Article

Bearing Non-Uniform Loading Condition Monitoring Based on Dual-Channel Fusion Improved DenseNet Network

Yanfei Zhang [1,2,*], Yang Liu [1], Lijie Wang [1], Dongya Li [2], Wenxue Zhang [3] and Lingfei Kong [1]

1   School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology,
     Xi'an 710048, China; 2210221250@stu.xaut.edu.cn (Y.L.); wanglijie@xaut.edu.cn (L.W.);
     lingfeikong@xaut.edu.cn (L.K.)
2   Luoyang Bearing Science & Technology Co., Ltd., Luoyang 471039, China; zyszzbldy@163.com
3   Hangzhou Wren Hydraulic Equipment Manufacturing Co., Ltd., Hangzhou 311100, China;
     jsfwb@wrenchina.com
*   Correspondence: yfzhang@xaut.edu.cn

Abstract: Misalignment or unbalanced loading of machine tool spindle bearings often results in skewed bearing operation, which makes the spindle more susceptible to failure. In addition, due to the weak impact signal of the bearing in skewed operation, a single feature information cannot accurately characterize the operation status of the bearing. To address the above problems, this paper proposes a method to monitor the uneven running state of bearing load based on a dual-channel fusion improved dense connection (DenseNet) network. First, the original signal is pre-processed by overlapping sampling method, and the dual-channel experimental data are obtained by frequency-domain and time-frequency-domain algorithms; then the processed data are input into the improved 1D-DenseNet and 2D-DenseNet models respectively for feature extraction; then the frequency-domain and time-frequency-domain features are fused by concat splicing operation, and the output belongs to each category The probability distribution is used to characterize the operating state of the bearings. Finally, the validity of the algorithm model is verified by using the Case Western Reserve University public rolling bearing data set, and an experimental bench is designed and built for experimental verification of the uneven bearing load operation. The comparative analysis of the experimental results in this paper shows that the algorithm can extract the features of the input signal more comprehensively and finally achieve 100% recognition accuracy.

Keywords: rolling bearing; condition monitoring; DenseNet network; dual channel; feature fusion

1. Introduction

As the "workhorse" of the equipment manufacturing industry, machine tools are one of the most important tools in industrial production, with applications covering the mechanical industry, automotive industry, electric power equipment, railroad locomotives and aerospace. As the core component of machine tools, the operating condition of the spindle directly affects the machining accuracy and efficiency of machine tools, while the bearing, as the spindle support component of machine tools, directly affects the rotational accuracy of the spindle due to its assembly accuracy and the same of the enemy [1–3]. Due to long-term service in variable loads, high temperatures, shock, and other harsh environments, and by manufacturing errors, assembly accuracy and human operation errors, and other factors, will lead to bearing deflection in the service process, which is very easy to cause bearing failure. Therefore, accurate and efficient real-time condition monitoring of bearings is important to ensure the healthy operation of machine tools and improve productivity [4].

Traditional bearing condition monitoring methods are often studied with the help of time-domain features of the signal, and a few methods consider feature extraction in both frequency and time-frequency domains and use the extracted feature information for

condition monitoring [5]. Previous mechanical fault diagnosis models based on deep learning have poor generalization ability and complex networks, Deng Mingyang et al. [6] combined frequency domain feature extraction self-encoder with variational self-encoder and proposed frequency domain feature variational self-encoder, which makes the extracted features more robust. Using the overall similarity of the same mode class vibration observation samples on the FFT amplitude spectrum feature waveform, Jiao Weidong et al. [7] proposed a fault diagnosis method based on the pattern matching of the frequency domain feature waveform. Kullbak-Leibler (KL) distance mutual parameter method can solve the blind deconvolution order uncertainty problem, Liu Feng [8] et al. proposed an improved time domain blind deconvolution order uncertainty method based on the combination of generalized morphological filtering and improved KL distance combination of improved time-domain blind deconvolution fault feature extraction algorithm, and the method can effectively extract rolling bearing fault features. Weimin Li [9] et al. proposed a diagnosis method based on a frequency-domain sparse classification algorithm, which effectively overcomes noise interference and avoids the problem of fault feature frequency estimation. For the cliffness, margin, and spectral cliffness, which are usually very sensitive to the data singularity of the signal due to chance factors, are easy to cause misjudgment in the condition monitoring of bearings, Wang Xiaoling [10] et al. proposed a frequency band entropy method based on time-frequency analysis and information entropy theory for rolling bearing fault monitoring. In order to make comprehensive use of the time-frequency domain information of vibration signal and the complexity characteristics of measuring time-frequency distribution, Jiaqi Li et al. [11] introduced two-dimensional multiscale entropy into the fault diagnosis of rolling bearings and proposed a rolling bearing fault diagnosis method based on two-dimensional time-frequency multiscale entropy and firefly algorithm optimized support vector machine. The above method only considers the fault feature extraction in the time domain, frequency domain, or time-frequency domain separately, which has certain limitations in bearing fault diagnosis and is difficult to reflect its fault state accurately, and it is difficult to guarantee the mapping relationship between its feature value and service state as the amount of data increases.

With the development of computer hardware, machine learning has become a very effective tool for classification, for which the classification problem is the basis and many applications have evolved from it. Machine learning is able to learn the laws and patterns of data with the help of computers in a large amount of data, and in the process of learning, the potential and valuable information within the data is mined deeply [12–14]. In order to improve the monitoring speed and monitoring accuracy, many scholars have introduced machine learning methods into the field of condition monitoring and achieved good results. Traditional shallow machine learning methods include support vector machines, decision trees, K-nearest neighbors, plain Bayes, artificial neural networks, etc., which, due to their need for large amounts of prior knowledge, lead to difficult feature extraction and selection [15,16].

In recent years, under the impact of the wave of artificial intelligence, people have started to introduce end-to-end deep learning methods into the fault diagnosis collar, and deep learning models have provided new ideas for fault diagnosis research by getting rid of the inevitable uncertainty of manual feature extraction methods [17]. A convolutional neural network (CNN) is a class of feedforward neural network that contains convolutional computation and has a deep structure, and is one of the representative algorithms of deep learning [18]. The algorithm has the capability of representational learning and is able to classify the input information in a translation-invariant manner according to its hierarchical structure. Janssens [19] proposed a three-layer CNN model for bearing fault detection using vibration signals, where a discrete Fourier transform is applied to the data before the model is trained and fed into the network model. Gu [20] proposed to feed the original vibration signals into 1-DCNN and Gu [21] proposed an improved convolutional neural network model with global mean pooling instead of the final fully connected layer of CNN for the purpose of reducing the number of parameters, which effectively improved the

computational speed of the model. To address the problem that traditional fault diagnosis methods require manual feature extraction and feature information is difficult to be fully mined, Chen Ke et al. [22] proposed an end-to-end bearing fault diagnosis method based on CNN, LSTM, and attention mechanism. To address the problem that the traditional bearing fault diagnosis method does not sufficiently extract key features under strong noise and variable load, Yang Xianglan et al. [23] proposed an ECA_ResNet-based bearing fault diagnosis method.

As the layers of a neural network become deeper, the path from the output to the input becomes longer, which leads to the disappearance of the change gradient in the process of backpropagation back to the input. To address this problem, densely connected neural network (DenseNet), as an improved algorithm of the CNN network, solves this problem by establishing dense connections between all the previous layers and the later layers to reuse the features, and some research have been conducted by related scholars. Huang et al. [24] proposed a densely connected convolutional network (DenseNet), which improves the learning efficiency by feature reuse. enhance the learning efficiency, which is the most advanced convolutional neural network architecture. c Shi [25] proposed a wear-induced internal leakage fault diagnosis method based on intrinsic modal functions (IMFs) and weighted densely connected convolutional networks (WDenseNets), using the weighted optimal IMF components as inputs to WDenseNet for fault identification and classification. Yufeng Wang [26] proposed an improved one-dimensional DenseNet network structure capable of handling one-dimensional spectral sequences to achieve multi-scale for signal feature extraction. Rexiang Niu et al. [27] proposed an improved fault diagnosis method for densely connected convolutional networks, which extracts features through multi-scale convolutional layers and achieves weighting of multi-scale feature channels using an attention mechanism to improve the generalization performance of the model. Qingrong Wang et al. [28] proposed a dual-channel cross-dense connected fault diagnosis model incorporating parallel ECA modules, and designed a multi-convolutional residual module and a multi-scale dense connected network for fault feature extraction to achieve interaction and integration of fault information. To address the problem of insufficient ability of shallow features to characterize the fault information of vibration signals, K. Wang et al. [29] proposed an intelligent fault diagnosis neural network model combining a style recalibration module and a densely connected convolutional neural network. Some of the above methods do not give full play to the powerful feature extraction capability of the DenseNet network, the model network structure is shallow, the network generalization capability is weak, or the frequency domain and time-frequency domain signal feature extraction is not considered.

In summary, this paper proposes a method to evaluate the bearing load inhomogeneous operating state based on a dual-channel fusion improved DenseNet network, firstly, data enhancement is performed on the original sample data by overlapping sampling method; secondly, frequency domain and time-frequency domain transformations are performed on the processed data to obtain the experimental data set of two channels; then the processed experimental data are input into the improved 1D-DenseNet and 2D-DenseNet models for feature extraction; finally, the frequency-domain and time-frequency-domain features are fused by concat splicing operation to achieve the classification and recognition of the load inhomogeneity of bearings. The method uses densely connected networks to build the base model, which greatly increases the depth of the network structure and enhances the generalization ability of the model. The effectiveness of the proposed method is verified through experiments, which provides a new idea for rolling bearing fault diagnosis.

