electronics

MDPI

Article

Rolling Bearing Fault Diagnosis Based on SVD-GST Combined with Vision Transformer

Fengyun Xie [1,2,3,*], Gan Wang [1], Haiyan Zhu [1,2,3], Enguang Sun [1], Qiuyang Fan [1] and Yang Wang [1]

[1] School of Mechanical Electrical and Vehicle Engineering, East China Jiaotong University, Nanchang 330013, China; wanggan813@163.com (G.W.); zhupetrelcao@163.com (H.Z.); sngzm999@163.com (E.S.); fqy18532671715@163.com (Q.F.); wwyy0130@163.com (Y.W.)
[2] State Key Laboratory of Performance Monitoring Protecting of Rail Transit Infrastructure, East China Jiaotong University, Nanchang 330013, China
[3] Life-Cycle Technology Innovation Center of Intelligent Transportation Equipment, Nanchang 330013, China
* Correspondence: xiefyun@163.com

Abstract: Aiming at rolling bearing fault diagnosis, the collected vibration signal contains complex noise interference, and one-dimensional information cannot be used to fully mine the data features of the problem. This paper proposes a rolling bearing fault diagnosis method based on SVD-GST combined with the Vision Transformer. Firstly, the one-dimensional vibration signal is preprocessed to reduce noise using singular value decomposition (SVD) to obtain a more accurate and useful signal. Then, the generalized S-transform (GST) is used to convert the processed one-dimensional vibration signal into a two-dimensional time–frequency image and make full use of the advantages of deep learning in image classification with higher recognition accuracy. In order to avoid the problem of limited sensory fields in CNN and the need for an RNN to compute step by step over time when processing sequence data, the use of a Vision Transformer model for pattern recognition classification is proposed. Finally, an experimental platform for the fault diagnosis of rolling bearings is built. The model is experimentally validated, achieving an average accuracy of 98.52% over multiple tests. Additionally, compared with the SVD-GST-2DCNN, STFT-CNN-LSTM, SVD-GST-LSTM, and GST-ViT fault diagnosis models, the proposed method has higher diagnostic accuracy and stability, providing a new method for rolling bearing fault diagnosis.

Keywords: singular value decomposition; generalized S-transform; Vision Transformer; rolling bearing; fault diagnosis

check for updates

Citation: Xie, F.; Wang, G.; Zhu, H.; Sun, E.; Fan, Q.; Wang, Y. Rolling Bearing Fault Diagnosis Based on SVD-GST Combined with Vision Transformer. *Electronics* **2023**, *12*, 3515. https://doi.org/10.3390/electronics12163515

Academic Editors: Séjir Khojet El Khil, Chiara Boccaletti and Monia Ben Khader Bouzid

Received: 9 August 2023
Revised: 15 August 2023
Accepted: 18 August 2023
Published: 19 August 2023

Copyright: © 2023 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

Rolling bearings are very common in rotating machine parts and are widely used in all kinds of mechanical equipment, including high-speed railways, airplanes, and automobiles [1]. Using fault diagnosis technology, the health status of rolling bearings can be judged more accurately, which can save on maintenance costs of the mechanical equipment and avoid unnecessary waste [2]. If rolling bearing failure occurs, it causes property loss and even a threat to the staff's life and health. Therefore, the fault diagnosis of rolling bearings is of great significance [3].

After years of development, rolling bearing fault diagnosis has gradually moved from traditional methods to intelligent fault diagnosis. Traditional rolling bearing fault diagnosis methods are usually more dependent on the practitioner's professional knowledge and work experience, so they can be influenced by the practitioner's own subjective judgment [4]. Conventional methods also require regular testing and maintenance of equipment by staff, which, for a large organization in continuous operation, requires a large annual investment to support the work [5]. Although traditional rolling bearing fault diagnosis technology has played an important role, with the continuous development and progress of science and technology, more intelligent fault diagnosis methods have emerged, one

after another. These methods meet the current needs of people because they improve the reliability and accuracy of fault diagnosis [6].

Intelligent fault diagnosis techniques utilizing artificial intelligence technology have gradually emerged over the years. The acquired data are analyzed using various equipment sensors. Using machine learning, deep learning, and other technologies, we can attain feature extraction and pattern recognition classification from these data [7] to determine whether there is a potential failure or an abnormal situation occurring in industrial equipment, as well as to determine which type of failure it is [8].

The use of deep learning techniques in the field of fault diagnosis is in line with the current trend of the positive impact of computers on people's lives. Compared to previous fault diagnosis methods, deep learning–based methods can realize an "end-to-end" fault diagnosis process, avoiding the troublesome feature extraction process [9]. In deep learning technology, neural network models are currently very popular, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory networks (LSTMs) [10]. These network models have also been widely used in the field of fault diagnosis. Liu et al. [11] proposed the possibility of combining GST and SVD to extract localized damage in rolling bearings and compared it with other commonly used methods to verify the feasibility of the model. Pan et al. [12] proposed an improved bearing fault diagnosis method combining a CNN and LSTM, where the input is the raw sampled signal without any preprocessing or conventional feature extraction. Huang et al. [13] used EMD for processing bearing vibration signals to reduce noise and then constructed a convolutional recurrent neural network as a rolling bearing fault diagnosis classifier using the envelope processed via EMD. Lu et al. [14] proposed a rolling bearing fault diagnosis method combining LSTM and a self-encoder. Self-encoders can automatically learn useful functions from vibration signals. LSTM is used to process time series data, and the LSTM network is used as an encoder and decoder of the self-encoder. Additionally, it has been shown experimentally that the proposed algorithm has good multi-class classification performance. In recent years, the Transformer model has shined in the field of deep learning. The Transformer model is a classic NLP model because of its excellent sequence modeling ability and the advantages of parallel computing [15]. Its biggest feature is the use of an attention mechanism to calculate its input and output, as well as balanced processing power. It does not adopt the sequential structure of traditional RNN sequence alignment and also avoids the drawbacks of the limited receptive field of CNNs, enabling it to capture global information [16].

Although the current research on deep learning fault diagnosis methods has achieved initial results, most of the existing research is aimed at one-dimensional vibration signal data as input. However, the research on the deep learning fault diagnosis method that takes two-dimensional data as input is not deep enough. There are few studies on the analysis of two-dimensional data in more complex situations [17]. Therefore, the potential advantages brought by the bearing fault diagnosis method that converts the vibration signal into two-dimensional image data and then uses the two-dimensional image data as the input of the deep learning model are worth exploring in depth. Visualizing the vibration signal can not only retain the information contained in the vibration signal with high quality but also optimize the preprocessing of the vibration signal data [18]. At the same time, the deep learning model has good recognition and processing characteristics for the converted two-dimensional image data, and the diagnosis accuracy rate is high.

