machines

MDPI

Article

Fault Diagnosis of Motor Vibration Signals by Fusion of Spatiotemporal Features

Lijing Wang [1,2,3], Chunda Zhang [1], Juan Zhu [3] and Fengxia Xu [1,3,*]

[1] School of Mechanical and Electrical Engineering, Qiqihar University, Qiqihar 161000, China; 02926@qqhru.edu.cn (L.W.); 2019911186@qqhru.edu.cn (C.Z.)
[2] School of Electrical and Electronic Engineering, Harbin University of Science and Technology, Harbin 150080, China
[3] Collaborative Innovation Center of Intelligent Manufacturing Equipment Industrialization of Heilongjiang Province, Qiqihar University, Qiqihar 161006, China; 02976@qqhru.edu.cn
* Correspondence: 01541@qqhru.edu.cn

Abstract: This paper constructs a spatiotemporal feature fusion network (STNet) to enhance the influence of spatiotemporal features of signals on the diagnostic performance during motor fault diagnosis. The STNet consists of the spatial feature processing capability of convolutional neural networks (CNN) and the temporal feature processing capability of recurrent neural networks (RNN). It is used for fault diagnosis of motor vibration signals. The network uses dual-stream branching to extract the fault features of motor vibration signals by a convolutional neural network and gated recurrent unit (GRU) simultaneously. The features are also enhanced by using the attention mechanism. Then, the temporal and spatial features are fused and input into the softmax function for fault discrimination. After that, the fault diagnosis of motor vibration signals is completed. In addition, several sets of experimental evaluations are conducted. The experimental results show that the vibration signal processing method combined with spatiotemporal features can effectively improve the recognition accuracy of motor faults.

Keywords: spatiotemporal feature fusion; convolutional neural network; gated recurrent unit; attention mechanism; fault diagnosis

check for updates

Citation: Wang, L.; Zhang, C.; Zhu, J.; Xu, F. Fault Diagnosis of Motor Vibration Signals by Fusion of Spatiotemporal Features. *Machines* **2022**, *10*, 246. https://doi.org/10.3390/machines10040246

Academic Editor: Alejandro Gómez Yepes

Received: 4 March 2022
Accepted: 24 March 2022
Published: 30 March 2022

1. Introduction

The asynchronous motor is the most widely used mechanical drive equipment in industrial production and has become an important component in fields such as machinery manufacturing [1–3] and intelligent transportation [4,5]. Due to the harsh working environment, overload, and complex electromagnetic relationships, the motor is prone to stator winding inter-turn short circuit, broken rotor strips, air gap eccentricity, and bearing wear [6–8]. During operation, the failure of asynchronous motors may cause huge economic losses and casualties. Therefore, it is very important to evaluate the working state of the motor and detect potential faults to prevent mechanical accidents. Fault diagnosis of motors plays an important role in equipment maintenance, which can improve the quality of machines and reduce maintenance costs.

The common way of motor fault diagnosis is to use vibration signals for analysis. Vibration signals can be collected using acceleration transducers. Abnormal vibration signals can characterize equipment faults, such as asymmetry of the shaft system [9], a loose connection of components [10], and damaged rotor bearings [11]. Therefore, the acquisition and analysis of vibration signals have also become a common fault diagnosis scheme in the field of rotating machinery [12,13]. Fault diagnosis methods based on vibration signals [14,15] mainly include two stages: feature extraction and pattern recognition. The key to the asynchronous motor fault diagnosis technique is extracting feature information from non-smooth vibration signals with time-varying characteristics. In the time domain, some works [15,16] acquired amplitude, root mean square, and kurtosis for the analysis and diagnosis of vibration signals. However, it was susceptible to environmental noise and the methods have limitations. Some works [17,18] used Fourier transform to convert the signal from the time domain to the frequency domain. But the frequency characteristics of the vibration signal over time cannot be extracted effectively. The time-frequency domain analysis was performed by wavelet transform [19], short-time Fourier transform [20,21], and empirical mode decomposition [22,23], which extracted both time-domain and frequency-domain features. But the above methods are only effective for specific features and have poor adaptivity and robustness.

With the rise of deep learning, some neural networks have been introduced into the field of fault diagnosis [24–26]. The vibration characteristics of the signal can be obtained adaptively by learning the nonlinear mapping between the hidden layers in the network. Deep learning-based methods are less interpretable [27] but have high recognition accuracy. Such methods overcome the disadvantages of traditional methods that require manual feature extraction and have poor adaptability. Shi et al. [28] used a long short-term memory neural network (LSTM) to extract the temporal features of bearing vibration signals. However, the local information of the signal in the spatial dimension was ignored and the full key information could not be maintained when the data sequence is too long. Gao et al. [29] combined one-dimensional convolution and adaptive noise cancellation techniques to suppress the strong interference components in the one-dimensional time series of gearboxes. However, the time-series feature of the vibration signal was not fully utilized due to the limitation of the convolutional neural network field of perception. Zhu et al. [30] reconstructed the one-dimensional time-domain sequence into a two-dimensional data format and used two-dimensional convolution to capture the spatial features of the vibration signal. However, the dependencies between the positions of the spatial features were ignored, resulting in some important features not playing a significant role. Due to the convolutional stride and weight connection, the convolutional neural networks [31,32] cannot accurately obtain the temporal features of the vibration signal. In contrast, recurrent neural networks [33] can handle the temporal features of the signal but do not consider the information of the spatial dimension.

At present, motor fault diagnosis only uses the temporal features or spatial features of vibration signals for analysis. In this paper, spatial features and temporal features are combined to construct a spatiotemporal feature fusion network (STNet). The network solves the problem of accuracy loss caused by excessively long signal sequences and the lack of dependencies of each position. STNet is constructed for fault diagnosis of motor vibration signals. The main contributions of this paper are listed as follows.

1. The STNet utilizes the spatial feature extraction capability of a CNN and the temporal feature extraction capability of a GRU to construct a dual-stream network. The network combines temporal and spatial features for fault diagnosis of vibration signals instead of a single temporal or spatial feature.
2. The time series of vibration signals is much longer than the text in natural language processing. Recurrent neural networks do not preserve all the critical information. Therefore, a GRU with an attention mechanism is designed to extract temporal features and effectively synthesize the state and vibration features at different moments.
3. When the CNN extracts the spatial information of vibration signals, channel and position attention make the network capture the dependencies of each position. The attention mechanism obtains rich contextual features to enhance diagnostic accuracy.

The structure of this paper is as follows. Section 2 presents the attention-based mechanism for the GRU to capture the temporal features of vibration signals. Section 3 enhances the data by local mean decomposition and extracts the spatial features of vibration signals using a CNN with channel and position attention. Section 4 proposes a spatiotemporal feature fusion network. Section 5 validates the model by experiments.

2. Temporal Feature

When BP neural networks process data, there is no interrelationship between the front and back inputs of the network. However, the vibration signal of a motor is a one-dimensional time series, and the temporal relationship between each sampling point has an important impact on the performance of the diagnosis. A recurrent neural network introduces memory units to interconnect the neurons in this layer based on the ordinary neural network. The state of the hidden layer is related to the input at this moment and the state of the hidden layer at the previous moment. Therefore, the relationship of the time dimension can be extracted from the original vibration sequence by recurrent neural networks.

2.1. Gated Recurrent Unit

The spatiotemporal feature fusion network extracts the temporal features of motor vibration signals through the variant (gated recurrent unit) of the recurrent neural network. A gated recurrent unit introduces a gating mechanism to improve recurrent neural networks. A GRU can selectively forget some unimportant information while memorizing the state of the previous moment. A GRU alleviates the gradient disappearance of recurrent neural networks and solves the problem of untimely update of network parameters. The GRU controls the input, output, and state information of the hidden layer by the update gate z_t and the reset gate r_t. The internal structure is shown in Figure 1.