2. Fundamentals

2.1. Convolutional Neural Network

2.1.1. Principle of Convolutional Neural Network (CNN)

CNN is a supervised learning neural network with a multilayer convolutional structure, introduced by Hubel and Wiesel with the concept of Receptive fields [30]. First proposed by LeCun for image processing [31], it differs from the traditional neural network

structure in that it consists of two main parts: a convolutional layer and a pooling layer. The convolutional layer is connected to the previous layer by local connectivity and weight sharing, which greatly reduces the number of required parameters; the downsampling layer is a method for feature dimensionality reduction, i.e., reducing the complexity of the network by reducing the input feature parameters, which not only improves the robustness of the neural network but also prevents the occurrence of overfitting [32,33].

The convolutional layer is the core of CNN, which mainly implements the feature extraction of the dataset, which is one of the most important differences from traditional neural networks. The features of each layer in the convolutional layer are obtained by convolving the convolutional kernel with the input features of the previous layer, and the parameters can be adjusted through training to obtain the optimal features. In practice, smaller convolutional kernels are used to reduce the amount of operations. Each convolutional kernel can be used as a tool for feature extraction, and a new feature map is generated by means of convolutional operations. The convolution operation is the process of letting the convolution kernel slide along the coordinate position of the input image or input signal horizontally or vertically for a certain number of steps to compute the data corresponding to it, and the computational equation is as follows [34]:

$$x_j^i = f(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_i^l) \tag{1}$$

where M represents the set of input feature maps, $*$ represents the convolution operation, k represents the convolution kernel, b represents the bias term, x_j^i represents the jth feature map of the ith layer, and $f(-)$ represents the nonlinear activation function used to enhance the representation of the data.

Pooling layer, also known as downsampling layer, can reduce the dimensionality of the input feature set, reduce the computational effort of the neural network, and increase the perceptual field of the posterior neurons to achieve the effect of effective control of overfitting. The pooling operation can be divided into maximum pooling and average pooling, among which maximum pooling is more widely used. Maximum pooling plans the input features into several regions and outputs the maximum value of each region separately.

2.1.2. ResNet Network

As the depth of the network keeps increasing, CNN models begin to suffer from a series of problems such as gradient disappearance and explosion, which in turn lead to a decrease in the accuracy of the model. Based on this, He et al. [35] proposed ResNet by borrowing the idea of cross-layer linking of high-speed networks, the core of which is the residual block. Since this network structure utilizes the residual technique, it is also known as the residual network, and the specific structure is shown in Figure 1:
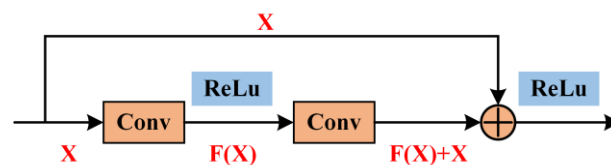

Figure 1. Structure of residual block.

The structural expression of the residual network is as follows:

$$H(x) = F(x) + x \tag{2}$$

where $H(x)$ denotes the output of the structure, x denotes the input, and $F(x)$ denotes the output obtained from the convolution layer. When the parameter of the convolution layer is 0, the formula is expressed as $H(x) = F(x)$. This is the core idea of the residual network, which achieves the transfer of feature information of each layer by constant connection,

which ensures the depth of the network and improves the performance of the network model at the same time.

2.2. DenseNet Network

A dense convolutional neural network is an improved convolutional neural network algorithm based on the residual network (ResNet), which aims to alleviate the problem of gradient disappearance and model degradation by using fewer parameters. The core idea of a dense convolutional network as a neural network with dense connectivity is cross-layer connectivity, where each layer of input in the network model takes as input the feature information output from all previous layers, while the features of that layer are also directly passed to all subsequent layers as input to ensure maximum information transfer between layers, making the network perform similar deep supervision in an implicit way [24].

The DenseNet network proposes a new structure by multiplexing the features, which not only slows down the gradient disappearance, but also has a smaller number of parameters, and it is connected in the form of cross-channel with the formula:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{3}$$

where: is the input to the x_0 network; x_l is the output of layer l in the network; x_{l-1} is the input to layer $l-1$ of the network. $H_l(\cdot)$ is the nonlinear transformation operation acting on layer l.

DenseNet mainly consists of convolutional layer, pooling layer, DenseBlock, TransitionLayer, and linear classification layer. As the network structure is based on dense connections between the layers, it is referred to as a densely connected network, as shown in Figure 2.
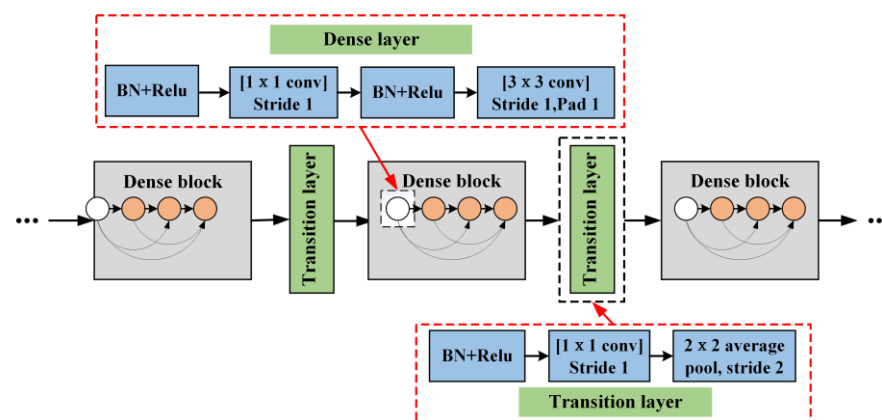

Figure 2. DenselNet model structure.

GrowthRate: The hyperparameter k is the network growth rate, which refers to the number of feature maps produced by each layer. An important feature of DenseNet is that k is very small for each layer, because each layer can be connected to all feature maps in the dense blocks that exist in it. The growth rate controls how much global information is added at each layer, and this information can be called anywhere in the network, which is the biggest difference between DenseNet and traditional neural networks.

DenseBlock: The network perceives the feature information locally through the first convolutional layer initially. Next, the data enters the dense block. A bottleneck layer structure is BN-Relu-Conv(1×1)-BN-Relu-Conv(3×3), which becomes Densenet-B. Each bottleneck layer generally contains a 1×1 convolution and a 3×3 convolution. The former serves to effectively reduce the number of feature maps, reduce the computational effort and achieve feature fusion for each channel, while the latter serves to perform feature extraction. A dense block can be composed of multiple bottleneck layers.

TransitionLayer: The network structure between two dense blocks is called the transition layer, and its structure is BN-Relu-Conv-Dropout-Pooling, which generally consistsa of 1 × 1 convolution and 2 × 2 pooling layer, the main role is to reduce the number of feature maps. θ denotes the compression factor, generally $\theta < 1$. If the forward thickening block generates n layers of feature maps, in order to compress the data, after the transition block, the number of feature maps as the input of the next thickening block becomes $\theta \times n$.

2.3. ECA-Net Module

The input time-frequency maps are learned by 2D-DenseNet dense network to obtain a large number of features, and the ECA-Net attention mechanism module is introduced to improve the classification efficiency of the fusion model, enhance the overall channel features, and improve the model performance [36].

The ECA-Net attention mechanism uses the global average pooling layer directly after the 1 × 1 convolutional layer, removing the fully connected layer. This module avoids dimensionality reduction and effectively captures cross-channel interactions. The module achieves good results with only a few parameters involved.

The ECA-Net module accomplishes cross-channel information interaction by one-dimensional convolution, and the size of the convolution kernel is adaptively varied by a function that allows more cross-channel interaction for layers with a larger number of channels, as shown in Figure 3.
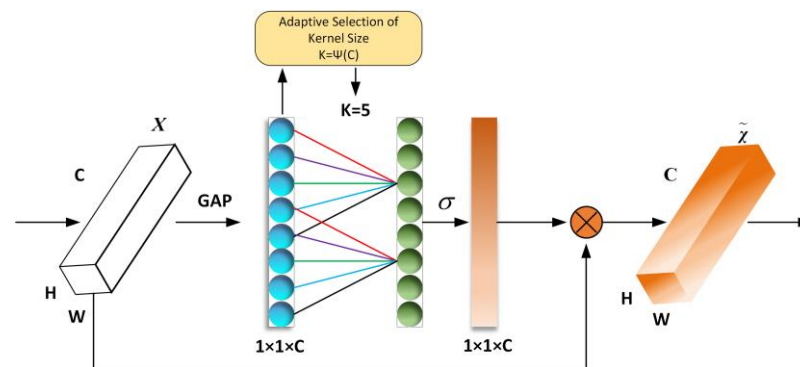

Figure 3. ECA-Net attention mechanism structure diagram.

The adaptive function is as follows (where $\gamma = 2$, $b = 1$):

$$k = \left| \frac{\log_2(c)}{\gamma} + \frac{b}{\gamma} \right| \tag{4}$$

The specific implementation process of the ECA-Net attention mechanism is as follows:

S1: Input feature maps with dimensions of H × W × C.

S2: Perform spatial feature compression on the input feature map. Implementation: in the spatial dimension, using global average pooling GAP to obtain the feature map of 1 × 1 × C.

S3: For the compressed feature map, channel feature learning is performed. Realization: through 1 × 1 convolution, learning the importance between different channels, at this time the output dimension is still 1 × 1 × C. The output dimension is still the same.

S4: Finally, the channel attention is combined with the feature map of channel attention 1 × 1 × C, the original input feature map H × W × C, perform channel-by-channel multiplication and finally outputs the feature map with channel attention.

According to the Efficient Channel Attention (ECA-Net) module shown in Figure 3. Considering the aggregated features obtained through the global average library (GAP), ECA-Net generates channel weights by performing a fast one-dimensional convolution of size k, where k is determined adaptively by mapping the channel dimension C.

2.4. LSTM-Attention Module

The addition of the LSTM-Attention module to the 1D-DenseNet densely connected network can effectively suppress gradient disappearance or explosion with good generalization ability.