To sum up, in order to effectively improve the problems of incomplete feature extraction, too-complex feature extraction, and external noise interference in the fault diagnosis process of rolling bearings, and to avoid the problems of limited receptive fields, CNNs and RNNs need to gradually calculate using time sequences when processing sequence data. This article proposes a rolling bearing fault diagnosis method that combines two-dimensional vibration images with Transformer models. The main contributions of this paper are as follows:

(1) A rolling bearing fault diagnosis method based on SVD-GST combined with the Vision Transformer is proposed. A fault diagnosis experimental platform is built, and the model is verified to have high accuracy and feasibility through experiments.
(2) In the process of using SVD noise reduction, the singular value energy difference spectrum is introduced to determine the order, which solves the problem of how to determine the effective order of the reconstruction matrix after the vibration signal of the rolling bearing is decomposed.
(3) It is verified that the Vision Transformer model can mine more hidden fault information and reduce information loss for the two-dimensional vibration images of rolling bearings obtained using GST.

The rest of this paper is composed as follows: Section 2 introduces the algorithm principle, including SVD, GST, and the Vision Transformer; Section 3 introduces the rolling bearing fault diagnosis model; Section 4 builds a fault diagnosis experimental platform and introduces the process of vibration signal acquisition; Section 5 analyzes the experimental results in various ways as well as compares them with other models; and Section 6 is the conclusion of this paper.

2. Principle Introduction

2.1. SVD

Singular value decomposition (SVD) is a very important matrix decomposition technique in the field of numerical analysis and linear algebra. It has applications in image processing, data reduction, and signal noise reduction [19]. By using SVD, we can remove noise from the vibration signal of the rolling bearing so that a cleaner and more accurate vibration signal can be obtained [20]. In SVD, the problem that needs to be solved at present is how to determine the effective order of the reconstruction matrix after the vibration signal of the rolling bearing is decomposed. Currently, the effective order is determined using methods such as the threshold method and singular entropy increment [21]. However, these methods require relatively high user experience, so the noise reduction effect is not obvious for the vibration signal of the rolling bearing. In order to solve this method, this paper introduces the singular value energy difference spectrum to determine the order so as to achieve the purpose of noise reduction.

The vibration signal of a rolling bearing is usually a one-dimensional signal, which cannot be directly subjected to SVD [22], so the one-dimensional vibration signal $X = \{x_1, x_2, x_3, \cdots, x_N\}$ must be converted into a two-dimensional matrix. Through the Hankel matrix, the signal can be represented as a low-rank approximation. This paper chooses to construct the Hankel matrix. The Hankel matrix $A_{m\times n}$ is shown in Formula (1):

$$A_{m\times n} = \begin{bmatrix} x(1) & \cdots & x(n) \\ \vdots & & \vdots \\ x(m) & \cdots & x(N) \end{bmatrix} = D_{m\times n} + W_{m\times n} \quad (1)$$

In Formula (1), $A_{m\times n}$ is expressed as a constructed Hankel matrix, $N = m + n + 1$. The noise signal is $W_{m\times n}$, and the useful vibration signal is $D_{m\times n}$. When $m = N/2$, the Hankel matrix noise reduction effect is generally more obvious. Because the size parameter m of the Hankel matrix is half the length of the original signal, it may produce a better separation effect, especially in noise reduction applications. This choice can suppress noise to a certain extent and preserve important features of the signal. This choice is mainly considered from the three aspects of capturing signal trend and periodic characteristics, suppressing high-frequency noise, and separating signal and noise.

On the problem of singular value order determination, the order is determined by the singular value energy distribution of the useful signal and the noise signal in the vibration signal of the rolling bearing. The signal energy is shown in Formula (2):

$$E = \sum_{i=1}^{q} \sigma_i^2 \tag{2}$$

In Formula (2), the signal energy is represented by E, σ_{i} represents the singular value, and the total order is q and ends at q. The singular value energy difference spectrum is described below and normalized, as shown in Formula (3):

$$p(i) = \frac{\sigma_i^2 - \sigma_{i+1}^2}{E} \tag{3}$$

In Formula (3), the $p(i)(i = 1, 2, \cdots, q)$ sequence represents the energy difference spectrum. From Formula (3), it can be seen that the energy changes the adjacent orders of the singular value. The singular value energy ratio of the useful signal is relatively large, so a large peak signal is formed. The signal after the peak is generated by the noise signal, and the singular value corresponding to this point is found in the energy difference spectrum. Then, take this point as the order of the reconstructed signal to realize the removal of the noise signal of the rolling bearing.

2.2. GST

Generalized S-transform (GST) is a form of the time–frequency analysis method, which is a combination of time-domain signal analysis and frequency-domain signal analysis [23]. GST provides more detailed and comprehensive signal characterization in the time–frequency domain by jointly analyzing the signal in the time and frequency domains, which can obtain the instantaneous frequency information of the signal [24]. The principle of GST is based on the ideas of short-time Fourier transform (STFT) and continuous wavelet transform (CWT). Its core concept is to perform local spectral analysis of the signal at different time points [25]. The specific principles are as follows:

$$S(f,\tau)_1 = \int_{-\infty}^{+\infty} h(t)w(t-\tau)\exp(-i\cdot 2\pi f t)\mathrm{d}t \tag{4}$$

In Formula (4), $S(f,\tau)$ represents S transformation, $h(t)$ represents the signal to be analyzed, and the translation amount is represented by τ. $w(t)$ represents the Gaussian window function. $w(t) = \frac{1}{\sigma(f)\sqrt{2\pi}}\exp\left(\frac{-t^2}{2\sigma(f)^2}\right)$, and $\sigma(f) = 1/|f|$. GST is modified on the S-transform formula. By adding the parameter m to adjust the Gaussian window width, the time–frequency resolution of the S-transform is improved. GST is shown in Formula (5).

$$S(f,\tau)_2 = \int_{-\infty}^{+\infty} h(t-\tau)w(t)\exp(-i\cdot 2\pi f t)\mathrm{d}t \tag{5}$$

In Formula (5), $w(t) = \frac{1}{\sigma_m(f)\sqrt{2\pi}}\exp\left(\frac{-t^2}{2\sigma_m(f)^2}\right)$, and $\sigma_m(f) = 1/|f|^m$. GST is performed on the one-dimensional vibration signal of the rolling bearing after noise reduction to obtain a two-dimensional time–frequency image. By imaging the vibration signal, the information contained in the vibration signal can be preserved at a high quality, and the deep learning model has good recognition and processing characteristics for the converted two-dimensional image data.