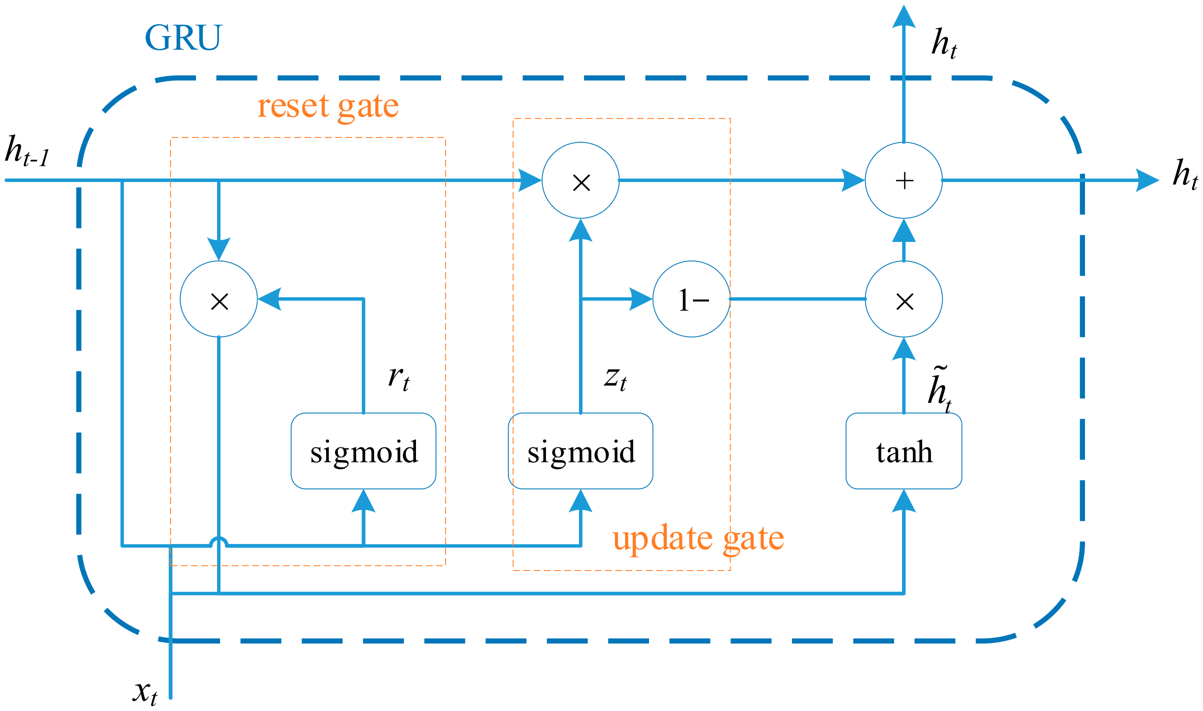


Figure 1. The GRU's internal structure.

The update gate z_t takes the current moment x_t and the previous moment information h_{t-1} by the weighting operation. Then the value between [0, 1] is obtained by the sigmoid function The value controls the effect of historical information on the state of the hidden layer at the current moment. The equation is as follows

$$z_t = \sigma(W_{tz} \cdot [h_{t-1}, x_t] + b_z) \tag{1}$$

where σ is the sigmoid function, W_{tz}, and b_z are the weights, h_{t-1} is the output at the previous moment, and x_t is the input at the current moment.

The reset gate r_t operates the current moment x_t and the previous moment information h_{t-1} with different weights, so that the model selectively forgets historical information that is irrelevant to the results. The equation is as follows

$$r_t = \sigma(W_{tr} \cdot [h_{t-1}, x_t] + b_r). \tag{2}$$

The status of the node at this moment is

$$\widetilde{h}_t = \tanh(W \cdot [r_t \times h_{t-1}, x_t] + b). \tag{3}$$

The final output of the hidden layer h_t is the sum of the information to be kept at the current moment and the information to be kept at the previous moment

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \widetilde{h}_t. \tag{4}$$

2.2. GRU Temporal Module Based on Attention Mechanism

The length of time series of motor vibration signals is much longer than the length of text in natural language processing. Although the GRU solves the problem of gradient disappearance in long sequence learning of recurrent neural networks, it still cannot retain all the key information when the time sequence is too long. Therefore, this paper not only selects the state output of the last moment of the GRU but also combines the state features of each moment of the GRU. Moreover, the attention mechanism is introduced to assign a weight coefficient to the output of the GRU at each moment. It makes the neural network pay attention to the data features of the output at different moments adaptively. The GRU temporal module based on the attention mechanism is shown in Figure 2.

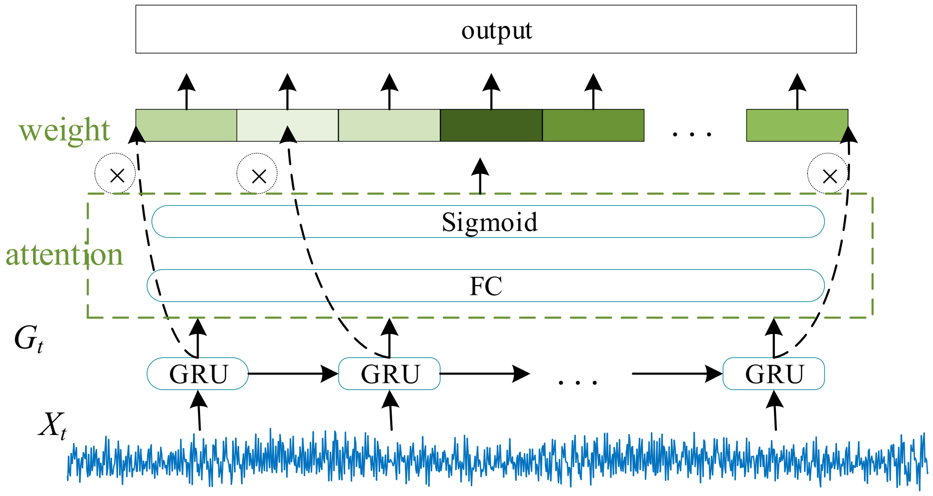


Figure 2. GRU temporal module based on an attention mechanism.

During the analysis of the vibration sequence, the output state of the GRU at the final moment determines the result of the fault diagnosis. However, the states of other moments also have many positive effects on the performance of the network. Therefore, the network not only relies on the output of the final moment but also considers the states of each moment in a comprehensive manner. The vibration signal X_t is fed into the GRU, which captures the vibration characteristics of the signal at each moment. The GRU outputs the state G_t at each moment as

$$G_t = GRU(X_t). \tag{5}$$

However, each momentary output of the GRU has a different degree of influence on the diagnosis results for different types of motor faults. Therefore, the states at each moment of the GRU are selected by the attention mechanism. The states with high relevance are kept and the states with low relevance are weakened. Then the weights of each moment state are obtained by the fully connected layer (FC) and sigmoid function. The weight parameters w_1 are

$$w_1 = \sigma(w(G_t) + b). \tag{6}$$

Finally, the output of GRU at each moment is multiplied by the weight parameter to obtain the output result O

$$O = weight \times G_t \tag{7}$$

3. Spatial Features

The GRU extracts the temporal features of vibration signals but ignores the spatially located information. This paper performs a time-frequency analysis of the vibration signal by local mean decomposition (LMD). The spatial features of the vibration signal after local mean decomposition are extracted by the convolutional neural network.

3.1. Local Mean Decomposition

The motor vibration signal is nonlinear and non-smooth. LMD adaptively decomposes the original vibration sequence into multiple instantaneous frequencies with physically meaningful product functions (PF). Each PF component is the product of a pure frequency modulation signal and an envelope signal, which can express the time-frequency distribution of the signal energy on the spatial scale. Then the vibration signal matrix is constructed and the original data is enhanced. The process of LMD for vibration signal processing is shown in Figure 3.

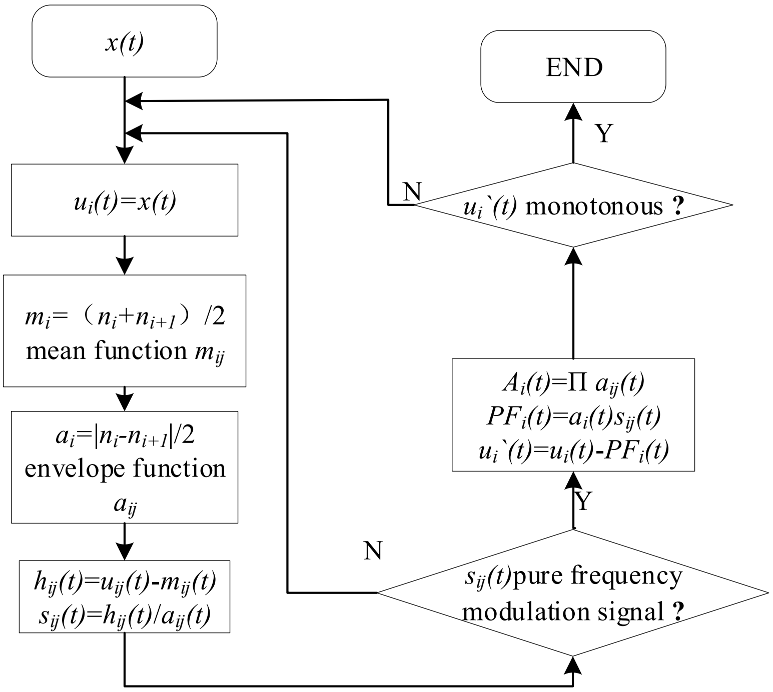


Figure 3. Local mean decomposition.