The LSTM network is improved from the standard RNN. The LSTM effectively alleviates the long-term dependence problem of the standard RNN through its internal complex gate operation and the introduction of cellular states [37]. The unique feature of LSTM is that it introduces a memory cell and gate mechanism to solve the gradient disappearance and gradient explosion problems in the traditional RNN. LSTM is unique in that it introduces memory cell and gate mechanism to solve the gradient disappearance and gradient explosion problems in traditional RNNs, and enhances the ability to model long-term dependence.

The equations for the forgetting gate f_t, the input gate i_t, the output gate o_t, the cell state c_t and the output h_t are shown in the following equations:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{5}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{7}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{8}$$

$$h_t = o_t \otimes \tanh(c_t) \tag{9}$$

where: x_t refers to the input at the current moment; h_{t-1} refers to the output at the previous moment; W refers to the weight matrix; b refers to the bias; $\sigma(x) = 1/(1 + e^x)$ is the sigmoid activation function; $\otimes$ refers to the dot product operation.

Self-Attention is an improvement of the attention mechanism, which not only can quickly filter out the key information and reduce the attention to other irrelevant information, but also can reduce the dependence on external information and be better at capturing the internal relevance of the input data [38]. By introducing the self-attention mechanism, the neural network solves the model information overload problem while also improving the accuracy and robustness of the network [39].

The computation of Self-Attention is divided into two steps. Step 1: Calculate the attention weights between any vectors of the input sequence; Step 2: Calculate the weighted average of the input sequence based on the attention weights. The specific operation is shown in the following equation:

$$Q = XW^q \tag{10}$$

$$K = XW^k \tag{11}$$

$$V = XW^v \tag{12}$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{dim}}\right)V \tag{13}$$

where: Q, K and V are the query matrix, key matrix and value matrix, respectively, obtained by multiplying the input X with the corresponding weight matrices W^q, W^k, W^v, respectively; dim denotes the dimensionality of Q, K and V.

In summary, vibration signal feature extraction by LSTM-Attention module can better capture the key features in the time series signal, and then improve the prediction accuracy of the model. This is very helpful for some application scenarios with high accuracy requirements, such as fault diagnosis and prediction.

3. Improved Evaluation Model of DenseNetnetwork Based Ontwo-Channel Fusion
3.1. Model Overview

In this paper, a dual-channel fusion DenseNet network model (Frequency and Time-Frequency domain fusion DenseNet, FTF-DFD) is constructed based on a densely connected neural network, and its structure is shown in Figure 4: Since the original vibration data samples are insufficient and cannot be directly used in the standard network model to obtain better evaluation results, this paper performs data enhancement on the original data by overlapping sampling method.
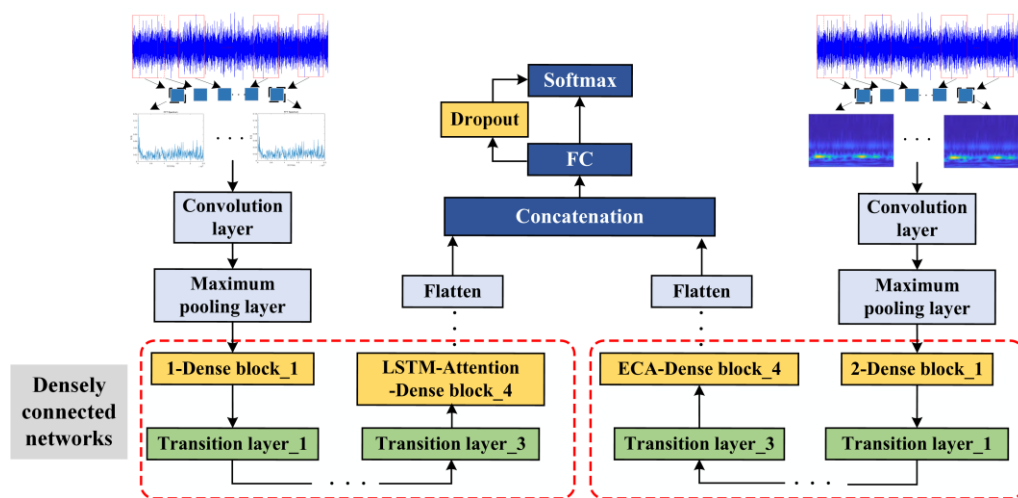
Figure 4. FTF-DenseNet model structure.

The model shown in Figure 4 is a two-channel DenseNet network structure consisting of the input layer, feature extraction module, and fault mode classification module. LSTM-Attention module for deep feature extraction; 1D-DenseNet model bottleneck layer structure is the same, both contain a combination of 1×1 and 1×3 size convolutional kernel, each dense block has a different number of bottleneck layers, this paper uses three groups of dense blocks, arranged according to the number of 3:2:1, adding a transition layer between every two dense blocks, which consists of a convolutional kernel size of 1×1 convolutional layer and a mean pooling layer of kernel size 1×2, which is used for downscaling and extracting global feature information, and finally, the output data of the network is linearized by squaring.

After the time domain signal is expanded by overlapping sampling, the wavelet time-frequency map is obtained by continuous wavelet transform, which is used as the input of the 2D-DenseNet model, and the ECA-Net attention mechanism is added after the last layer of dense blocks to effectively extract the model accuracy, strengthen the overall channel characteristics, and improve the model performance. The model structure of 2D-DenseNet is similar to that of 1D-DenseNet, which transforms the convolutional kernel from 1D to 2D, each bottleneck layer contains a combination of 1×1 and 3×3 size convolutional kernels, and the transition layer consists of a convolutional layer with 1×1 convolutional kernel size and an average pooling layer with 2×2 kernel size.

After the time-domain signal is overlapped sampled for data expansion, the wavelet time-frequency map is obtained by continuous wavelet transform, which is used as the input of the 2D-DenseNet model, and the ECA-Net attention mechanism is added after the last layer of dense blocks to effectively extract the model accuracy, strengthen the overall channel characteristics, and improve the model performance. 2D-DenseNet is similar to

the model structure of 1D-DenseNet The model structure of 2D-DenseNet is similar to that of 1D-DenseNet, which transforms the convolutional kernel from 1D to 2D, each bottleneck layer contains a combination of 1×1 and 3×3 size convolutional kernels, and the transition layer consists of a convolutional layer with 1×1 convolutional kernel size and an average pooling layer with 2×2 kernel size.

The output feature data of the 1D-DenseNet model and 2D-DenseNet model are stretched into feature vectors, and the splicing operation is performed by concat, and the fused feature information is input into the fully connected network layer and SoftMax classifier, and the probability distribution belonging to each category is output to achieve the fault classification recognition of bearings. The frequency domain and time-frequency domain fusion method proposed in this paper has a good fusion effect and improves the classification accuracy of the model. The specific parameters of the model are shown in Table 1.

Table 1. FTF-DenselNet network parameters.

Model Name	1D-DenseNet		2D-DenseNet	
	Structure Type	Convolution Kernel	Structure Type	Convolution Kernel
Input layer	One-dimensional FFT spectrum	-	Two-dimensional time-frequency diagram	-
Convolutional layer	Conv	1×7	Conv	7×7
Pooling layer	Maxpooling	1×3	Maxpooling	3×3
Dense block 1	$\begin{Bmatrix} BN - Relu - Conv \\ BN - Relu - Conv \end{Bmatrix} \times 1$	$\begin{Bmatrix} 1 \times 1 \\ 1 \times 3 \end{Bmatrix} \times 1$	$\begin{Bmatrix} BN - Relu - Conv \\ BN - Relu - Conv \end{Bmatrix} \times 1$	$\begin{Bmatrix} 1 \times 1 \\ 3 \times 3 \end{Bmatrix} \times 1$
Transition layer	BN-Relu-Conv-Pooling	$\begin{Bmatrix} 1 \times 1 \\ 1 \times 2 \end{Bmatrix} \times 1$	BN-Relu-Conv-Pooling	$\begin{Bmatrix} 1 \times 1 \\ 2 \times 2 \end{Bmatrix} \times 1$
Dense block 2	$\begin{Bmatrix} BN - Relu - Conv \\ BN - Relu - Conv \end{Bmatrix} \times 3$	$\begin{Bmatrix} 1 \times 1 \\ 1 \times 3 \end{Bmatrix} \times 3$	$\begin{Bmatrix} BN - Relu - Conv \\ BN - Relu - Conv \end{Bmatrix} \times 3$	$\begin{Bmatrix} 1 \times 1 \\ 3 \times 3 \end{Bmatrix} \times 3$
Transition layer	BN-Relu-Conv-Pooling	$\begin{Bmatrix} 1 \times 1 \\ 1 \times 2 \end{Bmatrix} \times 1$	BN-Relu-Conv-Pooling	$\begin{Bmatrix} 1 \times 1 \\ 2 \times 2 \end{Bmatrix} \times 1$
Dense block 3	$\begin{Bmatrix} BN - Relu - Conv \\ BN - Relu - Conv \end{Bmatrix} \times 1$	$\begin{Bmatrix} 1 \times 1 \\ 1 \times 3 \end{Bmatrix} \times 1$	$\begin{Bmatrix} BN - Relu - Conv \\ BN - Relu - Conv \end{Bmatrix} \times 1$	$\begin{Bmatrix} 1 \times 1 \\ 3 \times 3 \end{Bmatrix} \times 1$
ECA-Net	None	-	$\times 1$	-
LSTM-Attention	$\times 1$	-	None	-
Fully connected layer	FC	-	FC	-
Fully connected layer			FC	
Output layer			SoftMax	

3.2. Data Pre-Processing

3.2.1. Normalization Process

In order to data reduce the effect of distribution changes, improve the convergence speed of the model and diagnostic accuracy, the data are normalized and preprocessed, and the results are mapped to the [0–1] interval through a linear transformation, assuming that the sample data $X = \{x_1, x_2, \ldots, x_n\}$, whose transformation equation is as follows.

$$y_i = \frac{x_i - min\{x_j\}}{max\{x_j\} - min\{x_j\}} \tag{14}$$

where, y_i is the normalized result, x_i is the i-th sample data, $max\{x_j\}$ is the maximum value of the sample data, and $max\{x_j\}$ is the minimum value of the sample data.