2.3. Vision Transformer

The Transformer model is a classic NLP model proposed by the Google team in 2017 [26]. The Transformer architecture has revolutionized the field of natural language processing and has become the backbone of many state-of-the-art models for a variety

of tasks, including machine translation, text generation, question answering, sentiment analysis, and more. Unlike other models, it uses the self-attention mechanism completely to calculate the input and output. It does not adopt the sequential sequence alignment structure of traditional RNNs and also avoids the problem of the limited receptive field of CNNs. This allows the Transformer to capture global information [27]. The Transformer's multi-attention mechanism enables the extraction of richer feature representations from raw data. This is particularly important for fault diagnosis tasks, as effective feature extraction can improve the accuracy and robustness of the model.

Compared with other deep learning models (CNN, LSTM, etc.), the Transformer model has the following advantages in terms of more intuitive explanations: (1) Attention mechanism. Traditional deep learning models are basically local perceptual information, and contextual information is limited to a certain location or time. But, when dealing with certain problems, it is very important to have a global understanding of the context. For the Transformer model, the entire sequence can be modeled through the attention mechanism to capture global information. (2) Parallel processing. Traditional models (such as LSTM models) must be processed step by step when dealing with timing issues, and the next step can only be performed after the last time step is processed. This can lead to the underutilization of computing resources. The Transformer's self-attention mechanism can process the information of all positions at the same time, which can fully and effectively improve efficiency. (3) Applicability to relationship modeling at different distances. The traditional model has problems such as gradient explosion. The Transformer model relies on the self-attention mechanism to perform weighted attention on location information at different distances, effectively dealing with long-distance dependencies. Figure 1 shows the Transformer model structure.

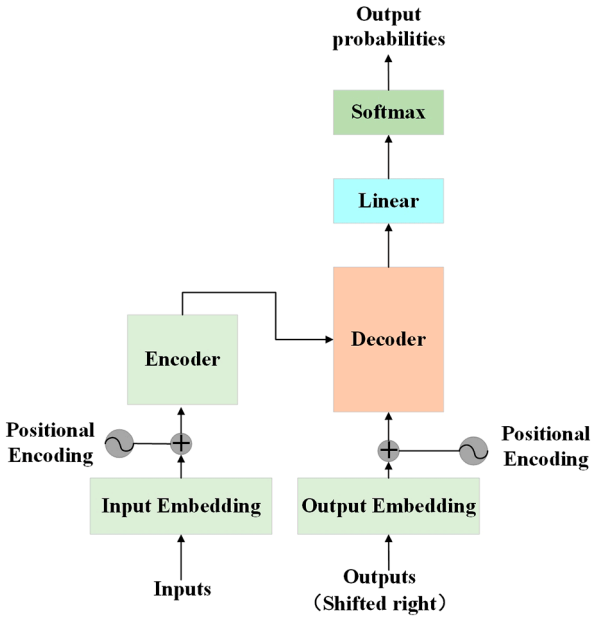


Figure 1. Transformer model structure.

Although the Transformer model is very good, there is a problem that it is not suitable for two-dimensional images. In order to solve this problem, the Vision Transformer (ViT) model came into being. ViT is an image classification model based on the Transformer architecture proposed by Alexey Dosovitskiy et al. [28] in 2020. The basic idea of ViT is to split the image into a series of small patches (patches), convert these small patches into sequence data, and then input the Transformer model for processing. The following are the main principles of ViT:

Embedding module. First, the input image is divided into patches of fixed size. These patches are images that do not overlap in spatial dimensions, similar to dividing an image

into a regular grid. Cut the image with size $[H, W, C]$ into size $[P, P, C]$, and the number of cut image blocks is N. Specifically, this is shown in Formula (6).

$$N = \frac{HW}{P^2} \tag{6}$$

In Formula (6), the height, depth, and width of the input image are H, C, and W, respectively, and the corresponding height and width after clipping are P. Each patch is mapped to a lower-dimensional vector space by a fully connected layer (often called an embedding layer). The parameters of this embedding layer are learned via model training so that each small block can be effectively represented as a vector. The vector length is $X_P = P \times P \times C$. Then, add a classification vector x_{cls}, and add a position code P containing spatial information as the input of the Transformer encoder layer [29].

$$z_0 = \left[x_{\text{cls}}; E\left(x_P^1\right); E\left(x_P^2\right); \cdots; E\left(x_P^M\right)\right] + P \tag{7}$$

In Equation (7), the input of the encoding is z_0. x_{cls} is a category token, and its purpose is to realize the classification task; E is a linear mapping matrix, and P is a position code.

Transformer encoder module. The Transformer encoder is composed of multiple self-attention mechanisms (self-attention) and feed-forward neural network layers, which can learn global and local context dependencies in sequence data [30]. Its structure is shown in Figure 2.

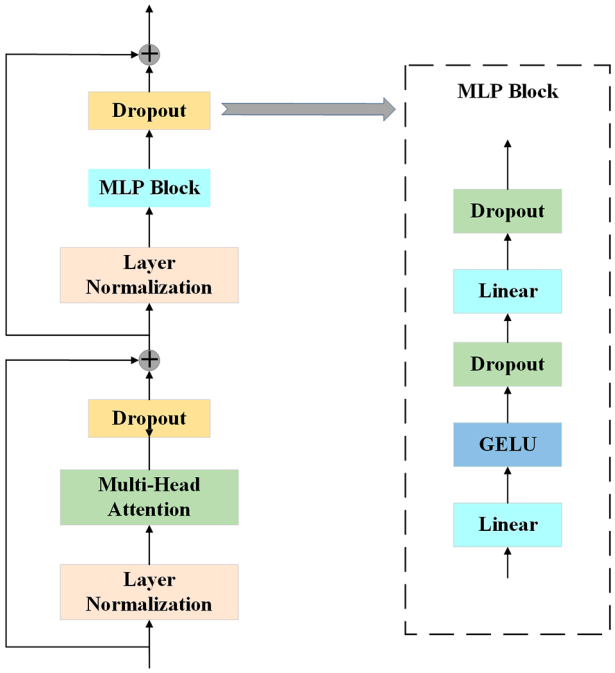


Figure 2. Transformer encoder model structure.

It can be seen from Figure 2 that the information data input into the Transformer encoder is first normalized. After the multi-head self-attention mechanism, the dropout is randomly inactivated, and then, the residual connection is used to fuse with the input information data. The processed data are then normalized. Then, enter the multi-layer perceptron, use the residual connection after dropout, and fuse with the input data again. The multi-layer perceptron (MLP block) consists of a fully connected layer, a GELU activation function, and a dropout module [31].