The original vibration signal $x(t)$ is decomposed by LMD and the mean value m_i of the adjacent local mean points is calculated. The curve is smoothed by the sliding average method to obtain the mean function m_{ij}. Then the envelope function a_{ij} is calculated. The mean function is separated from the original vibration signal to obtain $h_{ij}(t)$. Additionally, $h_{ij}(t)$ is demodulated to obtain $s_{ij}(t)$. If $s_{ij}(t)$ is a pure frequency modulation signal, the PF component $PF_i(t)$ and the residual signal $u_i'(t)$ are calculated based on the instantaneous amplitude function $a_i(t)$. If $u_i'(t)$ is a monotonic function, the decomposition ends and all PF components are obtained. The results of data decomposition are shown in Figure 4, where the original data $X(t)$ is decomposed into five PF components by LMD.

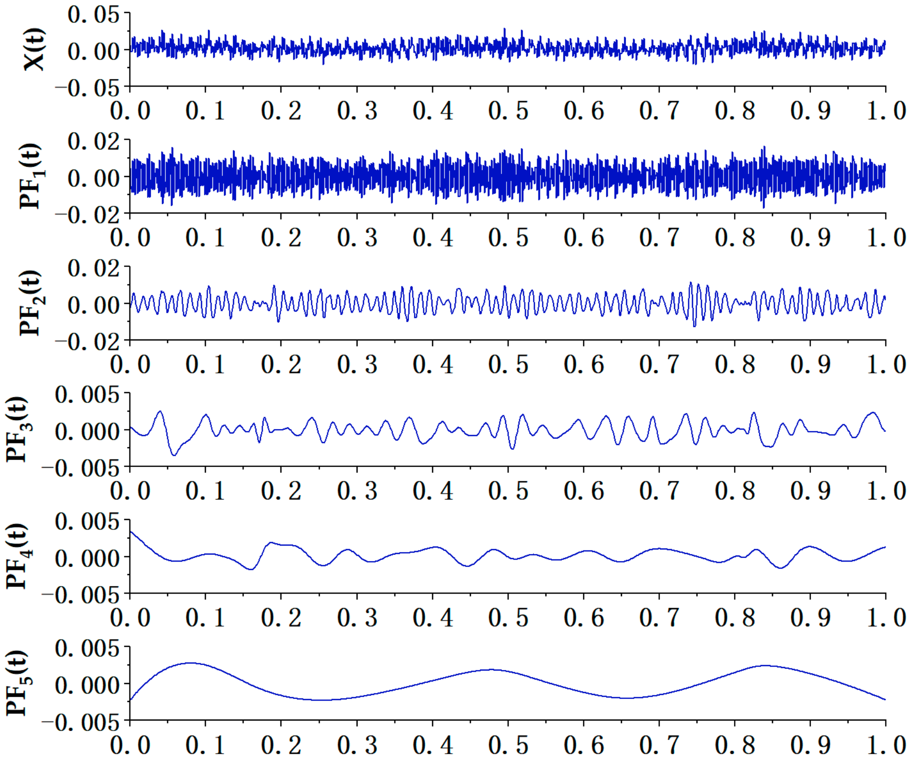


Figure 4. LMD motor vibration signal decomposition.

Convolutional neural networks are often used to process two-dimensional image signals, while the motor vibration signal $X(t)$ is a one-dimensional time-series signal, as follows

$$X(t) = [x_1, x_2, x_3, \cdots, x_t]. \tag{8}$$

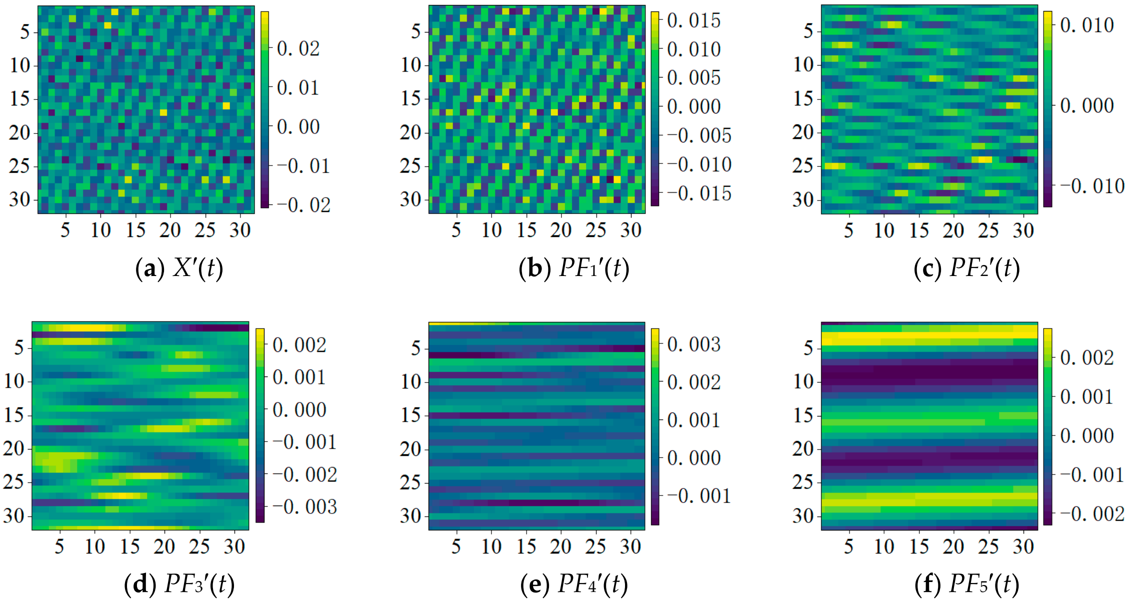
Therefore, the vibration signal is converted into a two-dimensional matrix $X'(t) \in \mathrm{R}^{M\times N}$

$$X'(t) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & \cdots & \cdots & x_{mn} \end{bmatrix}. \tag{9}$$

Each PF component is converted into two-dimensional data as shown in Figure 5. The PF components are concatenated with the two-dimensional data $X'(t)$ of the original vibration signal in the channel dimension. The final input matrix of the convolutional neural network is obtained. The method enhances the feature representation of the vibration signal in the spatial dimension.

3.2. CNN Module Based on Attention Mechanism

The convolutional neural network takes the multidimensional matrix of the motor vibration signal as input and adaptively extracts the spatial features of the signal. The different features have different effects on the fault diagnosis results. As shown in Figure 5, the same vibration signal decomposes with different PF components. It leads to huge differences between the different channels of the input 3D matrix $X_{in} \in \mathrm{R}^{c\times M\times N}$. The different channels have different effects on the diagnosis results for different fault types. Therefore, the attention mechanism is added to the channel dimension to make the model adaptively extract different channel features.



(**a**) $X'(t)$ (**b**) $PF_1'(t)$ (**c**) $PF_2'(t)$

(**d**) $PF_3'(t)$ (**e**) $PF_4'(t)$ (**f**) $PF_5'(t)$

Figure 5. Two-dimensional vibration matrix visualization. (**a**) is the original vibration signal. (**b**–**f**) are the PF components.

The structure of the channel attention is shown in Figure 6, where the input matrix X_{in} is convolved to obtain x $\in$ R$^{c \times m \times n}$ and $\otimes$ represents element-by-element multiplication.

$$x = w_i \otimes X_{in} + b_i \tag{10}$$

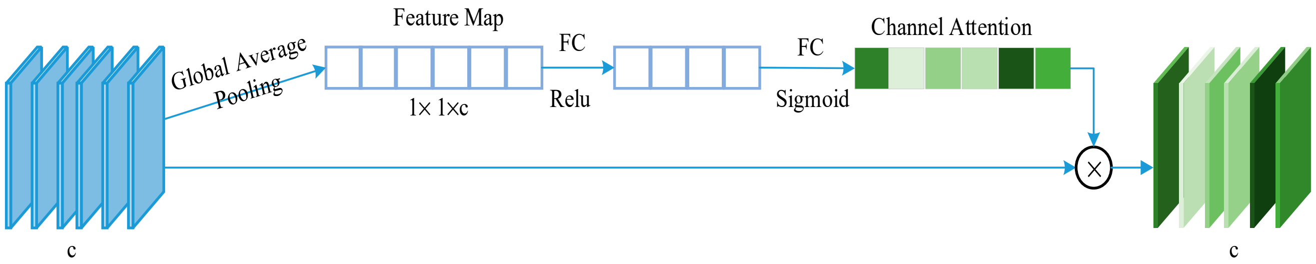


Figure 6. Channel attention module.