3.2.2. Data Enhancement

In the field of data-driven deep learning, having large enough training samples is the key to improve the accuracy of the model and effectively reduce the overfitting of the model. In this paper, we propose to increase the training samples by using overlapping sampling with moving sliding windows, as shown in Figure 5, which can effectively increase the training samples while maintaining the periodicity and continuity of the one-dimensional time-series vibration signals and avoiding problems such as signal loss caused by isometric sampling and sampling.
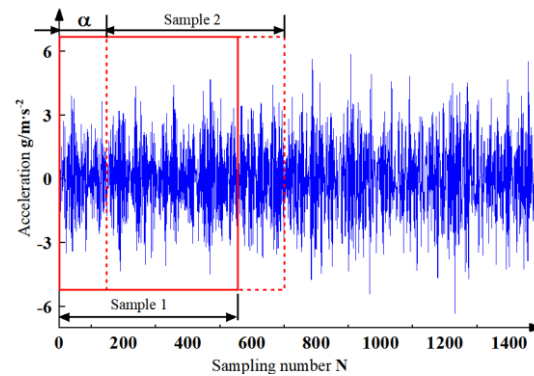

Figure 5. Schematic diagram of data enhancement.

From Figure 5, if the total length of data in a certain state is $L = 245{,}759$, the length of data for each sample is $l = 1024$. if no enhancement is applied, the number of samples A that can be segmented by the current vibration signal is:

$$A = [\frac{L}{l}] \tag{15}$$

Using the moving sliding window overlap method with offset $\alpha = 100$ for data sampling, the length of the overlapping part of the data is 924; the number of samples obtained from the current signal that can be split is (the maximum total number of samples that can be split per group of data is 2448)

$$B = [\frac{L - l}{\alpha} + 1] \tag{16}$$

Then the multiplier of sample expansion after using data enhancement γ is

$$\gamma = \frac{B}{A} = \frac{l(L - l + \alpha)}{\alpha L} \tag{17}$$

The expansion of the original data samples is achieved by overlapping sampling to avoid the loss of detailed features. Different offsets α, data lengths l, expansion multipliers γ and sample numbers B are set to achieve the performance detection of the model under different sample numbers and prove the practicality of the model.

3.2.3. Data Conversion

FFT is an efficient algorithm of Discrete Fourier Transform (DFT), called Fast Fourier Transform (FFT), which improves the DFT algorithm according to the odd, even, imaginary and real characteristics of DFT, and its basic principle is still Fourier Transform, which will not be discussed here. By calling the fft function in python, the frequency domain characteristics of the signal can be obtained, and then the frequency distribution of different signals can be analyzed.

The signal after data enhancement is analyzed in the time-frequency domain, where the time-frequency map of the original signal is obtained by continuous wavelet variation

(CWT), which can clearly and accurately represent the time-frequency distribution of the vibration.

The continuous wavelet transform provides the best resolution results for non-periodic signals without leakage effects. Continuous wavelet transform $CWT(\alpha, \tau)$ can be calculated by the following equation:

$$CWT(\alpha, \tau) = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{+\infty} s(u)\psi(t) \times \left(\frac{u - \tau}{\alpha}\right) du \quad (18)$$

where, α is the scale, $s(u)$ is the original signal, τ is the translation, and $\psi(t)$ is the mother wavelet.

The continuous wavelet variation (Continuous Wavelet Transform, CWT) chooses Complex Morlet Wavelet (Cmor) as the wavelet basis, and the Cmor wavelet basis function is obtained by improving the Morlet wavelet basis function. It is a complex wavelet basis function with dual resolution properties in both frequency and time domains, which is widely used in the field of signal processing and wavelet analysis. The Cmor wavelet basis function has a similar shape to the Gaussian function but has better frequency localization properties in the frequency domain. It has better time-frequency localization properties in both time and frequency domains, and is suitable for processing non-stationary signals and analyzing transient phenomena in signals.

3.3. Model Training

The optimization algorithm used for the TADAT-based rolling bearing fault diagnosis model is chosen as Adam. The Adam algorithm adaptively adjusts the learning rate of each parameter, and different learning rates can be used for different parameters, thus making the training more efficient and stable.

The Adam algorithm dynamically corrects the training steps of each parameter using first-order moment estimation and second-order moment estimation of the gradient with the following update rules:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + \varepsilon}} m_t \quad (19)$$

where, θ_{t+1}, θ_t denotes the model parameters at step $t + 1$ and step t, respectively; η is the learning rate; v_t denotes the value of the unbiased second-order moment estimate; m_t denotes the value of the unbiased first-order moment estimate; and ε is a very small positive number, generally taken as $\varepsilon = 10^{-8}$, preventing the denominator from being 0.

The TADAT-based rolling bearing fault diagnosis model diagnoses work conditions based on features, which belongs to the classification problem in supervised learning, so cross entropy is chosen as the loss function and optimized.

Cross entropy is mainly used to calculate the distance between the correct probability of labeling and the probability of prediction, and the smaller the value of cross entropy, the closer the prediction result is to the actual result, and the formula is as follows:

$$loss = -\sum_{\theta} p(\theta) lg q(\theta) \quad (20)$$

where, θ denotes the individual learning parameters; $p(\theta)$ denotes the probability of correct labeling; and $q(\theta)$ is the prediction probability.

For two probability distributions $p(\theta)$ and $q(\theta)$, define the K-L scatter of $p(\theta)$ and $q(\theta)$ as follows:

$$KLD = \sum_{\theta} p(\theta) lg \frac{p(\theta)}{q(\theta)} \quad (21)$$

When calculating the cross-entropy loss using KL scatter, the true labels need to be transformed into probability distributions, usually using methods such as one-hot encoding or smoothed labels. In this paper, a smoothed target label is used instead of the traditional one-hot encoded label, thus reducing the impact of the noise and uncertainty of the label on the model and obtaining a loss function *ce_loss*.

In *ce_loss* its L2 regularization penalty term is introduced to penalize the size of the model parameters to prevent overfitting. The strength of the regularization penalty can be controlled by adjusting the value of alpha to establish the regularized loss function as follows:

$$impro\_loss = ce\_loss + alpha * \sum_{\theta} \theta^2 \tag{22}$$

where, θ denotes each learning parameter; alpha is the regularization parameter.

After pre-processing, the data set is divided into training set, validation set and test set in the ratio of 7:2:1, the model uses the parameters with the highest training and validation accuracy as the final parameters, the optimizer uses the Adam optimizer with fast and stable convergence, the loss function is the regularized loss function, the initial learning rate is 0.01, and the learning rate decays by half for every 10 iterations. Normalization batch normalization is used to accelerate the convergence speed of the neural network, Dropout operation is added to prevent overfitting, and the number of training iterations of the model is set to 100; finally, the Softmax function is used to classify the target and output the probability distribution of each category; as shown in Table 2:

Table 2. Model parameters.

Parameter Category	Parameter Setting
Training set:Validation set:Test set	5600:1600:800
Optimizer	Adam
Number of training sessions	100
Learning Rate	0.02
Batch Size	64
alpha	0.05
smothing	0.1

In order to obtain the appropriate Batch Size parameter for the model, the mid-load experiment was used as the basis for comparison by setting different Batch Size parameter values, mainly setting three different sets of values of 32, 64, and 128, as shown in Figure 6, and it was found that the accuracy of the model improved the fastest when the Batch Size was set to 64, reaching 94% accuracy after 20 iterations, and After 60 iterations, the accuracy rate is stabilized at 100%, while the accuracy rate of the model test fluctuates more when the Batch Size is set to 32 and 128. It was concluded that the best iteration of model accuracy was achieved when the Batch Size was set to 64.
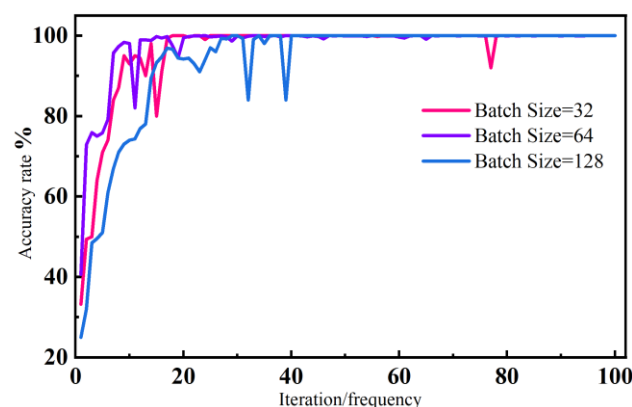

Figure 6. Comparison of Batch Size under medium load condition.

4. Experimental Verification

4.1. Environment Description

A non-equilibrium bearing load test stand was developed and built to further study the monitoring function of this technology in the process of double bearing operation, as

shown in Figure 7. The test stand mainly consists of a motor, a precision spindle, a rolling bearing and an acceleration sensor, and the maximum speed read by the electric spindle is 10,000 r/min. The mechanical spindle is connected to the electric spindle through a flexible coupling, and the motor operation is controlled by a servo control system.
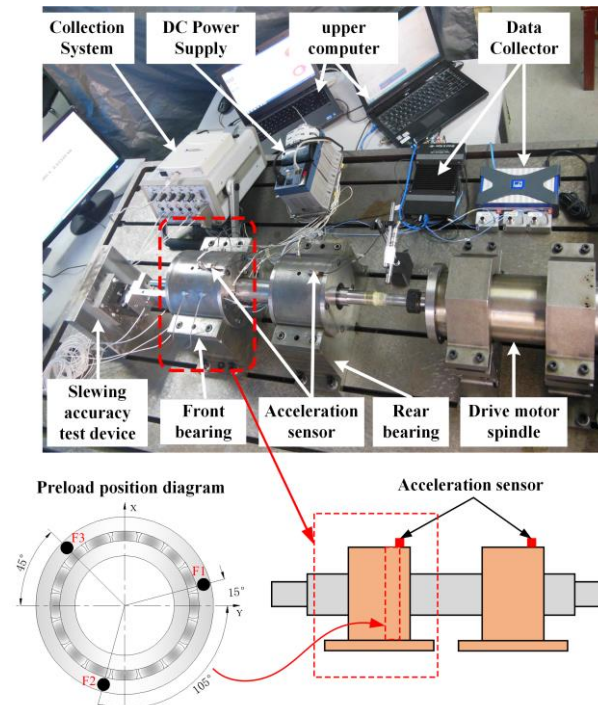

Figure 7. Structure of non-uniform preload test rig.