The last is the classification module. The output sequence of the Transformer encoder is classified and predicted through a fully connected layer, and the classification result of the image is obtained.

3. Fault Diagnosis Model

3.1. Vision Transformer Model

The input of the original Transformer in the NLP field is a one-dimensional word sequence, but the picture is two-dimensional, so ViT changes the input format. In this paper, the rolling bearing vibration signal is converted into a two-dimensional time–frequency image, which can not only retain rich feature information but also facilitate the input of the ViT model. The framework of the ViT model is shown in Figure 3.

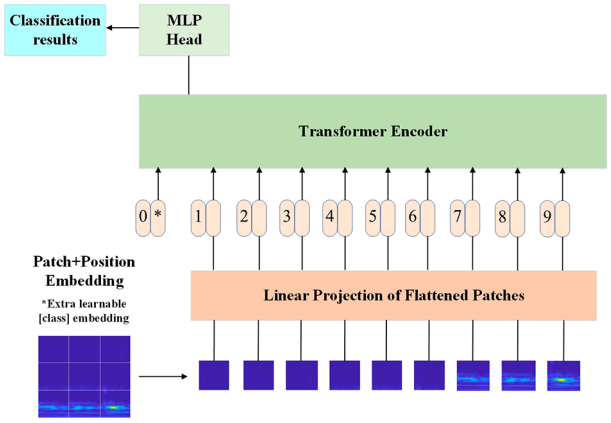


Figure 3. Vision Transformer model framework.

ViT divides the input time–frequency image into multiple non-overlapping patches. These patches are turned into one-dimensional vectors through the embedding layer. Then, compress the dimension of the changed patch sequence through linear projection, which also realizes feature transformation [32]. The above two processes are also called the token process so that the token vector can be obtained, which solves the problem that the dimension is relatively large in a one-dimensional vector. In order to facilitate the subsequent classification of rolling bearing fault types, a learnable class token is introduced to retain image location information [33]. Then, input the class token and the token vector into the Transformer encoder. The Transformer encoder module is introduced in the previous part of the principle. Finally, the output corresponding to the class position is input to the MLP head (composed of the fully connected layer, GELU, and dropout layer) to predict the classification output. The ViT model parameters are shown in Table 1.

Table 1. Vision Transformer model parameters.

Patch Size	Layers	Hidden Size D	MLP Size	Heads	Params
16 × 16	12	768	3072	12	86 M

It can be seen from Table 1 that the input patch size is 16 × 16. Layers represent the number of times the encoder block is stacked. Hidden size D indicates the length of the token vector. MLP size represents the number of fully connected layer nodes. Heads indicates the number of heads. Compared with the standard Transformer, ViT only needs to use its encoder part for image classification and does not require a decoder. The concept of image tokenization operation and the class token is introduced, and some features of the Transformer are retained. Compared with CNNs, the advantages are more obvious when the number of data is larger. At the same time, there is no disadvantage to a limited receptive field, and the feature information is more comprehensively captured.

3.2. Rolling Bearing Fault Diagnosis Model

The overall model of rolling bearing fault diagnosis based on SVD-GST combined with the Vision Transformer is shown in Figure 4.

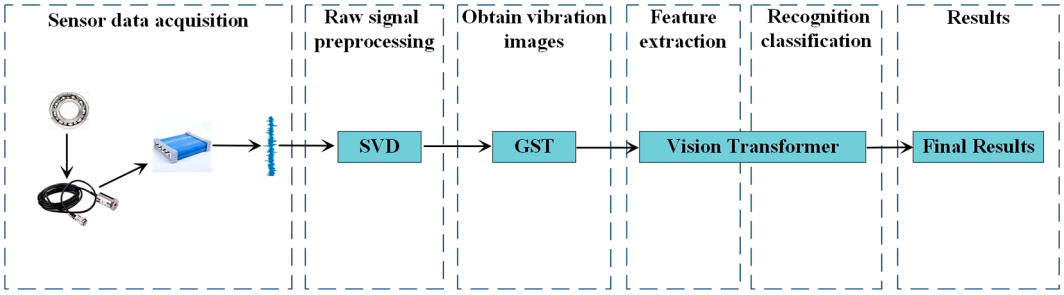


Figure 4. Overall model framework for rolling bearing fault diagnosis.

From Figure 4, we can see that the overall process is divided into the following parts: (1) first, obtain the vibration signal of the rolling bearing through the acceleration sensor, and use the data acquisition card to convert the signal to the PC; (2) use the SVD algorithm to perform noise reduction preprocessing on the collected vibration signal; (3) use the generalized S-transform to convert the one-dimensional vibration signal into a two-dimensional time–frequency image; (4) use the Vision Transformer to perform feature extraction and pattern recognition on time–frequency images and output training results; and (5) after training, use the trained model to classify rolling bearing faults.

4. Fault Diagnosis Platform Construction

The main components of the rolling bearing fault diagnosis experimental platform are as follows: rolling bearings (6406), a magnetic powder brake (FZ-A-12), a three-phase asynchronous motor (YE3-100L2-4), a piezoelectric acceleration sensor (CAYD051V), a frequency converter (G7R5/P011T4), a data acquisition card (YE6231), and a PC. In this experiment, five fault categories of rolling bearing inner ring faults, rolling element faults, cage fracture faults, outer ring faults, and normal rolling bearings were designed. The specific fault form is shown in Figure 5.

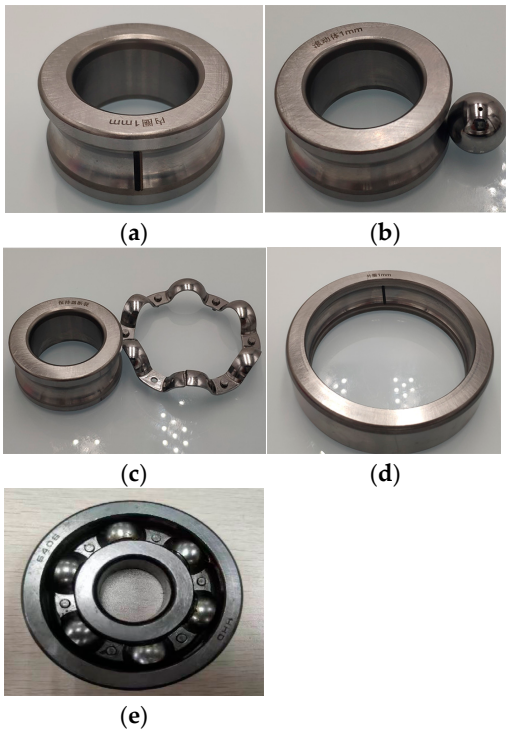


Figure 5. Rolling bearing fault types: (**a**) inner ring fault; (**b**) rolling element fault; (**c**) cage fracture fault; (**d**) outer ring fault; and (**e**) normal state.