Then the $m \times n$ dimensions are compressed to 1×1 by global average pooling. The global feature distribution of the input matrix in the channel dimension is captured to obtain the feature map

$$map = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} x(i,j) \tag{11}$$

The feature maps are adjusted nonlinearly by the fully connected layer (FC). The module uses the sigmoid function to obtain the attentional weights of the channel dimensions C_{atte}

$$C_{atte} = \sigma(w_s \cdot (\mathrm{Relu}(w_r \cdot map + b_r)) + b_s) \tag{12}$$

Finally, the input features X_{in} are multiplied with the channel weights to rescale the features in the channel dimension.

The channel dimension completes the rescaling of the original features, and the channel attention adjusts the different channel features. However, there are also large differences in the data of different fault types of vibration signals in the same channel, as shown in Figure 7. Convolutional neural networks also need to consider the influence of different location features on the diagnosis results when extracting features. Therefore, this paper makes the network focus on the features of vibration signals in spatial dimensional features by position attention.

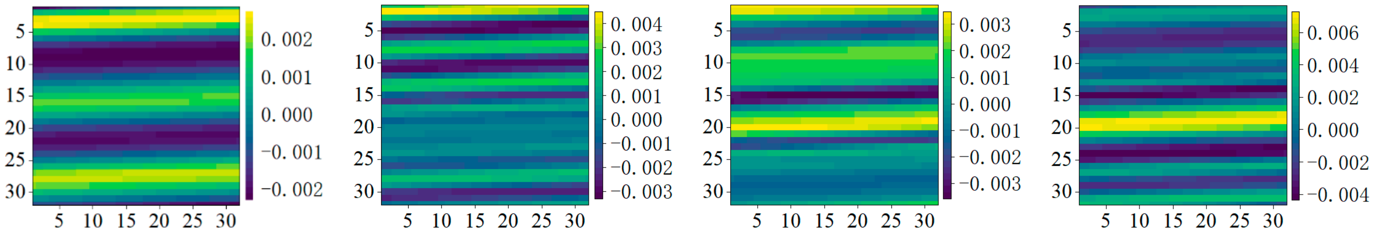


Figure 7. Data visualization of different fault types in the same channel.

The structure of the position attention is shown in Figure 8. The input features $x \in \mathrm{R}^{c\times m\times n}$ are computed separately for max pooling and average pooling to obtain feature maps $f_{max} \in \mathrm{R}^{1\times m\times n}$ and $f_{avg} \in \mathrm{R}^{1\times m\times n}$. Then the feature maps are concatenated in the channel dimension. Finally, the feature maps adopt convolutions and a sigmoid activation function to obtain the position attention P_{atten}

$$P_{atten} = \sigma(\mathrm{conv}(\mathrm{concat}(f_{max}, f_{avg}))). \tag{13}$$

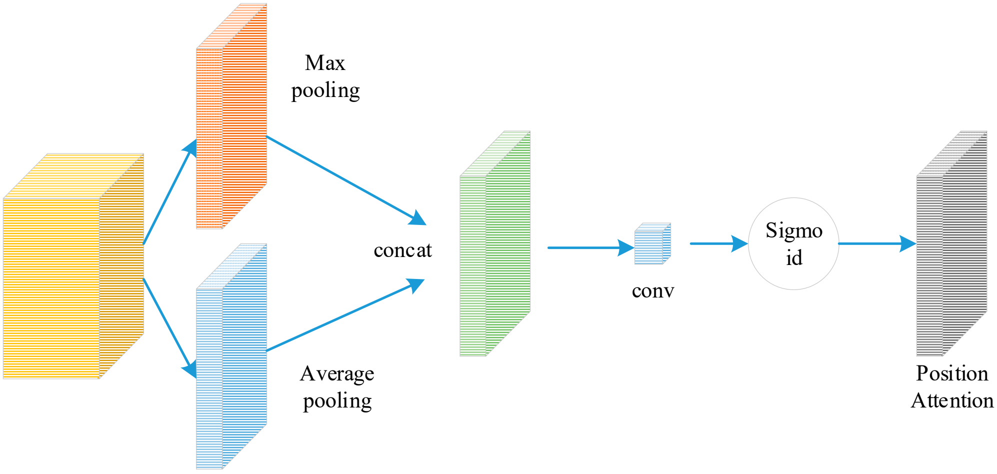


Figure 8. Position attention module.

4. Spatiotemporal Feature Fusion Network

The structure of the spatiotemporal feature fusion network is shown in Figure 9. The STNet uses a GRU to extract the temporal features of one-dimensional vibration signals. The GRU branch introduces the attention mechanism to synthesize the effect of each moment state on the performance in the long sequence signal. Meanwhile, the original vibration sequence is decomposed by LMD for time-frequency analysis. The original vibration data and each PF component are converted into multidimensional matrices as the input of the CNN. The CNN branch adaptively extracts the spatial features of the input matrix by convolutions. Meanwhile, considering the channel features and the influence of different fault features, the CNN branch adds channel attention and position attention to selectively enhance the spatial features of the signal. The attention mechanism acquires rich contextual information. Finally, the spatial and temporal features of the vibration signal are fused, and the softmax layer classifies the fused features.

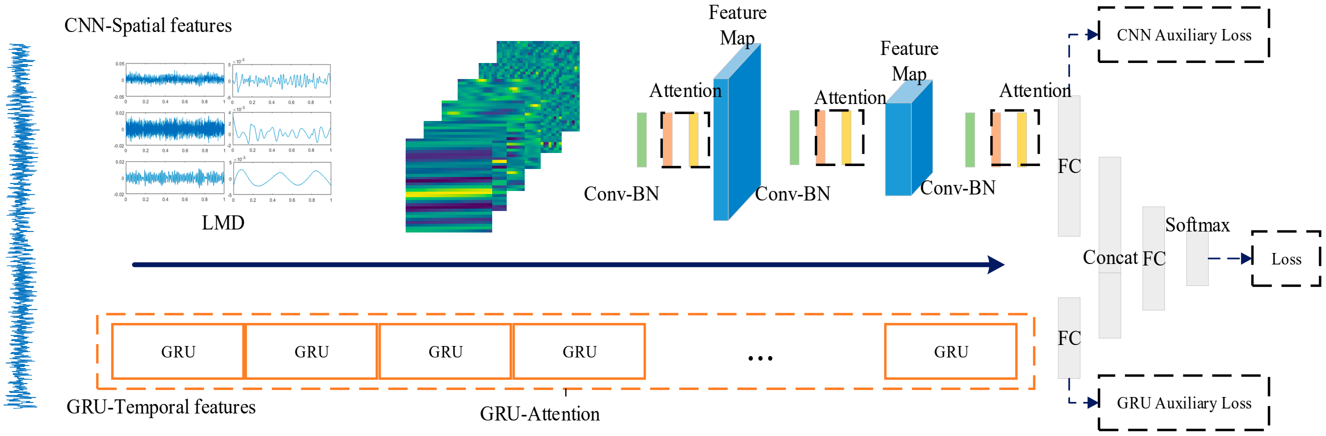


Figure 9. Spatiotemporal feature fusion network.

STNet is a dual-stream network consisting of a GRU branch and CNN branch. The specific network layers are shown in Table 1, where Conv-BN denotes the convolution layer and batch normalization layer, and FC is the fully connected layer. The input of the CNN branch is the vibrational signal matrix with the size of 6 × 32 × 32. The network uses the convolution kernel with the size of 3 × 3 to extract features. The padding type of the convolution kernel is “SAME”. Then, the kernel is normalized by the BN layer with a Relu activation function. The CNN branch recalibrates the original features by channel attention and position attention. The spatial resolution of the feature map at each stage becomes half that of the previous stage, and the number of channels becomes twice that of the previous stage. The network obtains a feature map with the size of 128 × 8 × 8 by three stages of feature extraction. The captured features are then fed into the fully connected layer with 1024 neurons. The input of the GRU branch is the original vibration signal with 1024 sampling points. The network obtains the temporal features through the 2-layer GRU attention unit, and the features are fed into the fully connected layer with 128 neurons. The fully connected layers of the CNN branch and GRU branch are concatenated, and the number of neurons is 1152. The network is nonlinearly adjusted by two fully connected layers. Finally, the diagnosis results of eight faults are output by the softmax function.