The test rig used four NSK7014C angular contact ball bearings, where F1, F2 and F3 were loaded 120° on the bearings, respectively, and the bearing bias operating condition was determined by setting different sizes of preload; the bearings were mounted back-to-back with a fixed speed of 4000 r/min, a sampling frequency of 8192 Hz and a sampling length of 512. Table 3 shows the bearing parameters.

Table 3. NSK7014C angular contact ball bearing parameters table.

Inner Ring Diameter/mm	Parameter Setting	Thickness/mm	Dynamic Load/KN	Static Load/mm
70	10	20	47	43

Software environment: The training and testing environment of this paper is 14 cores, 16 G memory, processor: 12th Gen Intel Core i7-12700H processor; programming environment Pytorch1.7.1.

The bearing non-uniform load test bench is designed to distinguish the operating condition of the bearing under unbalanced operation, so that the bearing failure caused by factors such as assembly or machining can be detected in time. Due to the limited conditions in the laboratory, the currently built test bench can only be used to verify the effectiveness and accuracy of the condition monitoring method, and cannot simulate the corresponding bearing failure state for verification experiments.

4.2. Example Analysis

4.2.1. Data Conversion

A total of twelve sets of data are collected through the bearing load non-uniform operation fault simulation test bench, including F1, F2, F3 loading and data under even

load conditions, which are mainly divided into four conditions of light load, medium load, heavy load and even load, and there are 12 types of bearing vibration data. There are four sets of experiments in total, the first three sets of experiments with 1400 samples in the training set, 400 samples in the test set and 200 samples in the validation set; the last set of experiments with all types of data input; the specific experimental data are shown in Table 4.

Table 4. Experimental dataset.

Experiment Name	Signal Type	Training Set	Validation Set	Test Set
First group of experiments (Light load comparison)	F1 (C2) = 400 N	1400	400	200
	F2 (C2) = 400 N	1400	400	200
	F3 (C2) = 400 N	1400	400	200
	$F_{1.2.3}$(C1) = 200 N	1400	400	200
Second group of experiments (mid-load comparison)	F1 (C4) = 800 N	1400	400	200
	F2 (C4) = 800 N	1400	400	200
	F3 (C4) = 800 N	1400	400	200
	$F_{1.2.3}$(C3) = 400 N	1400	400	200
Third group of experiments (Heavy load comparison)	F1 (C6) = 1200 N	1400	400	200
	F2 (C6) = 1200 N	1400	400	200
	F3 (C6) = 1200 N	1400	400	200
	$F_{1.2.3}$(C5) = 600 N	1400	400	200
Fourth group of experiments	Enter 12 types of data	700–12	200–12	100–12

The vibration signals at F1 positions C2, C4, C6 and $F_{1.2.3}$ (C1) working conditions were subjected to signal analysis, and 1024 data points were taken as one sample and subjected to FFT transform with wavelet transform. The analysis in Figure 8 shows that the spectrograms of the four working conditions data at 3044 Hz have the maximum amplitude variation, which reflects the main frequency components of the signal in the frequency domain. As the load at the F1 position of the bearing gradually increases, the vibration characteristics of the bearing will be more intense, and more frequency components and amplitudes will be generated, and more noise and spurious frequencies may appear in the FFT analysis. Since the frequency and amplitude of the bearing vibration will change with the load, the amplitude of the main frequency components in the FFT spectrum is relatively small when the load is higher. Under the average load, the frequency and amplitude of bearing vibration are relatively stable, so the amplitude of the main frequency components in the FFT spectrum is relatively large. The spectrum analysis shows that the spectrum gradually increases with C6, C4, C2, and $F_{1.2.3}$ (C1). The signal analysis of the time-frequency diagram also proves this point. In the energy distribution in the frequency range of 2 to 3 kHz, the $F_{1.2.3}$ (C1) condition has the largest energy, while the C6 condition has the least energy.

4.2.2. Model Testing

The experiments were conducted with the input sample length of 1024, Batch Size of 64, training iterations of 100, learning rate of 0.002, and optimizer choice of Adam. The method was initially tested on the Case Western Reserve University bearing dataset.

The test selected the bearing data at the drive end, sampling frequency of 48 kHz, and load of 0 hp, the fault form contains outer ring fault, inner ring fault and rolling body fault three kinds of fault parts as shown in Figure 9, the fault type is specifically divided into 7 mils, 14 mils and 21 mils three kinds of fault diameter, plus the normal state, a total of ten kinds of bearing state data. The specific sample composition information is shown in Table 5. A total of 70% of the samples are selected as the training set, 20% as the validation set, and 10% as the test set.
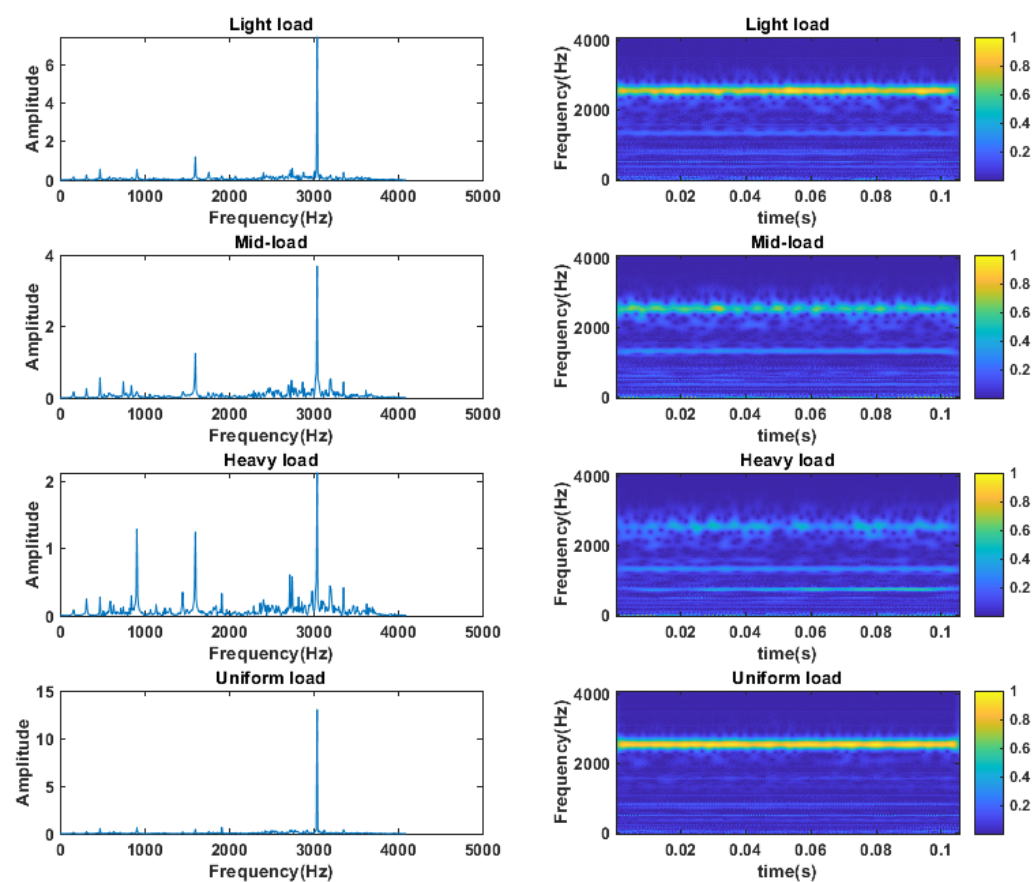
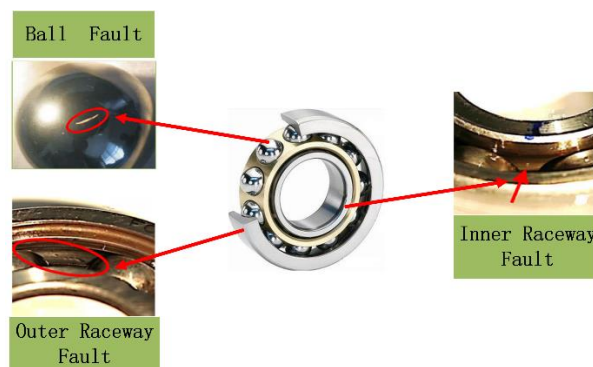**Figure 8.** Signal analysis of three operating conditions at F1 position with equal load.

Figure 9. Bearing fault form distribution chart.

Table 5. Sample composition information.

Sample Type	Sample Length	Sample Size	Type Tags
Ball Fault (7 mils)	864	400	B007
Ball Fault (14 mils)	864	400	B014
Ball Fault (21 mils)	864	400	B021
Inner Raceway Fault (7 mils)	864	400	IR007
Inner Raceway Fault (14 mils)	864	400	IR014
Inner Raceway Fault (21 mils)	864	400	IR021
Outer Raceway Fault (7 mils)	864	400	OR007
Outer Raceway Fault (14 mils)	864	400	OR014
Outer Raceway Fault (21 mils)	864	400	OR021
Normal	864	400	normal

The time domain of the bearing vibration signal contains a large number of high and low-frequency components, which have different sensitivities for the diagnosis of bearing faults. Therefore, converting the time domain signal to the frequency domain signal for analysis can better capture the characteristics of bearing faults as shown in Figure 10. The fault signal with fault type 7 mils at 0 hp is taken for spectral analysis with the normal signal, and 1024 data points are taken as a sample for FFT transformation, and four states of the bearing, such as normal state, rolling element fault, inner ring fault, and outer ring fault, can be found. The difference of amplitude in the high-frequency band is large. Figure 11 shows that this mechanical vibration signal mainly contains energy in the frequency range of 0~5 kHz, in which the inner ring fault has obvious energy intensity transformation in both low and high-frequency bands, and its distribution is very dense because it is a different type of fault, which can effectively identify the frequency components and time domain features of the signal and provide useful information for applications such as signal feature extraction, classification, and diagnosis.