Figure 5a–e, respectively, show the five states of rolling bearing inner ring failure, rolling element failure, cage fracture failure, outer ring failure, and normal rolling bearing. The physical map of the rolling bearing fault diagnosis experiment platform is shown in Figure 6, and the experimental process is shown in Figure 7.

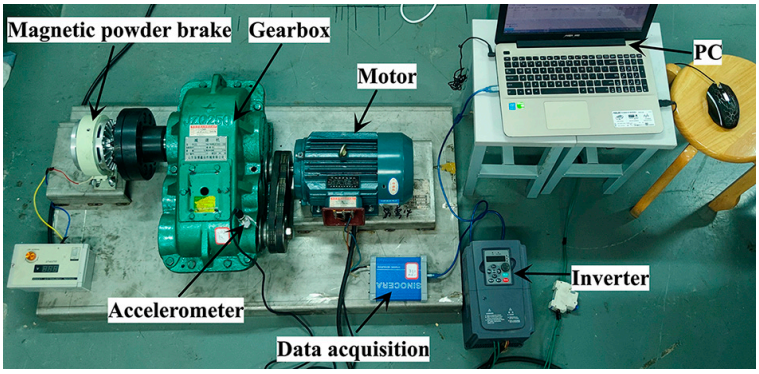


Figure 6. The physical picture of the rolling bearing fault diagnosis experiment platform.

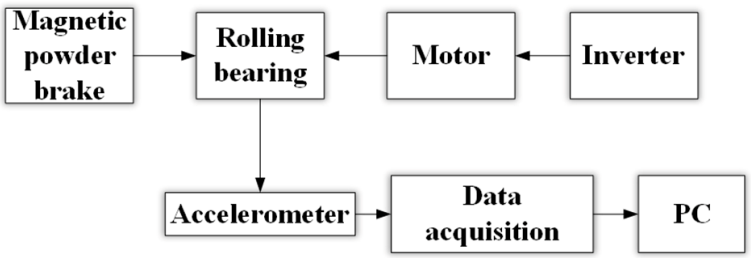


Figure 7. Fault diagnosis experiment flowchart of rolling bearing.

The specific experimental steps are as follows: To ensure safety, first, add an air switch between the power supply and the inverter. Connect the inverter to the motor. The motor and the gearbox are connected by a belt (the rolling bearing in the gearbox is tested in this experiment). The intermediary between the gearbox and the magnetic powder brake is through a coupling. An acceleration sensor is installed on the end cover of the gearbox, the vibration signal is obtained through the sensor, and the signal data information is transmitted to the PC using a data acquisition card.

This experiment is a no-load experiment, so the magnetic powder brake is closed during the experiment. The speed of the three-phase asynchronous motor is 900 r/min. The sampling frequency is 6 kHz. The experimental data information is shown in Table 2.

Table 2. Experimental data information.

Category	Rolling Bearing Status	Motor Speed (Hz)	Brake Load (A)	Length	Number of Data Sets
1	cage fracture fault	30	0	1024	1000
2	normal state	30	0	1024	1000
3	inner ring fault	30	0	1024	1000
4	rolling element fault	30	0	1024	1000
5	outer ring fault	30	0	1024	1000

In this experiment, 1024 points comprise a set of data lengths. Each state collects 1000 groups, so there are 5000 groups in total. The training set, verification set, and test set are divided according to 7:2:1, that is, 3500 training sets, 1000 verification sets, and 500 test sets. The specific division is shown in Table 3.

Table 3. Data set division.

Gearbox Status	Number of Training Sets	Number of Validation Sets	Number of Test Sets	Label
cage fracture fault	700	200	100	0
normal state	700	200	100	1
inner ring fault	700	200	100	2
rolling element fault	700	200	100	3
outer ring fault	700	200	100	4
total number of samples	3500	1000	500	

5. Result Analysis

5.1. Rolling Bearing Vibration Signal Preprocessing

This article uses SVD to denoise the vibration signal of rolling bearings. In order to solve the problem of determining the effective order of the reconstruction matrix after decomposing the vibration signal of rolling bearings, a singular value energy difference spectrum was introduced to determine the order, thus achieving the purpose of noise reduction in vibration signals. Take the first 500 singular values for analysis this time. Figure 8 shows the singular value distribution curve of the vibration signal of rolling bearings. Figure 9 shows the singular value energy difference spectrum of the vibration signal of rolling bearings. Figure 10 shows the vibration signal of rolling bearings before and after noise reduction.

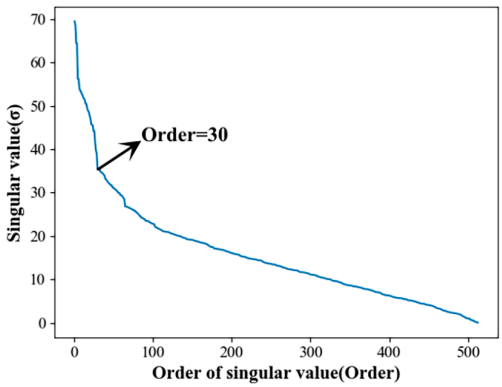


Figure 8. Singular value distribution curve of rolling bearing vibration signal.

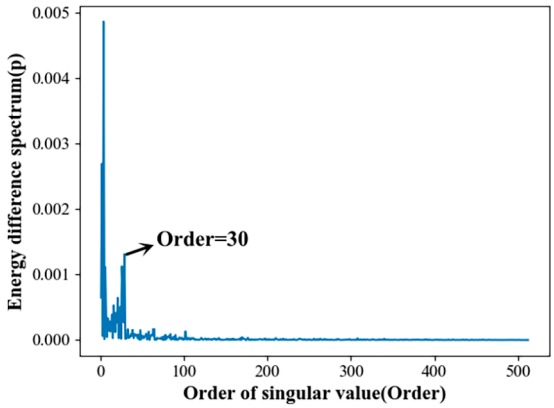


Figure 9. Singular value energy difference spectrum of rolling bearing vibration signal.

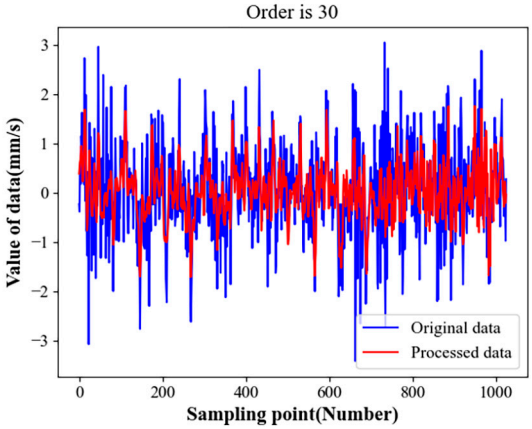


Figure 10. Rolling bearing vibration signal before and after noise reduction display.