When the STNet extracts features, there are significant differences between the spatial features extracted by the CNN and the temporal features extracted by the GRU. Therefore, the CNN auxiliary loss function and GRU auxiliary loss function are added respectively during the training process. The auxiliary loss function supervises the temporal features and spatial features extracted by the network separately to reduce the generation of invalid information. The auxiliary loss function not only promotes the backpropagation of the network but also enhances the canonical representation of temporal and spatial features. The final loss function (L_{total}) of the network is shown as follows

$$L = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{14}$$

$$L_{\text{total}} = \alpha L_{\text{CNN}} + \beta L_{\text{GRU}} + L_{\text{loss}} \tag{15}$$

where M is the number of categories; y_{ic} is the symbolic function; p_{ic} is the probability that sample i belongs to c; α and β are the weights of the auxiliary loss function.

Table 1. The STNet's structure.

Layer	Node	Stride	Output Size	Layer	Node	Stride	Output Size
CNN Branch				GRU Branch			
			6 × 32 × 32				1024
Conv-BN	32	2	32 × 16 × 16	FC	990	-	990
Channel-Position Attention	-	1	32 × 16 × 16	GRU	330	-	330
Conv-BN	64	2	64 × 8 × 8	Attention	-	-	330
Channel-Position Attention	-	1	64 × 8 × 8	GRU	110	-	110
Conv-BN	128	2	128 × 8 × 8	Attention	-	-	110
Channel-Position Attention	-	1	128 × 8 × 8	FC	128	-	128
FC	1024	-	1024				
Concat (1152) FC (512)-FC (128) Softmax (8)							

5. Experiments

5.1. Data

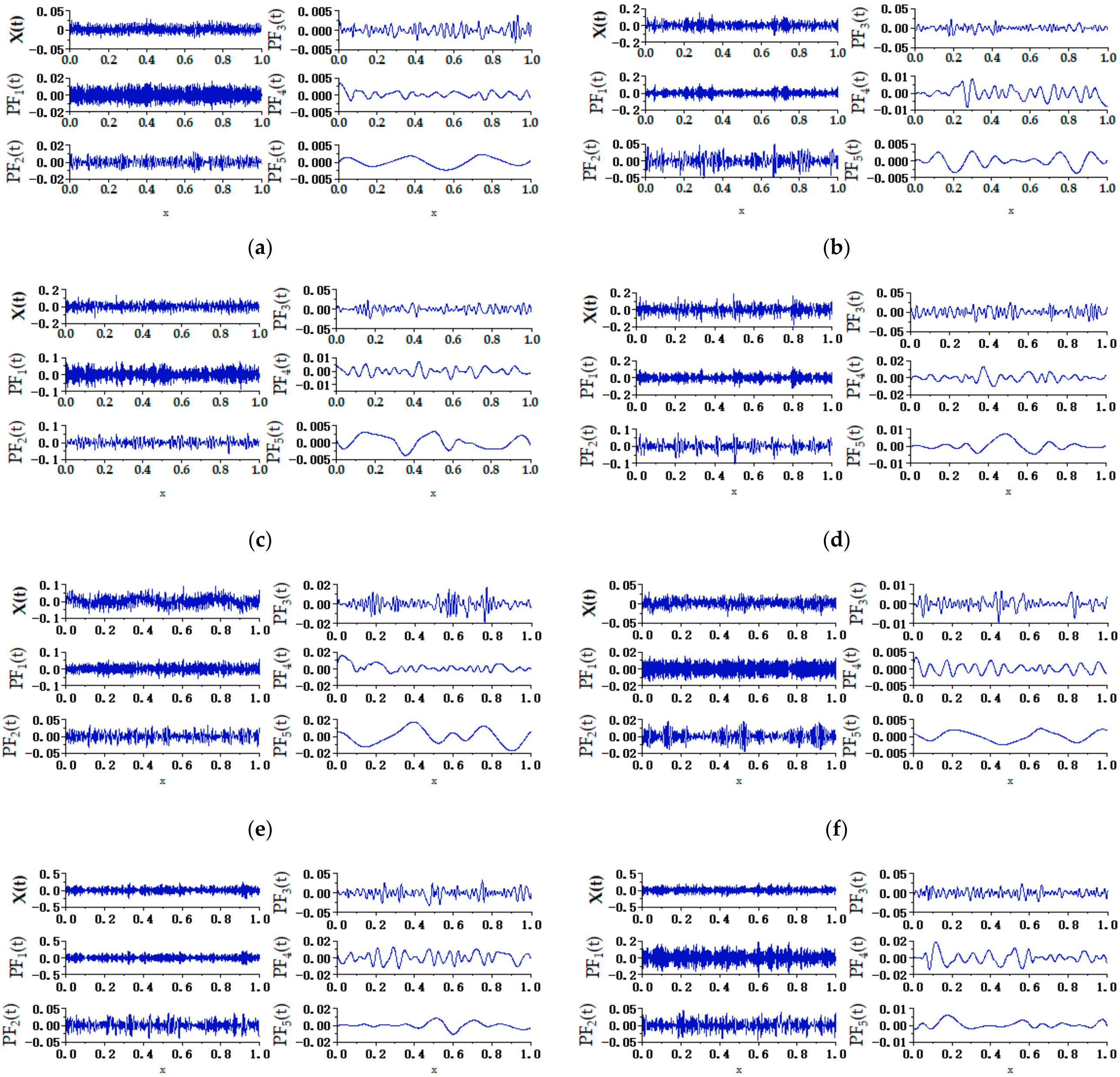
The main types of faults in the experimental motor vibration data are inter-turn short circuit, air gap eccentricity, rotor broken strips, bearing seat damage, bearing wear, etc. There are 8 kinds of samples, the number of samples is 8000, and the number of sampling points per second is 1024, as shown in Table 2. The deep learning framework is PaddlePaddle 1.8.4. The CPU of the training platform is Intel Xeon Gold 6171C. The GPU is Nvidia Tesla V100 (16G). GPU acceleration is performed by CUDA 10.1, and the experimental dataset is divided into training and test sets (7:3).

Table 2. Fault types.

Label	Types	Numbers
0	Normal	1000
1	2 turns short circuit	1000
2	4 turns short circuit	1000
3	8 turns short circuit	1000
4	Air gap eccentricity	1000
5	Broken rotor strip	1000
6	Bearing seat damage	1000
7	Bearing wear	1000