The ten bearing condition data in Table 5 were input into the model for condition diagnosis of the bearings. From the model output accuracy versus loss function curve in Figure 12, it can be seen that the model can reach 98% accuracy after 20 iterations, and after 50 iterations, the model can finally reach 100% accuracy.

In order to clearly represent the extraction ability of features in the model, we use the t-SNE technique to downscale and visualize the features in the input and output layers to indirectly represent the extraction ability of features in the model, where different colors and numbers indicate different fault categories and horizontal and vertical coordinates indicate different dimensions.
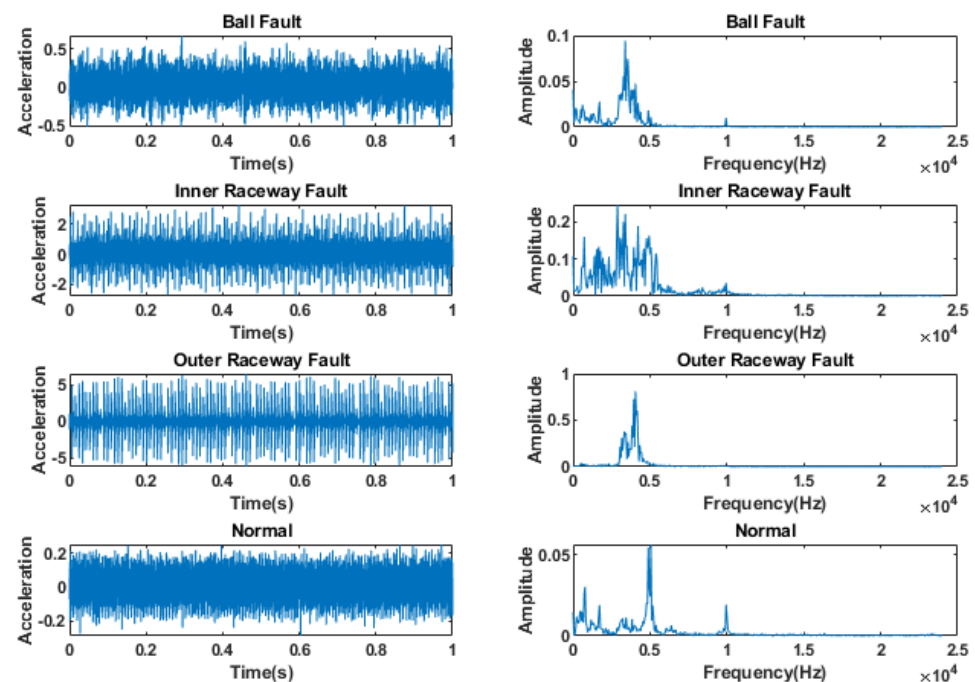

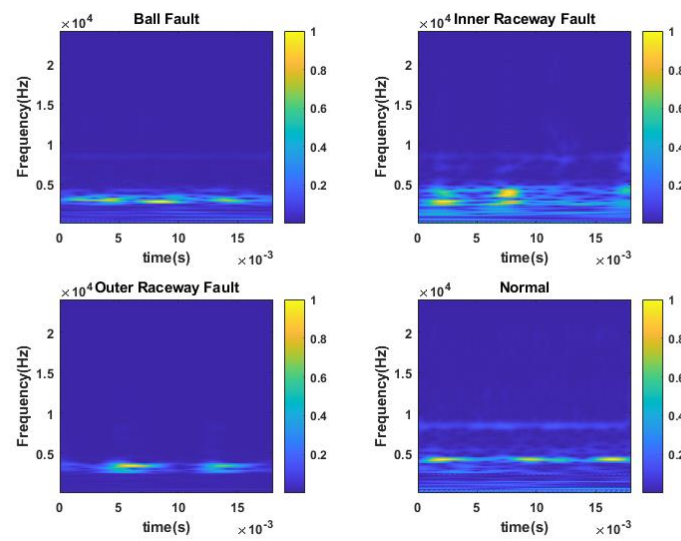**Figure 10.** Time domain waveforms and frequency spectrum of CWRU bearing data.

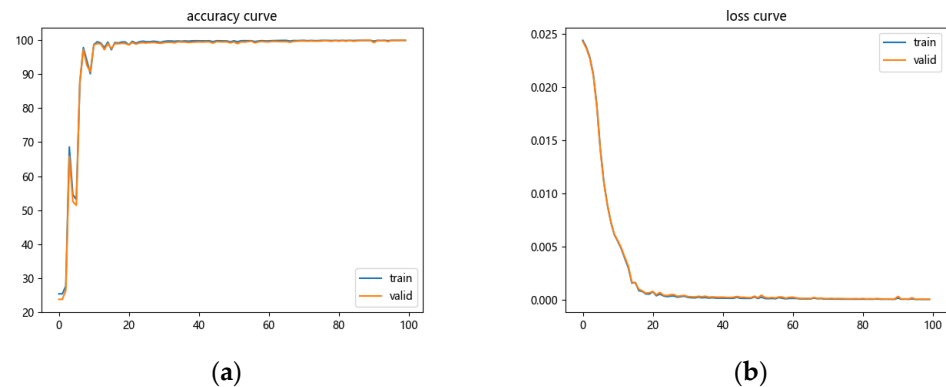**Figure 11.** Wavelet time−frequency diagram of CWRU bearing data.

Figure 12. (**a**) Accuracy curve; (**b**) Loss curve.

As shown in Figure 13, the input layer is disorganized and various features are mixed together. After the three-stage densely connected network and the two-domain feature fusion, the extraction of features by the model is basically completed, and the separation and convergence of all kinds of features are basically completed, and the visual classification results of the output layer show that the model has a good classification effect.
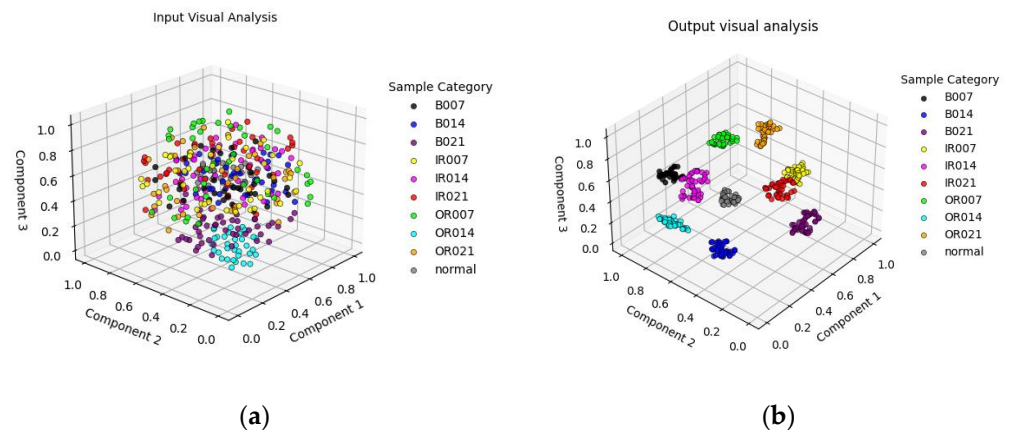

Figure 13. (**a**) Input feature visualization; (**b**) Output feature visualization.

4.3. Comparison Experiments

At the position of bearing F1, three working conditions of light load (OC_1), medium load (OC_2) and heavy load (OC_3) were measured, and three data sets M1, M2, and M3 were established to contain the information of the above three working conditions, and 300 samples were taken for each working condition. In the network model, the Batch Size is set to 64, the number of training iterations is set to 100, the optimizer is Adam, the initial learning rate is 0.01, and the learning rate decays by half every 10 iterations, and the loss function is selected as *impro_loss*.

The diagnostic res CNN: The model structure is the input layer, Conv layer, MaxPool layer, ReLu activation function, BN layer, flat layer, Dropout layer, fully connected layer, and SoftMax output layer. The input data is a two-dimensional time-frequency map, and the middle layer is a two-layer convolutional pooling network, which is stretched by the flat layer and then passed through the fully-connected layer and the SoftMax output layer to achieve the classification of work conditions.

Improved-FTF-CNN: The model adopts the fusion of frequency domain and time-frequency domain, in which the 1D and 2D models, the three-level dense connection network is used, and the ratio of dense blocks are 3:2:1, relying on the feature fusion through concat, and finally the classification of working conditions through the SoftMax output layer.

DenseNet: The input layer of the model is fed with a two-dimensional time-frequency map, and the intermediate structure uses three groups of dense blocks, according to the number 3, 2, and 1. The output layer uses the SoftMax output layer for the classification of working conditions.

Improved-FTF-DenseNet: the base structure of the model is the same as the improved-FTF-CNN network structure, and the intermediate feature extraction structure replaces the CNN module in it with the DenseNet network structure, and the rest of the network results remain unchanged.