In Figure 8, after the singular value order of the overall curve is 30, there is a turning point in the gradient descent trend, and the decline rate becomes relatively gentle compared with the previous one. Since the curve is not too obvious, the singular value energy difference spectrum in Figure 9 can be clearer, which is why the energy difference spectrum is introduced. In Figure 9, when the order is 30, it can be clearly seen that the peak signal changes abruptly, indicating that this is the dividing point between the vibration signal of the rolling bearing and the useful signal. The useful signal is located before the order is 30, and the noise signal is located after the order is 30. Therefore, the singular value order of the reconstructed signal is 30; that is, the first 30 singular value signals are reconstructed. In Figure 10, it can be seen that the periodicity of the signal after noise reduction is more prominent overall, and the quality and readability of useful signals are improved.

5.2. Two-Dimensional Time–Frequency Image Acquisition

Perform a generalized S-transform on the denoised vibration signal of the rolling bearing to convert the one-dimensional vibration signal into a two-dimensional time–frequency image. Figure 11 shows the time–frequency images of rolling bearings under five fault states.

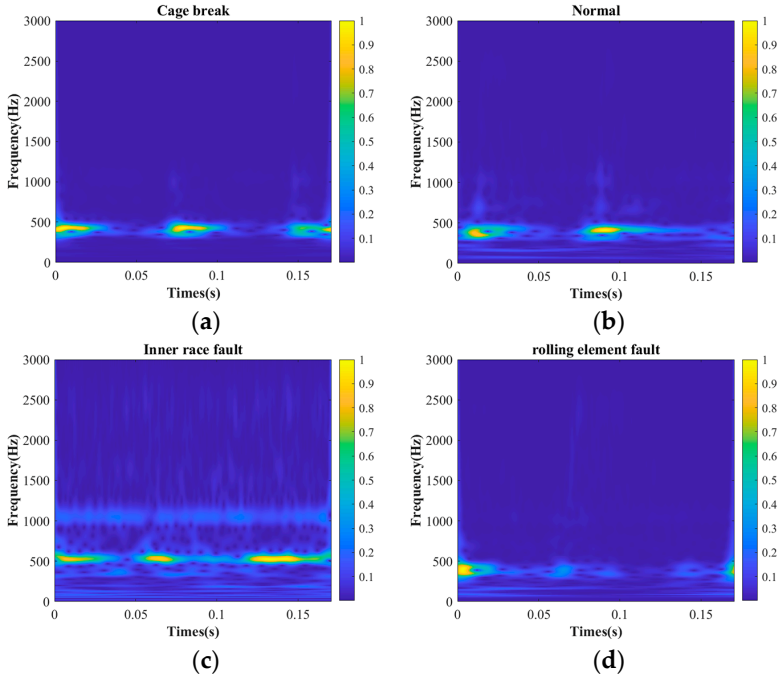
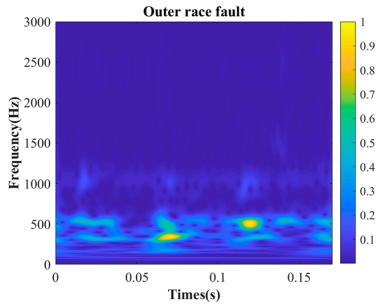


Figure 11. *Cont.*



(**e**)

Figure 11. Two-dimensional time–frequency diagram of rolling bearing: (**a**) cage fracture fault; (**b**) normal state; (**c**) inner ring fault; (**d**) rolling element fault; (**e**) outer ring fault.

Figure 11a–e are the two-dimensional time–frequency diagrams of five fault states of cage fracture, normal state, inner ring fault, rolling element fault, and outer ring fault, respectively. The generalized S-transform can provide high time–frequency resolution by interpolating the signal on the time–frequency plane. This makes it better able to capture subtle changes in the signal in both the time and frequency domains.

5.3. Fault Diagnosis Model Analysis

This experiment is carried out on the Windows 11 operating system. The programming language is Python, and the Pytorch deep learning framework is used. The processed two-dimensional time–frequency image is input into the Vision Transformer model, and the process of feature extraction and pattern recognition training and classification is completed in this model. Figure 12 shows the loss value curves corresponding to the training set and the verification set. Figure 13 is the conversion curve of the training set and verification set accuracy.

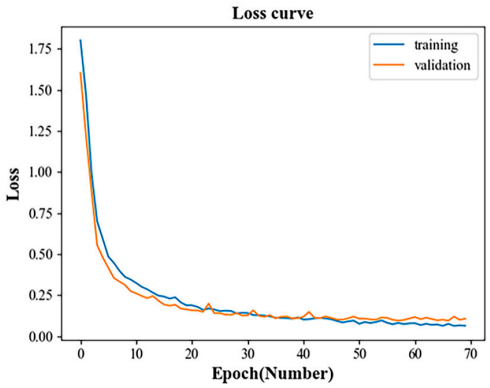


Figure 12. Loss value change curve.

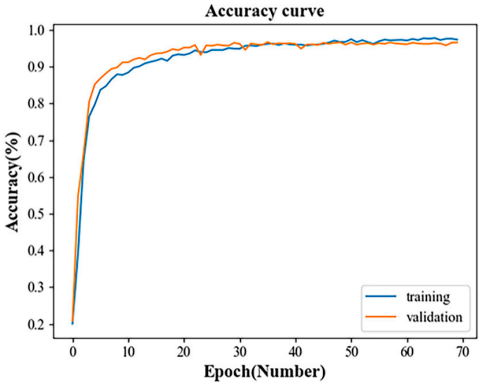


Figure 13. Accuracy transformation curve.

It can be seen from Figure 12 that as the number of training samples increases, the loss value gradually decreases until it finally decreases to a more stable value. In the 0–10 training period, the loss value changes most obviously, indicating that the convergence accelerates during this period. When the number of iterations reaches 10–40 times, the loss transformation is relatively slow, indicating that as the number of iterations increases, the training sample and the verification sample have a convergence trend. After 40 iterations, the training samples and verification samples are basically not fluctuating and have overlapped, indicating that the training has ended at this time.