5.2. Experiment Analysis

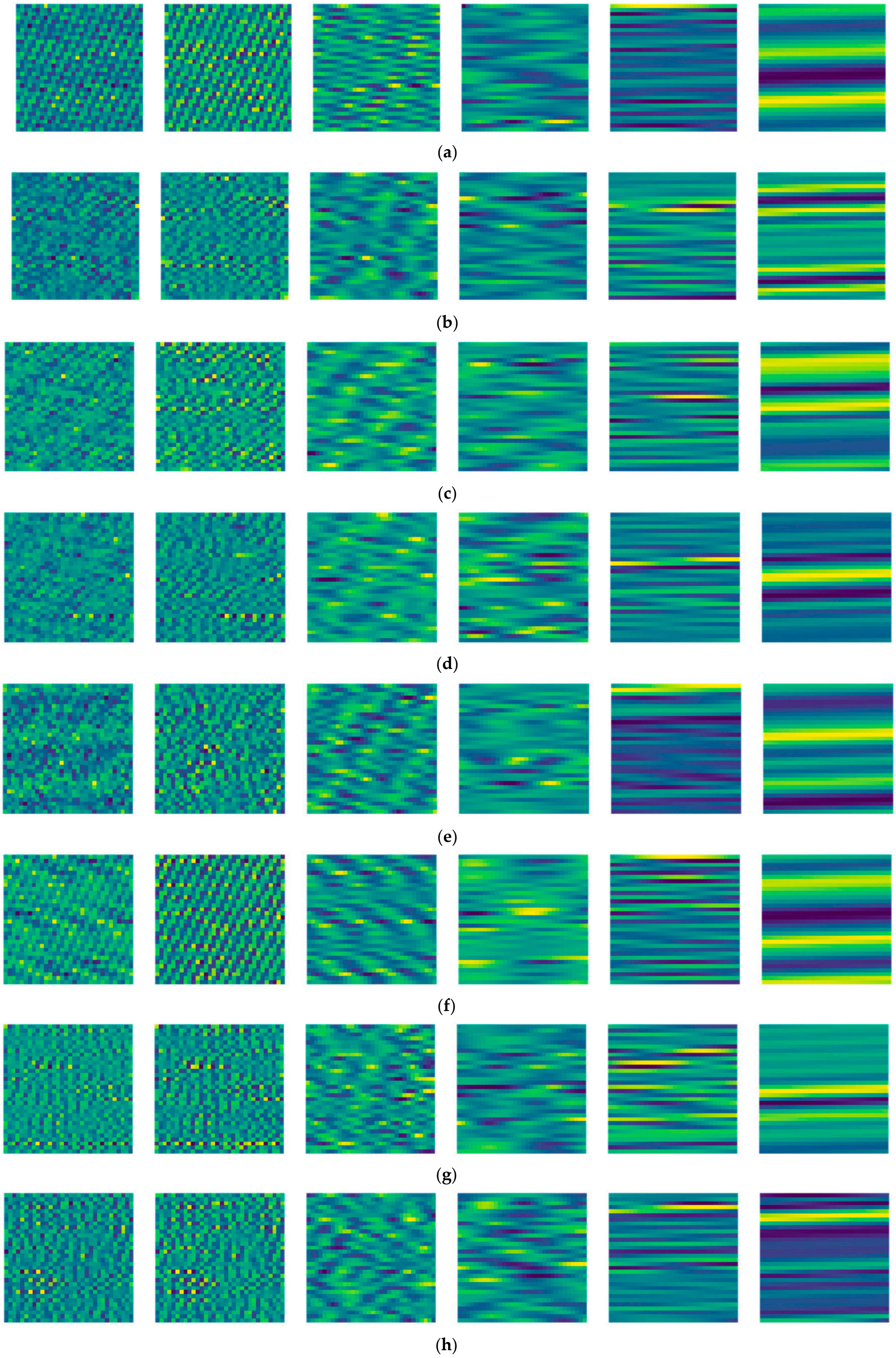
When the network extracts features using the GRU, only the features in the time domain of the vibration signal are captured. However, the vibration signal also contains rich features in the frequency domain. Therefore, the original vibration data is decomposed by LMD. The decomposition results of each fault type are shown in Figure 10. When abnormal vibration occurs in the accelerometer, each PF component can show the amplitude modulation and frequency modulation signals of the abnormal vibration. The vibration signal is enhanced so that the CNN extracts the vibration features by the original vibration sequence and each PF component.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**)

Figure 10. Visualization of local mean decomposition of fault signals. (**a**) normal; (**b**) 2 turns short circuit; (**c**) 4 turns short circuit; (**d**) 8 turns short circuit; (**e**) air gap eccentricity; (**f**) broken rotor strip; (**g**) bearing seat damage; (**h**) bearing wear.

A convolutional neural network has unique superiority in two-dimensional image recognition due to the special structure of local weight sharing and the presence of the local perceptual field. The visualization results of each fault signal transformed into the two-dimensional matrix are shown in Figure 11. The original vibration signal is 1024 sampling points, and the size of the transformed 2D matrix is 32 × 32. Similarly, each PF component is also transformed into a two-dimensional matrix and connected to the two-dimensional matrix of the original vibration signal in the channel dimension. Finally, the input size of the CNN branch is 6 × 32 × 32. The visualization results of the two-dimensional matrix show that the PF component matrices of different faults have large differences in different dimensions, and the fault features extracted by the CNN would have a positive effect on the performance of diagnosis.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

(**f**)

(**g**)

(**h**)

Figure 11. Two-dimensional matrix visualization of fault data. (**a**) normal; (**b**) 2 turns short circuit; (**c**) 4 turns short circuit; (**d**) 8 turns short circuit; (**e**) air gap eccentricity; (**f**) broken rotor strip; (**g**) bearing seat damage; (**h**) bearing wear.

The input size of the CNN branch is 6 × 32 × 32, and the sequence length of the GRU branch input is 1024. The number of network training epochs is 100. The batch size is 600. The model parameters are updated using the Adam optimization algorithm. The learning rate adjustment strategy is "Poly", with an initial learning rate of 0.001 and a power of 0.9. The loss function is the cross-entropy loss function. The weight of the CNN network auxiliary loss function is 0.1. The weight of the GRU network auxiliary loss function is 0.9. The evaluation index is the accuracy rate. The loss and accuracy curves of the training set and test set with the number of epochs are shown in Figure 12.

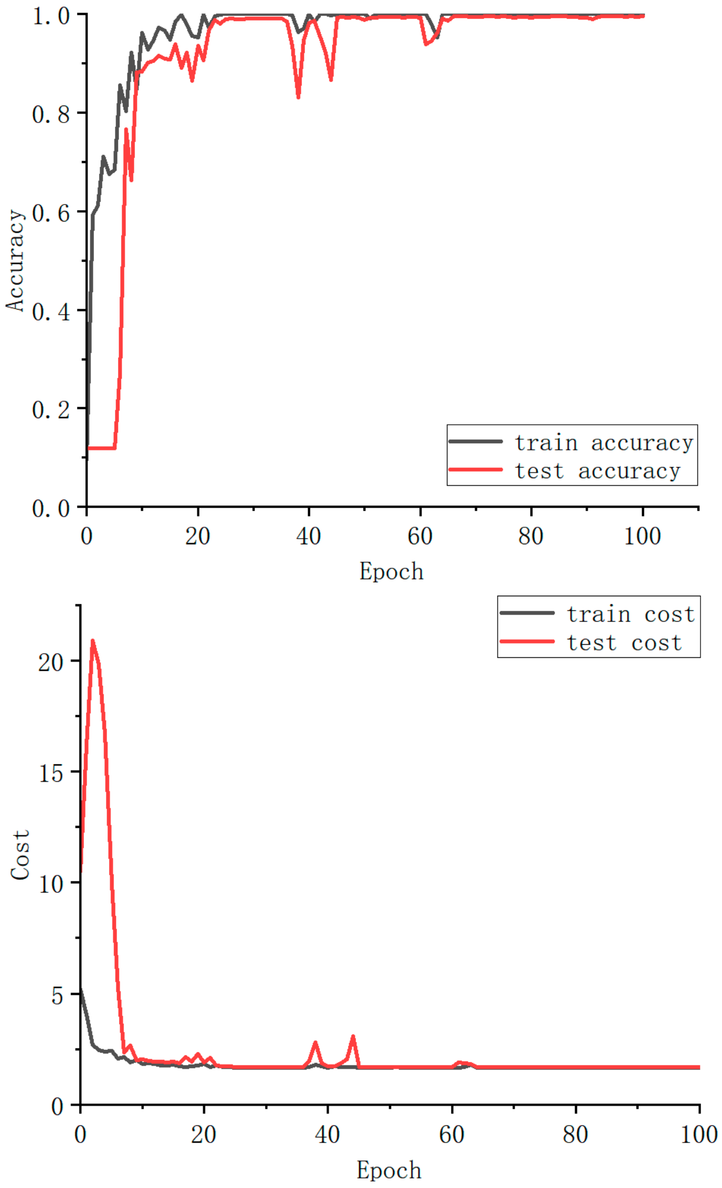


Figure 12. Training process loss and accuracy variation.

The test set loss increases sharply in the first 10 rounds of training, but the training set and test set losses gradually decrease with the increase of iterations. It indicates that the model is converging and approaching 0. After 60 epochs, the training set loss and test set loss are close to overlapping. The waveforms do not have large fluctuations and there are no overfitting problems.

The model is validated for each type of fault after training, and the results are shown in Table 3. The number of error samples for inter-turn short circuit fault is three, and the number of error samples for bearing seat damage is three. The recognition accuracy of each type of fault is above 99%. The model has high recognition accuracy.

Table 3. The result of each category of fault identification.

Label	Types	Accuracy
0	Normal	100%
1	2 turns short circuit	99.67%
2	4 turns short circuit	99.33%
3	8 turns short circuit	100%
4	Air gap eccentricity	100%
5	Broken rotor strip	100%
6	Bearing seat damage	99%
7	Bearing wear	100%

To verify the performance of each module in the STNet, five ablation experiments are set up. The results are shown in Table 4. The accuracy of the temporal features extracted from the vibration signal using the GRU is 98.58%, while the accuracy of the spatial features captured from the vibration signal using the CNN is 98.83%. The CNN + GRU model with the fusion of temporal and spatial features improves the accuracy by 0.39% and 0.04%, respectively. Compared with the single branch, it indicates that both temporal and spatial features of the vibration signal are indispensable parts for fault diagnosis. The CNN + GRU + attention model with the attention module on the CNN branch and GRU branch improves the accuracy by 0.59% compared to the model without attention. The attention mechanism considers the importance of different features and makes the important features play a significant role in the network. The final accuracy of the STNet with auxiliary loss function is 99.75%. The auxiliary loss function facilitates the network backpropagation to update the parameters and enhances the feature representation of each branch.

