distribution evolves, with faults of the same category beginning to cluster together. Ultimately, following the comprehensive feature extraction by the CNN-ViT model, a superior separation of sample features across different fault categories is achieved, highlighting the effectiveness of the model's feature extraction capabilities.

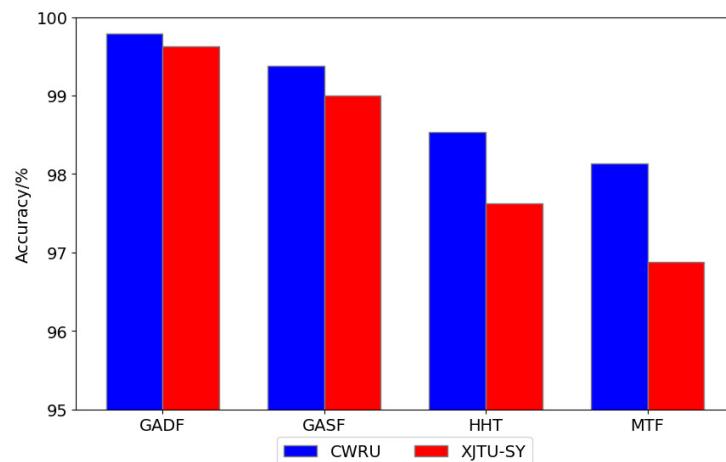


**Figure 13.** t-SNE visualization of feature extraction process for GADF images of XJTU-SY dataset.

From the above experiments, it can be seen that the CNN-ViT architecture proposed in this paper achieves better fault diagnosis results on public datasets by introducing convolution operators from CNNs to extract local features of images and self-attention mechanisms from Transformers to capture global representations. To further verify the reliability of the image encoding method proposed in this paper, we compared it with other image encoding methods. One method converts the original one-dimensional signal into a time-frequency map based on the Hilbert–Huang Transform (HHT), and another method encodes the original one-dimensional signal into feature images based on the Markov Transition Field (MTF). The training and testing sets were divided in the same proportion. The training and testing sets were then input into the CNN-ViT model for feature extraction and fault classification. The results, as shown in Figure 14, indicate that the recognition rates of GAF images are higher than those of HHT images and MTF images, with the highest recognition rate achieved by GADF images. This demonstrates that the GADF image encoding method has better signal reconstruction ability and can more effectively express vibration signal characteristics, making it more suitable for bearing fault classification and diagnosis compared to other image encoding methods.

To verify the advantages of the CNN-ViT model in bearing fault diagnosis, comparative experiments were conducted with various fault classification models, and the results are presented in Tables 3 and 4. The LSTM and SVM algorithms, which rely on one-dimensional vibration signals as input, struggle to meet the rigorous demands of high-precision fault diagnosis, with their diagnostic accuracy generally trailing behind that of deep learning models employing GAF-encoded images of these one-dimensional signals. This evidences that encoding vibration signals into GAF images markedly enhances the overall recognition accuracy of the model. Furthermore, in order to comprehensively evaluate the performance, the study incorporated the VGG16 model, embodying a classic CNN architecture, and the ViT model for comparative experiments. Whether tested on the CWRU dataset or the XJTU-SY dataset, the CNN-ViT model proposed here achieved slightly higher classification accuracy compared to instances where the CNN and ViT architectures were used independently under the same input conditions. These results demonstrate that the

proposed CNN-ViT model has strong feature learning and fault diagnosis capabilities, allowing it to accurately identify bearing status.



**Figure 14.** Fault diagnosis test results of different image coding methods.

**Table 3.** Experimental results of CWRU dataset.

Method	Data Type	Accuracy
CNN	GADF	99.58%
CNN	GASF	98.96%
Vision Transformer	GADF	99.17%
Vision Transformer	GASF	98.54%
CNN-ViT	GADF	99.79%
CNN-ViT	GASF	99.38%
LSTM	Original signal	95.83%
SVM	Original signal	96.45%

**Table 4.** Experimental results of XJTU-SY dataset.

Method	Data Type	Accuracy
CNN	GADF	99.13%
CNN	GASF	98.50%
Vision Transformer	GADF	99.00%
Vision Transformer	GASF	98.25%
CNN-ViT	GADF	99.63%
CNN-ViT	GASF	99.00%
LSTM	Original signal	94.75%
SVM	Original signal	95.50%