The diagnostic results of the above methods are shown in Figure 14, and the average value of five experiments is taken as the model evaluation result, and the average value of six groups of experiments is taken for model performance evaluation, which is shown in Table 6. the network structure of the CNN model is relatively simple and cannot extract accurate features, and the training time is the shortest, with an average accuracy rate of 87.43%; the improved-FTF-CNN model, compared with the simple CNN network, has an accuracy rate has significantly improved, and is 4.83% higher than the CNN model; the DenseNet model can improve the complexity of the model due to the dense connection structure, and after adjusting its parameters, the final accuracy can reach 92.57%; the improved-FTF-DenseNet can reach a final average accuracy of 93.88% through the model of dual-channel fusion, which is lower than the method of this paper by 3.18%.
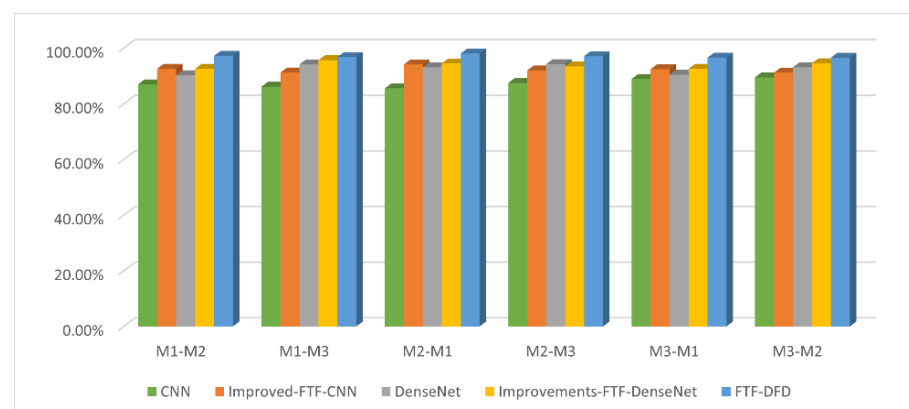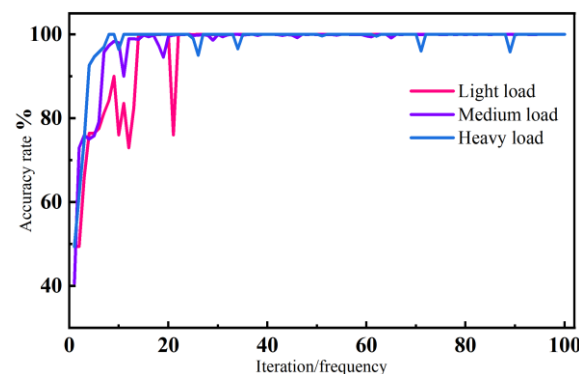

Figure 14. Model comparison chart.

Table 6. Diagnostic results of different methods.

Network Structure	Accuracy	Training Time/min
CNN	87.43%	0.5 min
Improved-FTF-CNN	92.26%	1.2 min
DenseNet	92.57%	3.5 min
Improvements-FTF-DenseNet	93.88%	4.2 min
FTF-DFD	97.06%	3.8 min

4.4. Uneven Bearing Load Experiment

The results of three sets of experiments of the FTF-DFD fault diagnosis model proposed in this paper are shown in Figure 15, which are the accuracy curves of the unbalanced experiments. The accuracy rates of the three different experimental conditions on the test set after 18 iterations of training all reach more than 98%, among which the accuracy curves of the light load experiments have a large abrupt change in the rising stage and the accuracy rate is not as fast as The accuracy curves of the medium-load and heavy-load experiments are not as fast as those of the medium-load and heavy-load experiments. The accuracy transformation curve is flatter under the medium-load experimental condition, and the accuracy can reach 100% on the test set after 50 iterations of training. The accuracy of all three unbalanced experiments reached 100% after 70 iterations of training, and it can be seen through the three sets of experiments that the model is more adaptable under medium-load and heavy-load working conditions.

Figure 15. Comparison of three working conditions.

The deeper the layers of the neural network model, the better the extraction effect for signal features. In this paper, the complexity of the model is increased by the densely connected network, and the learning ability of the model is enhanced to extract the one- and two-dimensional features of the original signal, and the model is made to obtain more feature information through the mode of two-channel fusion, so as to improve the accuracy of the model for monitoring the load inhomogeneous state.

The comparison results of the first three sets of experiments are shown in Figure 16. Through the confusion matrix and classification result graphs of the three sets of experiments, it can be seen that the FTF-DFD model proposed in this paper achieves the recognition of four types of position information, F1, F2, F3 and $F_{1.2.3}$, respectively, under light load, medium load and heavy load conditions, and all of them achieve 100% recognition accuracy.

(**a**) Light load experiment

(**b**) Mid-load experiments

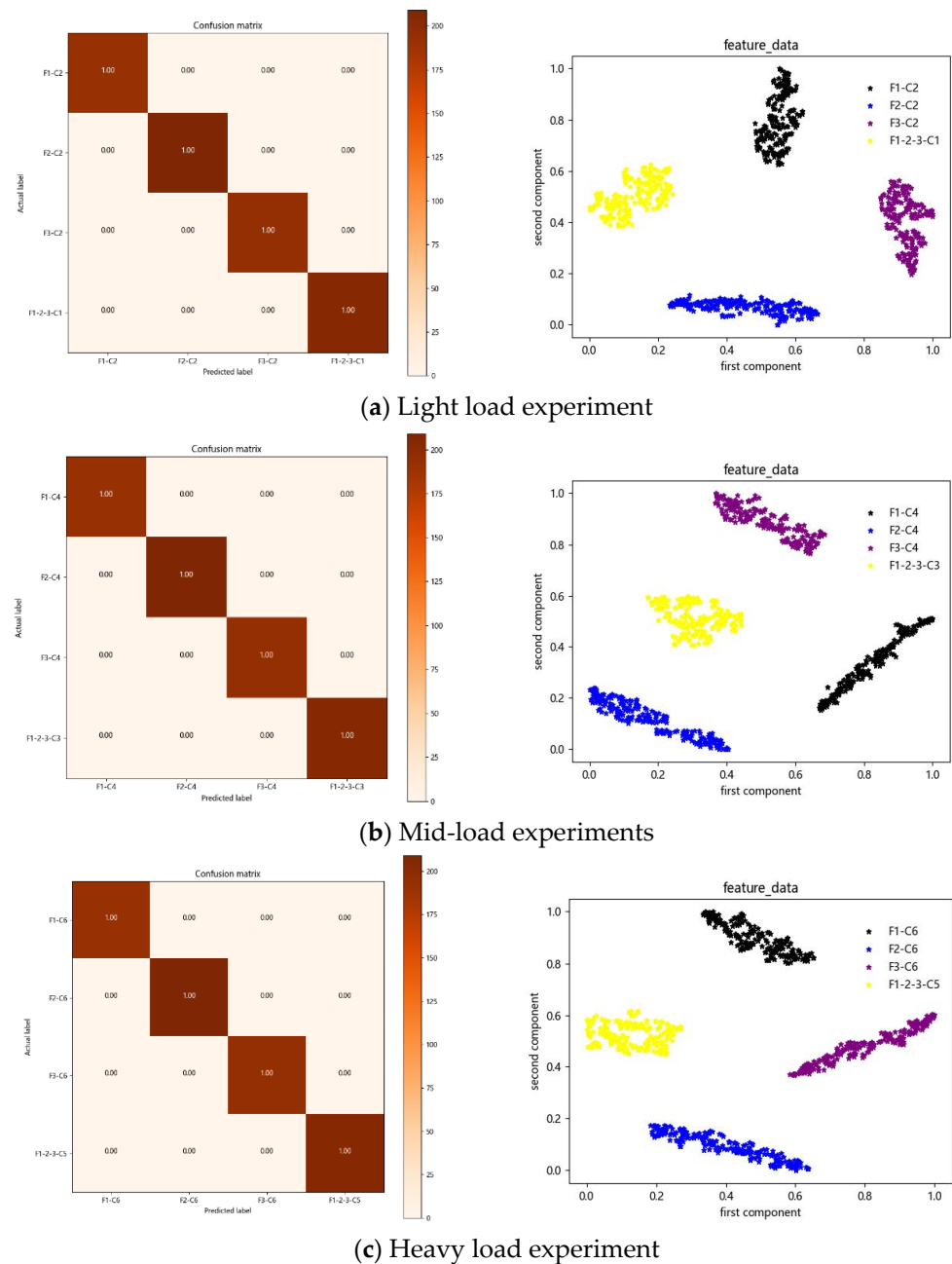(**c**) Heavy load experiment

Figure 16. Comparison chart of the first three groups of experiments. Confusion matrix and visualization of classification results.

The fourth set of experiments took the 12 sets of work condition data involved in this paper and input them into the evaluation model after the data sample expansion of overlapping sampling, and the final output was divided into 12 clusters by t-SNE visualization. According to the input features in Figure 17a, it can be seen that compared with the Western Reserve University 10 classification task features are completely mixed, the original input data of this experiment are mainly divided into two parts, F1 and F3 positions at all the working condition data are mixed together, and F2 is mixed with all the working condition data at position $F_{1,2,3}$, indicating that the original feature distributions of these data are closer and cannot be easily distinguished from each other. The classification result graph of the output of Figure 17b shows that the experimental data of all kinds of working conditions of the model are improved from the chaotic state to the aggregated state, and the classification task of 12 working conditions is completed effectively, and all the working condition information is completely distinguished. The experiments are conducted

by expanding different sample sizes, and it can be seen that the feature distributions in the two groups of working conditions, F2 (C2) and $F_{1.2.3}$ (C5), are always closer, indicating the similarity of the feature components in these two groups of working conditions. This proves that the algorithm in this paper can effectively realize the working condition recognition under the uneven bearing load.
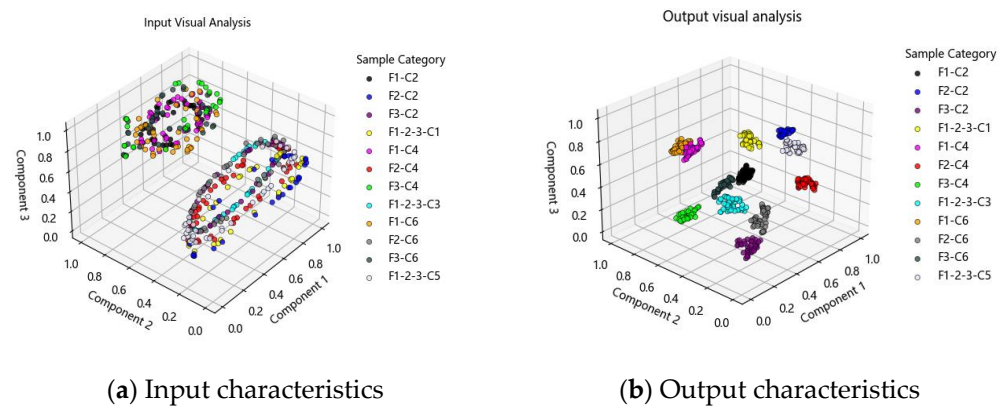

(**a**) Input characteristics

(**b**) Output characteristics

Figure 17. Visual classification result chart.