It can be seen from Figure 13 that as the number of training samples increases, the accuracy of the training samples and verification samples continues to increase until finally reaching a stable value. In the 0–10 training times, the accuracy rate rises very rapidly, indicating that the model is learning rapidly at this time. Within 10–40 training times, the training samples and verification samples start to rise slowly, indicating that the model training is basically becoming mature at this time. After 40 training times, the accuracy curve can basically be maintained at about 97%. It is smooth and stable without large fluctuations, indicating that the model has been trained.

5.4. Analysis of Fault Diagnosis Results

In order to demonstrate the recognition effect of the model on various faults of rolling bearings, this experiment uses a confusion matrix to represent the classification results. The details are shown in Figure 14.

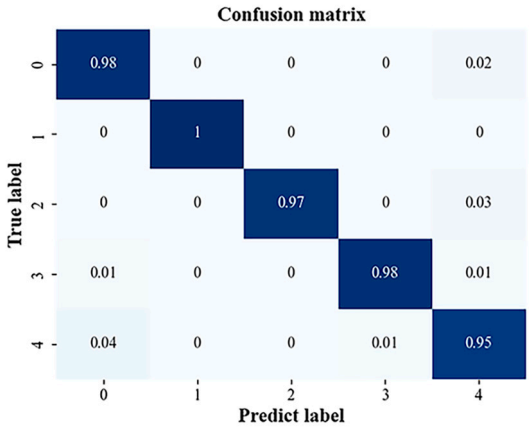


Figure 14. Diagnosis results confusion matrix.

It can be seen in Figure 14 that the model proposed in this paper has the highest recognition rate for label 1 (normal state), reaching 100%. The identification rates of label 0 (cage fracture) and label 3 (rolling element failure) can reach 98%. In total, 2% of cage fracture faults are misidentified as outer ring faults. In the process of rolling element fault identification, 1% of them are misidentified as cage fracture faults and outer ring faults. The model can identify label 2 (inner ring fault) with a recognition rate of 97%, of which 3% are mistaken for outer ring faults. The model has the lowest recognition accuracy rate for label 4 (outer ring fault), which is 95%, of which 4% are misidentified as cage fractures and 1% are misidentified as inner ring faults. Overall, the model recognition rate is high, up to 97.6%, which also proves that the model proposed in this paper is feasible.

5.5. Comparison of Models

In order to illustrate the advanced nature of the rolling bearing fault diagnosis model based on SVD-GST combined with the Vision Transformer (SVD-GST-ViT), under the same experimental conditions, the model was compared with the popular CNN and LSTM models in deep learning. The training, validation, and test sets are divided according to 7:2:1. All models undergo the same data preprocessing. For each model, set the same

hyperparameters, such as the number of iterations, batch size, etc. In order to prevent the accidental occurrence of a single recognition rate, the above methods are trained and tested 10 times, respectively. Table 4 shows the average accuracy and standard deviation of 10 tests of different methods:

Table 4. The average accuracy of different methods.

Fault Diagnosis Model	10 Average Accuracy %	Standard Deviation %
SVD-GST-2DCNN	95.24	1.2933
STFT-CNN-LSTM	92.50	0.6520
SVD-GST-LSTM	94.28	1.7863
GST-ViT	91.06	0.9834
SVD-GST-ViT	98.52	0.4266

As can be seen in Table 3, the average recognition accuracy of the proposed model (SVD-GST-ViT) method is the highest, reaching 98.52%. Compared with the method of directly using GST-ViT, it increases by 7.46%, indicating that the original vibration signal can remove redundancy and enhance features after SVD decomposition. Compared with SVD-GST-2DCNN, STFT-CNN-LSTM, and SVD-GST-LSTM, it also increases by 3.28%, 6.02%, and 4.24%, respectively. The standard deviation of the model proposed in this paper is the smallest, which is 0.4266%. Compared comprehensively, the model has high accuracy and stability.

6. Conclusions

This paper proposes a rolling bearing fault diagnosis model based on SVD-GST combined with the Vision Transformer. SVD is used for noise reduction processing to solve the problem that in the fault diagnosis of rolling bearings, the collected vibration signals contain interference from complex noise and redundant components, which affects subsequent feature extraction and pattern recognition. A generalized S-transform is proposed to convert a 1D vibration image into a 2D time–frequency image. It solves the problem that the recognition rate is difficult to further improve because of the loss of signal information in the one-dimensional data processing and industrial practice of the bearing fault diagnosis method based on deep learning, making full use of the advantages of deep learning in image classification and prediction with higher recognition accuracy. At the same time, in order to avoid the problems of limited receptive fields in CNNs and the need for step-by-step calculations in time sequences when an RNN processes sequence data, the Vision Transformer model is proposed. The experimental results show that the multiple average accuracy rate of the fault diagnosis model adopted in this paper is 98.52% for different fault states of rolling bearings. Compared with other model methods, it can effectively improve the fault identification effect of rolling bearings.

For future research on rolling bearing fault diagnosis, multimodal data fusion can be considered. This article only uses the vibration signal of the rolling bearing. In addition, other sensor data, such as current and temperature, can also be considered to obtain more comprehensive fault diagnosis information. The advantages of the Transformer model can be fully utilized in dealing with multimodal data problems. In conclusion, the research direction of applying the Transformer model to rolling bearing fault diagnosis is worthy of further exploration.

Author Contributions: Conceptualization, F.X.; data curation, Q.F.; formal analysis, F.X.; methodology, G.W. and H.Z.; project administration, G.W.; software, F.X. and G.W.; supervision, H.Z.; validation, E.S.; visualization, G.W.; writing—original draft, G.W. and Q.F.; writing—review & editing, F.X., E.S. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (52265068; 52162045), the Natural Science Foundation of Jiangxi Province (20224BAB204050; 20224BAB204040),

the Carrier and Equipment Key Laboratory Project of the Ministry of Education (KLCEZ2022-02), and the Project of Jiangxi Provincial Department of Education (GJJ2200627).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding authors upon request.