#### 4. Conclusions

In this paper, a fault diagnosis method that can be used for edge devices is proposed. The one-dimensional vibration signal is encoded into a two-dimensional image by GAF, the original signal from the sensor is reconstructed, and the time correlation is retained. A CNN-ViT model for rolling bearing condition identification was constructed. Experimental results on CWRU and XJTU-SY bearing datasets show that the proposed fault diagnosis method achieves average fault recognition accuracies of 99.79% and 99.63%, respectively. Compared to other intelligent bearing fault diagnosis methods, it exhibits better feature extraction and classification performance, providing a reliable solution for fault diagnosis on edge devices. However, this study has certain limitations. Firstly, the evaluation of the model is focused on the CWRU and XJTU-SY datasets, which implies a dependence on specific datasets. Therefore, future research should consider validation on a more diverse

range of industrial datasets to ensure the general applicability of the research findings. The fault diagnosis method proposed in this paper involves many hyperparameters, most of which are set based on experimental experience, posing the risk of overfitting to the training data and compromising the model's generalization to unseen instances. In future work, we will explore intelligent hyperparameter tuning algorithms, including Bayesian optimization or neural architecture search, to systematically explore the vast hyperparameter space. This will not only mitigate the potential risk of overfitting but also accelerate the model development cycle, making it more suitable for deployment on resource-constrained edge devices.

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

**Funding:** This research was funded by National Key R&D Program of China (2022ZD0119303) and The Fundamental Research Funds for the Central Universities (WUT: 104972024KFYd0003).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are publicly available.

**Acknowledgments:** We would like to thank all the workmates who participated in the work in the School of Information Engineering, Wuhan University of Technology.

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

## References

- Li, X.; Zhang, W. Deep learning-based partial domain adaptation method on intelligent machinery fault diagnostics. *IEEE Trans. Ind. Electron.* **2020**, *68*, 4351–4361. [[CrossRef](#)]
- Chen, J.; Li, Z.; Pan, J.; Chen, G.; Zi, Y.; Yuan, J.; Chen, B.; He, Z. Wavelet transform based on inner product in fault diagnosis of rotating machinery: A review. *Mech. Syst. Signal Process.* **2016**, *70*, 1–35. [[CrossRef](#)]
- Kompella, K.D.; Mannam, V.G.R.; Rayapudi, S.R. DWT based bearing fault detection in induction motor using noise cancellation. *J. Electr. Syst. Inf. Technol.* **2016**, *3*, 411–427. [[CrossRef](#)]
- Li, X.; Ma, J.; Wang, X.; Wu, J.; Li, Z. An improved local mean decomposition method based on improved composite interpolation envelope and its application in bearing fault feature extraction. *ISA Trans.* **2020**, *97*, 365–383. [[CrossRef](#)]
- Guo, T.; Deng, Z. An improved EMD method based on the multi-objective optimization and its application to fault feature extraction of rolling bearing. *Appl. Acoust.* **2017**, *127*, 46–62. [[CrossRef](#)]
- Wang, L.H.; Zhao, X.P.; Wu, J.X.; Xie, Y.Y.; Zhang, Y.H. Motor fault diagnosis based on short-time Fourier transform and convolutional neural network. *Chin. J. Mech. Eng.* **2017**, *30*, 1357–1368. [[CrossRef](#)]
- Liu, H.; Li, L.; Ma, J. Rolling bearing fault diagnosis based on STFT-deep learning and sound signals. *Shock Vib.* **2016**, *2016*, 6127479. [[CrossRef](#)]
- Mohiuddin, M.; Islam, M.S.; Islam, S.; Miah, M.S.; Niu, M.B. Intelligent Fault Diagnosis of Rolling Element Bearings Based on Modified AlexNet. *Sensors* **2023**, *23*, 7764. [[CrossRef](#)]
- Shen, W.; Xiao, M.; Wang, Z.; Song, X. Rolling bearing fault diagnosis based on support vector machine optimized by improved grey wolf algorithm. *Sensors* **2023**, *23*, 6645. [[CrossRef](#)]
- Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Wang, R. Efficient kNN classification with different numbers of nearest neighbors. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 1774–1785. [[CrossRef](#)]
- Quiroz, J.C.; Mariun, N.; Mehrjou, M.R.; Izadi, M.; Misron, N.; Radzi, M.A.M. Fault detection of broken rotor bar in LS-PMSM using random forests. *Measurement* **2018**, *116*, 273–280. [[CrossRef](#)]
- Zheng, J.; Pan, H.; Cheng, J. Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines. *Mech. Syst. Signal Process.* **2017**, *85*, 746–759. [[CrossRef](#)]
- Huang, J.; Hu, X.; Yang, F. Support vector machine with genetic algorithm for machinery fault diagnosis of high voltage circuit breaker. *Measurement* **2011**, *44*, 1018–1027. [[CrossRef](#)]
- Yan, X.; Jia, M.; Zhao, Z. A novel intelligent detection method for rolling bearing based on IVMD and instantaneous energy distribution-permutation entropy. *Measurement* **2018**, *130*, 435–447. [[CrossRef](#)]
- Sun, H.; Fan, Y. Fault diagnosis of rolling bearings based on CNN and LSTM networks under mixed load and noise. *Multimed. Tools Appl.* **2023**, *82*, 43543–43567. [[CrossRef](#)]