Table 4. Ablation experiments.

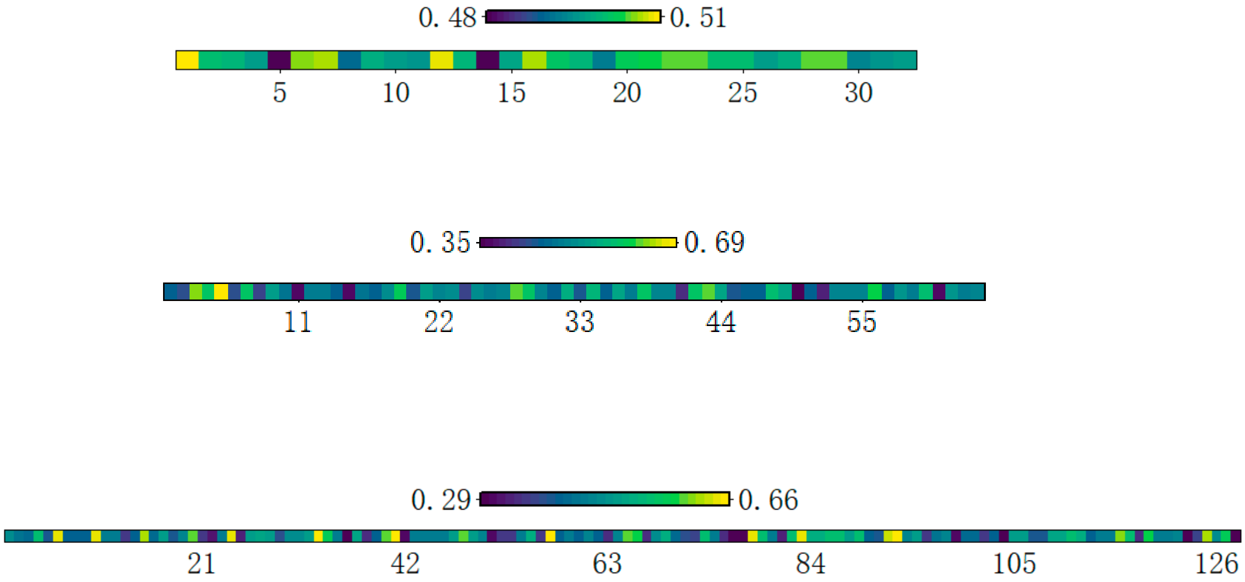
Model	Accuracy
GRU	98.58%
CNN	98.83%
CNN + GRU	98.97%
CNN + GRU + Attention	99.56%
CNN + GRU + Attention + Auxiliary Loss	99.75%

To further investigate the effect of the attention module on the network performance, the attention matrices of the GRU branch and the CNN branch are visualized. Figure 13a represents the channel attention for the three-stage feature extraction in the CNN branch with channel dimensions of 32, 64, and 128. The shallow layer of the CNN branch requires sufficient feature extraction of the vibration signal to preserve all feature information as much as possible. Therefore, the attention varies from 0.48 to 0.51, which is not a large range. Due to the number of network layers increasing and the number of channels increasing, the redundant features are increased. The network needs to suppress the redundant channels, while the effective channel features are enhanced. So, the range of variation of channel attention increases. Figure 13b represents the position attention of the three-stage feature extraction in the CNN branch with dimensions of 16 × 16, 8 × 8, and 8 × 8. The position attention becomes more and more focused because the local features of the convolutional neural network are extracted. Figure 13c represents the attention of the output features of the second GRU in the GRU branch. The GRU module outputs the prediction results of multiple time series. The output represents the impact of each moment on the diagnostic results. It retains the results with high relevance by the attention mechanism, so the GRU attention does not fluctuate greatly.

To further verify the fault diagnosis capability of the STNet, it is compared with BP, 1D-CNN, multichannel-CNN, and inception-LSTM models. The experimental results are shown in Table 5. The BP network diagnoses the fault types by nonlinear mapping without considering the temporal and spatial features of the signal. Therefore, the recognition accuracy is only 96.12%. The 1D-CNN model uses 1D convolution to obtain the abstract features and local features of the vibration signal. The 1D-CNN model improves the accuracy by 2.12% compared to the BP network. The multichannel-CNN model weights different receptive fields and captures contextual information at different scales. The inception-LSTM model extracts temporal information under several different receptive fields with an accuracy of 99.34%. Compared with BP, 1D-CNN, multichannel-CNN, and inception-LSTM models, the STNet obtains the highest accuracy of 99.75%. The STNet combines spatial features and temporal features instead of single features, compared with BP, 1D-CNN, and multichannel-CNN models. Compared with the inception-LSTM model, STNet uses the attention mechanism to select features adaptively. Therefore, both temporal and spatial features have a positive impact on the performance of diagnosis during the analysis of vibration signals. The number of parameters of STNet is 9.2876 M and the number of floating-point operations (FLOPs) is 0.02 G.

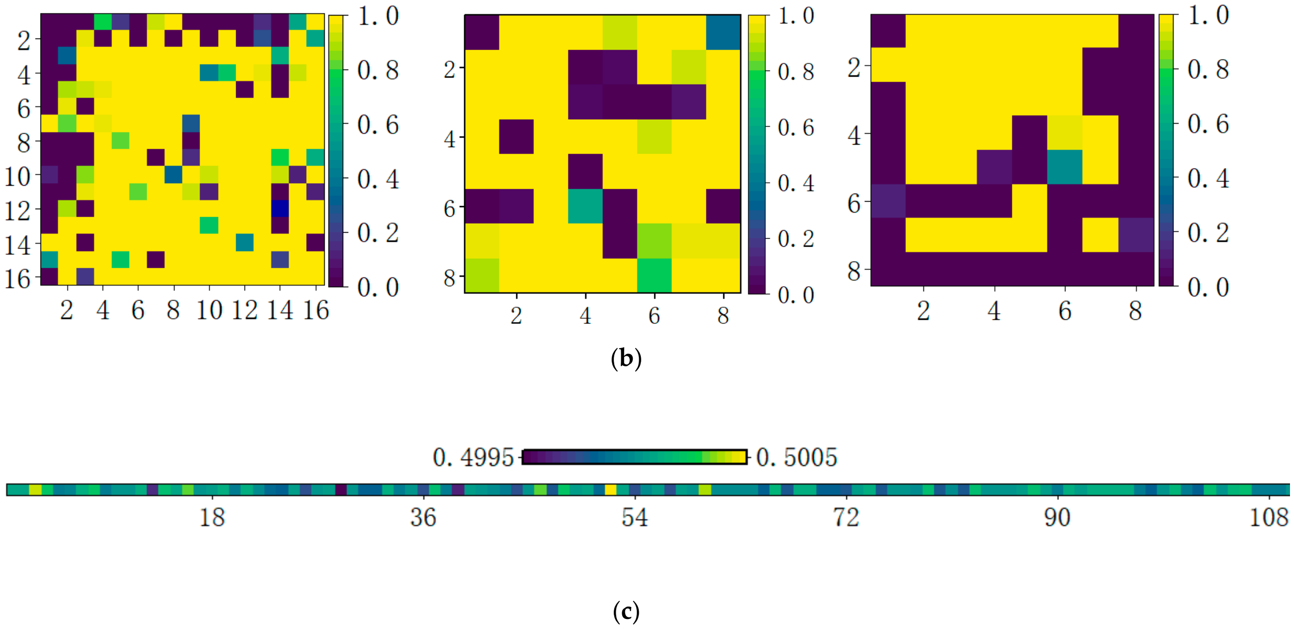
Table 5. Model comparison experiments.

Model	Accuracy
BP	96.12%
1D-CNN	98.24%
Multichannel-CNN	99.17%
Inception-LSTM	99.34%
STNet	99.75%



(**a**)

Figure 13. *Cont.*



(**b**)

(**c**)

Figure 13. Attention visualization. (**a**) CNN branching channel attention visualization; (**b**) CNN branching position attention visualization; (**c**) GRU branch attention visualization.