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

References

1. Zhang, Q.; Gao, J.; Dong, H.; Mao, Y. WPD and DE/BBO-RBFNN for solution of rolling bearing fault diagnosis. *Neurocomputing* **2018**, *312*, 27–33. [CrossRef]
2. Li, L.; Meng, W.; Liu, X.; Fei, J. Research on Rolling Bearing Fault Diagnosis Based on Variational Modal Decomposition Parameter Optimization and an Improved Support Vector Machine. *Electronics* **2023**, *12*, 1290. [CrossRef]
3. Zhou, J.; Xiao, M.; Niu, Y.; Ji, G. Rolling Bearing Fault Diagnosis Based on WGWOA-VMD-SVM. *Sensors* **2022**, *22*, 6281. [CrossRef] [PubMed]
4. Fuan, W.; Hongkai, J.; Haidong, S.; Wenjing, D.; Shuaipeng, W. An adaptive deep convolutional neural network for rolling bearing fault diagnosis. *Meas. Sci. Technol.* **2017**, *28*, 095005. [CrossRef]
5. Meng, Z.; Zhan, X.; Li, J.; Pan, Z. An enhancement denoising autoencoder for rolling bearing fault diagnosis. *Measurement* **2018**, *130*, 448–454. [CrossRef]
6. Li, X.; Jiang, H.; Wang, R.; Niu, M. Rolling bearing fault diagnosis using optimal ensemble deep transfer network. *Knowl. Based Syst.* **2021**, *213*, 106695. [CrossRef]
7. Jia, F.; Lei, Y.; Guo, L.; Lin, J.; Xing, S. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing* **2018**, *272*, 619–628. [CrossRef]
8. Cheng, C.; Wang, J.; Chen, H.; Chen, Z.; Luo, H.; Xie, P. A Review of Intelligent Fault Diagnosis for High-Speed Trains: Qualitative Approaches. *Entropy* **2021**, *23*, 1. [CrossRef] [PubMed]
9. Duan, L.; Xie, M.; Wang, J.; Bai, T. Deep learning enabled intelligent fault diagnosis: Overview and applications. *J. Intell. Fuzzy Syst.* **2018**, *35*, 5771–5784. [CrossRef]
10. Li, X.; Zhang, W.; Ding, Q.; Sun, J.Q. Multi-layer domain adaptation method for rolling bearing fault diagnosis. *Signal Process.* **2019**, *157*, 180–197. [CrossRef]
11. Liu, X.; Zhao, X.; He, K. Feasibility Study of the GST-SVD in Extracting the Fault Feature of Rolling Bearing under Variable Conditions. *Chin. J. Mech. Eng.* **2022**, *35*, 1–14. [CrossRef]
12. Pan, H.; He, X.; Tang, S.; Meng, F. An improved bearing fault diagnosis method using one-dimensional CNN and LSTM. *Stroj. Vestn./J. Mech. Eng.* **2018**, *64*, 443–452.
13. Huang, M.; Huang, T.; Zhao, Y.; Dai, W. Fault diagnosis of rolling bearing based on empirical mode decomposition and convolutional recurrent neural network. In *Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2021; Volume 1043, p. 42015.
14. Lu, Z.; Qin, Y.; Cheng, X.; Zhang, S.; Zeng, Y. Bearing Fault Diagnosis Method of Bearing Based on LSTM Auto-Encoder. In *International Conference on Electrical and Information Technologies for Rail Transportation*; Springer: Singapore, 2021; pp. 582–591.
15. Tang, X.; Xu, Z.; Wang, Z. A Novel Fault Diagnosis Method of Rolling Bearing Based on Integrated Vision Transformer Model. *Sensors* **2022**, *22*, 3878. [CrossRef]
16. Ding, Y.; Jia, M.; Miao, Q.; Cao, Y. A novel time–frequency Transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. *Mech. Syst. Signal Process.* **2022**, *168*, 108616. [CrossRef]
17. Wang, J.; Mo, Z.; Zhang, H.; Miao, Q. A deep learning method for bearing fault diagnosis based on time-frequency image. *IEEE Access* **2019**, *7*, 42373–42383. [CrossRef]
18. Kim, J.H. Time frequency image and artificial neural network based classification of impact noise for machine fault diagnosis. *Int. J. Precis. Eng. Manuf.* **2018**, *19*, 821–827. [CrossRef]
19. Yang, B.; Liu, R.; Chen, X. Fault diagnosis for a wind turbine generator bearing via sparse representation and shift-invariant K-SVD. *IEEE Trans. Ind. Inform.* **2017**, *13*, 1321–1331. [CrossRef]
20. Yang, Z.-X.; Zhong, J.-H. A Hybrid EEMD-Based SampEn and SVD for Acoustic Signal Processing and Fault Diagnosis. *Entropy* **2016**, *18*, 112. [CrossRef]
21. Cheng, H.; Zhang, Y.; Lu, W.; Yang, Z. A bearing fault diagnosis method based on VMD-SVD and Fuzzy clustering. *Int. J. Pattern Recognit. Artif. Intell.* **2019**, *33*, 1950018. [CrossRef]
22. Tian, Y.; Ma, J.; Lu, C.; Wang, Z. Rolling bearing fault diagnosis under variable conditions using LMD-SVD and extreme learning machine. *Mech. Mach. Theory* **2015**, *90*, 175–186. [CrossRef]
23. Ma, Q.; Huang, D.; Yang, J. Adaptive stochastic resonance in second-order system with general scale transformation for weak feature extraction and its application in bearing fault diagnosis. *Fluct. Noise Lett.* **2018**, *17*, 1850009. [CrossRef]

24. Qin, Y.; Shi, X. Fault Diagnosis Method for Rolling Bearings Based on Two-Channel CNN under Unbalanced Datasets. *Appl. Sci.* **2022**, *12*, 8474. [CrossRef]
25. Liu, X.; He, Y.; Wang, L. Adaptive Transfer Learning Based on a Two-Stream Densely Connected Residual Shrinkage Network for Transformer Fault Diagnosis over Vibration Signals. *Electronics* **2021**, *10*, 2130. [CrossRef]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762v7.
27. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
29. Xu, Z.; Tang, X.; Wang, Z. A Multi-Information Fusion ViT Model and Its Application to the Fault Diagnosis of Bearing with Small Data Samples. *Machines* **2023**, *11*, 277. [CrossRef]
30. Liang, P.; Yu, Z.; Wang, B.; Xu, X.; Tian, J. Fault transfer diagnosis of rolling bearings across multiple working conditions via subdomain adaptation and improved vision transformer network. *Adv. Eng. Inform.* **2023**, *57*, 102075. [CrossRef]
31. Chen, Z.; Chen, J.; Liu, S.; Feng, Y.; He, S.; Xu, E. Multi-channel Calibrated Transformer with Shifted Windows for few-shot fault diagnosis under sharp speed variation. *ISA Trans.* **2022**, *131*, 501–515. [CrossRef] [PubMed]
32. Jin, Y.; Hou, L.; Chen, Y. A time series transformer based method for the rotating machinery fault diagnosis. *Neurocomputing* **2022**, *494*, 379–395. [CrossRef]
33. Xu, Y.; Li, Y.; Wang, Y.; Zhong, D.; Zhang, G. Improved few-shot learning method for transformer fault diagnosis based on approximation space and belief functions. *Expert Syst. Appl.* **2021**, *167*, 114105. [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.