16. Zhang, Q.; Wu, H.; Tao, J.; Ding, W.; Zhang, J.; Li, J. Fault Diagnosis of Rolling Bearing Based on CNN with Attention Mechanism and Dynamic Learning Rate. In Proceedings of the 2021 International Conference on Sensing, Measurement & Data Analytics in the Era of Artificial Intelligence (ICSMD), Nanjing, China, 21–23 October 2021; pp. 1–7.
17. Ayas, S.; Ayas, M.S. A novel bearing fault diagnosis method using deep residual learning network. *Multimed. Tools Appl.* **2022**, *81*, 22407–22423. [[CrossRef](#)]
18. Xu, Y.; Li, Z.; Wang, S.; Li, W.; Sarkodie-Gyan, T.; Feng, S. A hybrid deep-learning model for fault diagnosis of rolling bearings. *Measurement* **2021**, *169*, 108502. [[CrossRef](#)]
19. Amarouayache, I.I.E.; Saadi, M.N.; Guersi, N.; Boutasseta, N. Bearing fault diagnostics using EEMD processing and convolutional neural network methods. *Int. J. Adv. Manuf. Technol.* **2020**, *107*, 4077–4095. [[CrossRef](#)]
20. He, K.; Xu, Y.; Wang, Y.; Wang, J.; Xie, T. Intelligent diagnosis of rolling bearings fault based on multisignal fusion and MTF-ResNet. *Sensors* **2023**, *23*, 6281. [[CrossRef](#)]
21. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Trans. Ind. Electron.* **2017**, *65*, 5990–5998. [[CrossRef](#)]
22. Yuan, Z.; Zhang, L.; Duan, L.; Li, T. Intelligent fault diagnosis of rolling element bearings based on HHT and CNN. In Proceedings of the 2018 Prognostics and System Health Management Conference (PHM-Chongqing), Chongqing, China, 26–28 October 2018; pp. 292–296.
23. Simonyan, K.; Zisserman, A. Deep convolutional networks for large-scale image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
25. Wang, Z.; Oates, T. Imaging time-series to improve classification and imputation. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
26. Fan, X.; Feng, X.; Dong, Y.; Hou, H. COVID-19 CT image recognition algorithm based on transformer and CNN. *Displays* **2022**, *72*, 102150. [[CrossRef](#)]
27. Yang, X.; Huang, X.; Gao, X.; Tao, Z. High-temperature power reduction state identification for wind turbines using feature correlation analysis and deep learning methods. *Forsch. Im Ingenieurwesen* **2022**, *86*, 225–239. [[CrossRef](#)]
28. CWRU Dataset; Case Western Reserve University Bearing Data Center. Available online: <https://csegroups.case.edu/bearingdatacenter> (accessed on 21 May 2024).
29. Lei, Y.; Han, T.; Wang, B.; Li, N.; Yan, T.; Yang, J. XJTU-SY rolling element bearing accelerated life test datasets: A tutorial. *J. Mech. Eng.* **2019**, *55*, 1–6.
30. Wang, B.; Lei, Y.; Li, N.; Li, N. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Trans. Reliab.* **2018**, *69*, 401–412. [[CrossRef](#)]
31. Yan, J.; Kan, J.; Luo, H. Rolling Bearing Fault Diagnosis Based on Markov Transition Field and Residual Network. *Sensors* **2022**, *22*, 3936. [[CrossRef](#)]
32. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

**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.