6. Conclusions

In the paper, the fault diagnosis for motor vibration signals has been investigated based on spatiotemporal feature fusion. The method has used gated recurrent units and convolutional neural networks to extract the temporal and spatial features of vibration signals. Since the time series of vibration signals were too long to retain all the key information, a GRU has extracted the temporal features by an attention mechanism to effectively synthesize the states of different time series and the vibration features at different moments. When extracting spatial features, the one-dimensional time-domain signal has been converted into a two-dimensional matrix using local mean decomposition and matrix transformation to extend the data dimensionality. The CNN model based on the attention mechanism adaptively has extracted the channel and location features of the signal. In the experimental evaluation of eight different vibration signals, the vibration signal processing method combined with spatiotemporal feature fusion has obtained 99.75% recognition accuracy. The method has improved the diagnostic performance effectively, which is important for the safe detection and stable operation of the system.

Author Contributions: Conceptualization, L.W. and F.X.; methodology, L.W.; software, C.Z.; validation, L.W., C.Z., J.Z. and F.X.; writing—review and editing, L.W., C.Z., J.Z. and F.X.; visualization, F.X.; supervision, F.X.; project administration, F.X.; funding acquisition, L.W. and F.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Heilongjiang Province Key R&D Program (GA21A304), the Fundamental Research Funds in Heilongjiang Provincial Universities (135509404 and 135409102), and the Open project of Agricultural multidimensional sensor information Perception of Engineering and Technology Center in Heilongjiang Province (DWCGQKF202105).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Yu, J.; Liu, X. One-dimensional residual convolutional auto-encoder for fault detection in complex industrial processes. *Int. J. Prod. Res.* **2021**, *196*, 1–20. [CrossRef]
2. Han, T.; Liu, C.; Yang, W.; Jiang, D. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. *Knowl. Based Syst.* **2019**, *165*, 474–487. [CrossRef]
3. Chen, Z.; Gryllias, K.; Li, W. Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network. *IEEE Trans. Ind. Inform.* **2019**, *16*, 339–349. [CrossRef]
4. Chen, H.; Jiang, B. A review of fault detection and diagnosis for the traction system in high-speed trains. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 450–465. [CrossRef]
5. Chen, H.; Jiang, B.; Ding, S.; Huang, B.S. Data-driven fault diagnosis for traction systems in high-speed trains: A survey, challenges, and perspectives. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 1–17. [CrossRef]
6. Hsueh, Y.-M.; Ittangihal, V.R.; Wu, W.-B.; Chang, H.-C.; Kuo, C.-C. Fault diagnosis system for induction motors by CNN using empirical wavelet transform. *Symmetry* **2019**, *11*, 1212. [CrossRef]
7. Kao, I.H.; Wang, W.J.; Lai, Y.H.; Perng, J.W. Analysis of permanent magnet synchronous motor fault diagnosis based on learning. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 310–324. [CrossRef]
8. Namdar, A.; Samet, H.; Allahbakhshi, M.; Tajdinian, M.; Ghanbari, T. A robust stator inter-turn fault detection in induction motor utilizing kalman filter-based algorithm. *Measurement* **2022**, *187*, 110181. [CrossRef]
9. Vinayak, B.; Anand, K.; Jagadanand, G. Wavelet-based real-time stator fault detection of inverter-fed induction motor. *IET Electr. Power Appl.* **2020**, *14*, 82–90. [CrossRef]
10. Ben Abid, F.; Sallem, M.; Braham, A. Robust interpretable deep learning for intelligent fault diagnosis of induction motors. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 3506–3515. [CrossRef]
11. Hasan, M.J.; Islam, M.M.M.; Kim, J.M. Bearing fault diagnosis using multidomain fusion-based vibration imaging and multitask learning. *Sensors* **2022**, *22*, 56. [CrossRef]
12. Karabacak, Y.E.; Özmen, N.G.; Gümüşel, L. Intelligent worm gearbox fault diagnosis under various working conditions using vibration, sound and thermal features. *Appl. Acoust.* **2022**, *186*, 108463. [CrossRef]
13. Mao, W.; Feng, W.; Liu, Y.; Zhang, D.; Liang, X. A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis. *Mech. Syst. Signal Process.* **2021**, *150*, 107233. [CrossRef]
14. Yan, X.; Liu, Y.; Jia, M.; Zhu, Y. A multi-stage hybrid fault diagnosis approach for rolling element bearing under various working conditions. *IEEE Access* **2019**, *7*, 138426–138441. [CrossRef]
15. Jan, S.; Lee, Y.-D.; Shin, J. Sensor fault classification based on support vector machine and statistical time-domain features. *IEEE Access* **2017**, *5*, 8682–8690. [CrossRef]
16. Cui, L.; Huang, J.; Zhang, F. Quantitative and localization diagnosis of a defective ball bearing based on vertical horizontal synchronization signal analysis. *IEEE Trans. Ind. Electron.* **2017**, *66*, 8695–8706. [CrossRef]
17. Shao, S.-Y.; Sun, W.-J.; Yan, R.-Q.; Wang, P.; Gao, R.X. A deep learning approach for fault diagnosis of induction motors in manufacturing. *Chin. J. Mech. Eng.* **2017**, *30*, 1347–1356. [CrossRef]
18. Lin, H.; Ye, Y.; Huang, B.; Su, J. Bearing vibration detection and analysis using enhanced fast Fourier transform algorithm. *Adv. Mech. Eng.* **2016**, *8*, 1687814016675080. [CrossRef]
19. Qu, J.; Zhang, Z.; Gong, T. A novel intelligent method for mechanical fault diagnosis based on dual-tree complex wavelet packet transform and multiple classifier fusion. *Neurocomputing* **2016**, *171*, 837–853. [CrossRef]
20. Li, L.; Han, N.N.; Jiang, Q.T. A chirplet transform-based mode retrieval method for multicomponent signals with crossover instantaneous frequencies. *Digit. Signal Process.* **2022**, *120*, 103262. [CrossRef]
21. Li, H.; Zhang, Q.; Qin, X.; Sun, Y.T. Fault diagnosis method for rolling bearings based on short-time Fourier transform and convolution neural network. *Shock Vib.* **2018**, *37*, 124–131.
22. Ali, J.B.; Fnaiech, N.; Saidi, L.; Chebel-Morello, B.; Fnaiech, F. Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals. *Appl. Acoust.* **2015**, *89*, 16–27.
23. Yu, X.; Dong, F.; Ding, E.; Wu, S.; Fan, C. Rolling bearing fault diagnosis using modified LFDA and EMD with sensitive feature selection. *IEEE Access* **2017**, *6*, 3715–3730. [CrossRef]
24. Zhang, W.; Li, C.; Peng, G.; Chen, Y.; Zhang, Z. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mech. *Syst. Signal Process.* **2018**, *100*, 439–453. [CrossRef]
25. Zhu, J.; Chen, N.; Peng, W. Estimation of bearing remaining useful life based on multiscale convolutional neural network. *IEEE Trans. Ind. Electron.* **2019**, *66*, 3208–3216. [CrossRef]
26. Li, D.; Zhang, M.; Kang, T.; Li, B.; Xiang, H.; Wang, K.; Pei, Z.; Tang, X.; Wang, P. Fault diagnosis of rotating machinery based on dual convolutional-capsule network (DC-CN). *Measurement* **2022**, *187*, 110258. [CrossRef]
27. Chen, H.; Liu, Z.; Alippi, C.; Huang, B.; Liu, D. Explainable Intelligent Fault Diagnosis for Nonlinear Dynamic Systems: From Unsupervised to Supervised Learning. *TechRxiv* **2022**. *Preprint*. [CrossRef]
28. Shi, H.; Guo, L.; Tan, S.; Bai, X.; Sun, J. Rolling bearing initial fault detection using long short-term memory recurrent network. *IEEE Access* **2019**, *7*, 171559–171569. [CrossRef]
29. Gao, J.; Guo, Y.; Wu, X. Gearbox bearing fault diagnosis based on SANC and 1-D CNN. *Shock Vib.* **2020**, *39*, 204–209.

30. Zhu, X.; Hou, D.; Zhou, P.; Han, Z.; Yuan, Y.; Zhou, W.; Yin, Q. Rotor fault diagnosis using a convolutional neural network with symmetrized dot pattern images. *Measurement* **2019**, *138*, 526–535. [CrossRef]
31. Guo, S.; Zhang, B.; Yang, T.; Lyu, D.; Gao, W. Multitask Convolutional Neural Network with Information Fusion for Bearing Fault Diagnosis and Localization. *IEEE Trans. Ind. Electron.* **2019**, *67*, 8005–8015. [CrossRef]
32. Cao, P.; Zhang, S.; Tang, J. Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning. *IEEE Access* **2018**, *6*, 26241–26253. [CrossRef]
33. Liu, H.; Zhou, J.; Zheng, Y.; Jiang, W.; Zhang, Y. Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. *ISA Trans.* **2018**, *77*, 167–178. [CrossRef] [PubMed]