5. Conclusions

In order to improve the condition monitoring performance of bearings with non-uniform loads, we propose a fault diagnosis method using the FTF-DFD model to identify the operating condition of spindles more accurately. First, the sample expansion of the original data is performed, and then the frequency domain and time-frequency domain conversion are performed. Then, the FTF-DFD model is constructed for the extraction of dual-domain feature information, and the overall iterative performance of the model is improved by the *Adam* dynamic adjustment strategy and the improved *ce_loss* loss function. Finally, the validity of the model was tested by the Case Western Reserve University data set, and an experimental bench for bearing non-uniform load operation was designed and built for validation, which was compared with the other four methods, and the following conclusions were drawn:

- Using the dual-channel model to extract the frequency domain features and time-frequency domain features of the original signal can reflect the vibration characteristics of the bearing more comprehensively and accurately, thus improving the accuracy of fault diagnosis.
- The condition monitoring model of the spindle bearing of FTF-DFD is established. The model has strong generalization performance, and the bearing condition under fault condition and variable load condition can be identified, and the condition detection rate is extremely high, reaching up to 100%.
- The overall iterative performance of the model is greatly improved, and the training time is reduced by using the Adam dynamic adjustment strategy in conjunction with the improved *ce_loss* loss function.
- This paper only validates the performance of the method for identifying non-uniform loads on bearings. In the future, the FTF-DFD model will be applied to other components of the spindle system, and the model will be migrated to other fields to complete further validation.

Author Contributions: Software, W.Z.; Formal analysis, D.L.; Resources, L.W.; Data curation, Y.L.; Writing—review & editing, Y.Z.; Supervision, L.K. All authors have read and agreed to the published version of the manuscript.

References

1. Jing, L.; Zhao, M.; Li, P.; Xu, X. A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. *Measurement* **2017**, *111*, 1–10. [CrossRef]
2. Jeschke, S.; Brecher, C.; Meisen, T.; Özdemir, D.; Eschert, T. *Industrial Internet of Things and Cyber Manufacturing Systems*; Springer International Publishing: Cham, Switzerland, 2017; pp. 3–19.
3. Yin, S.; Li, X.; Gao, H.; Kaynak, O. Data-based techniques focused on modern industry: An overview. *IEEE Trans. Ind. Electron.* **2014**, *62*, 657–667. [CrossRef]
4. Chen, Z. Research on Intelligent Diagnosis Method of Mechanical Equipment Based on Deep Migration Learning. Ph.D. Thesis, South China University of Technology, Guangzhou, China, 2020.
5. Wang, Y.J. Research on Rolling Bearing Vibration Signal Feature Extraction and State Evaluation Method. Ph.D. Thesis, Harbin Institute of Technology, Harbin, China, 2015.
6. Deng, M.Y.; Li, C.Z.; Yang, H. Research on bearing fault diagnosis based on frequency domain feature variational self-encoder. *Comput. Meas. Control* **2023**, *31*, 70–75.
7. Jiao, W.D.; Ding, X.M.; Yan, T.Y.; Yan, Y.Y. Research on fault diagnosis method based on frequency domain feature waveform pattern matching. *J. East China Jiaotong Univ.* **2021**, *38*, 73–81.
8. Liu, F.; Wu, X.; Pan, N.; Zhou, J. Application of improved time-domain blind deconvolution algorithm for bearing fault diagnosis. *Mech. Strength* **2016**, *38*, 207–214.
9. Li, W.; Ma, J.; Yu, F. A rolling bearing fault diagnosis method based on frequency domain sparse classification. *Bearing* **2016**, *33*, 58–61.
10. Wang, S.L.; Chen, J.; From, F. Application of time-frequency-based band entropy method in rolling bearing fault identification. *Vib. Shock* **2012**, *31*, 29–33.
11. Li, J.Q.; Zheng, J.D.; Pan, H.Y.; Tong, J.; Feng, K.; Ni, Q. A two-dimensional time-frequency multi-scale entropy method for rolling bearing fault diagnosis. *Mech. Sci. Technol.* **2023**, *111*, 1–10.
12. Zhou, J.; Zhu, J.W. Research on machine learning classification problems and algorithms. *Software* **2019**, *40*, 205–208.
13. Zhang, R.; Wang, Y.B. Research on machine learning and its algorithms and development. *J. Commun. Univ. China* **2016**, *23*, 10–18.
14. Pei, S. Research on Classification Algorithm Based on Machine Learning. Master's Thesis, North Central University, Taiyuan, China, 2016.
15. Wu, X.M.; Wu, Y.Y.; Wang, X.; Li, C.F.; Zhang, F.H. Application of machine learning in bearing fault diagnosis. *Equip. Manuf. Technol.* **2022**, *327*, 118–126.
16. Wang, J. Research and Application of Text Classification Algorithm Based on Machine Learning. Master's Thesis, University of Electronic Science and Technology, Chengdu, China, 2015.
17. Zhang, X.Y.; Luan, Z.Q.; Liu, X.L. A review of rolling bearing fault diagnosis research based on deep learning. *Equip. Manag. Maint.* **2017**, *414*, 130–133.
18. Lu, X.; Zhang, C.; Gao, J.; Xu, Y.; Shao, X. Bearing fault diagnosis algorithm based on convolutional neural network and CatBoost. *Mechatron. Eng.* **2023**, *40*, 1–10.
19. Janssens, O.; Slavkovikj, V.; Vervisch, B.; Stockman, K.; Loccufier, M.; Verstockt, S.; Van de Walle, R.; Van Hoecke, S. Convolutional neural network based fault detection for rotating machinery. *J. Sound Vib.* **2016**, *377*, 331–345. [CrossRef]
20. Gu, X.; Tang, X.H.; Lu, J.G.; Li, S.W. Adaptive Fault Diagnosis Method for Rolling Bearings Based on I-DCNN-LSTM. *Mach. Tool Hydraul.* **2020**, *48*, 107–113.
21. Gong, W.F.; Chen, H.; Zhang, Z.H.; Zhang, M.L.; Guan, C.; Wang, X. Intelligent fault diagnosis forrolling bearing based on improved convolutional neural network. *J. Vib. Eng.* **2020**, *33*, 400–413.
22. Chen, K.; Huang, M.; Li, Y. Bearing fault diagnosis method based on CNN-LSTM and attention mechanism. *J. Beijing Univ. Inf. Sci. Technol.* **2022**, *37*, 26–31.
23. Yang, X.; Sun, S.; Wang, G.; Shi, N.; Xie, Y. Bearing fault diagnosis method based on ECA_ResNet. *Bearing* **2023**, 1–8.
24. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
25. Shi, C.; Ren, Y.; Tang, H.; Mupfukirei, L.R. A fault diagnosis method for electro-hydraulic directional valve based on intrinsic mode functions and weighted densely connected convolutional networks. *Meas. Sci. Technol.* **2021**, *32*, 084015. [CrossRef]
26. Wang, Y. Research on The Fault Diagnosis Method of Fusing Multiple Sensors. Master's Thesis, Beijing Jiaotong University, Beijing, China, 2019.
27. Niu, R.X.; Ding, H.; Shi, R.; Meng, X.L. Improved densely connected convolutional networks for rolling bearing fault diagnosis. *Vib. Shock* **2022**, *41*, 252–258.

28. Wang, Q.R.; Wang, Y.; Zhu, C.F.; Zhou, Y.T. Fault diagnosis of rolling bearings with dual-channel cross-dense connection. *Mech. Sci. Technol.* **2023**, 1–9.

29. Wang, K.; Liu, X.; Yang, J.Q.; Dong, Z.S. Fault diagnosis of rolling bearings with variable operating conditions based on improved DenseNet model. *Comb. Mach. Tools Autom. Mach. Technol.* **2022**, *580*, 78–81.

30. Hubel, D.H.; Wiesel, T.N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **1962**, *160*, 106. [CrossRef] [PubMed]

31. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Handwritten digit recognition with a back-propagation net-work. *Adv. Neural Inf. Process. Syst.* **1989**, *2*, 1–9.

32. Li, H.Y.; Su, T.B. A hand-drawn sketch recognition method based on Bayesian network and convolutional neural network. *J. Southwest Norm. Univ.* **2019**, *44*, 96–102.

33. Wu, D.H.; Ren, G.Q.; Wang, H.G.; Zhang, Y. A review of mechanical fault diagnosis methods based on convolutional neural networks. *Mech. Strength* **2020**, *42*, 1024–1032.

34. Wang, T.Y.; Gong, L.M.; Wang, P.; Qiao, H.U.; Ren, D. KD-DenseNet-based fault diagnosis model for rotating machinery. *Vib. Shock* **2020**, *39*, 39–45.

35. 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.

36. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.

37. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef]

38. Zhao, H.; Jia, J.; Koltun, V. Exploring self-attention for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 13–19 June 2020; pp. 10076–10085.

39. Zhang, H.; Zhang, Q.; Shao, S.Y.; Niu, T.; Yang, X. Attention-based LSTM network for rotary machine remaining useful life prediction. *IEEE Access* **2020**, *8*, 188–199. [CrossRef]