machines

MDPI

Article

Bearing Fault Diagnosis for Time-Varying System Using Vibration–Speed Fusion Network Based on Self-Attention and Sparse Feature Extraction

Fulin Chi, Xinyu Yang, Siyu Shao * and Qiang Zhang

Air and Missile Defense College, Air Force Engineering University, Xi'an 710000, China
* Correspondence: cathygx.sy@gmail.com

Abstract: Nowadays, most deep-learning-based bearing fault diagnosis methods are studied under the condition of steady speed, while the performance of these models cannot be fully played under time-varying conditions. Therefore, in order to facilitate the practical application of a deep learning model in bearing fault diagnosis, a vibration–speed fusion network is proposed, which utilizes a transformer with a self-attention module to extract vibration features and utilizes a sparse autoencoder (SAE) network to extract sparse features from speed pulse signal. The vibration–speed fusion network enables the efficient fusion of different signals in a high-dimensional vector space with a high degree of model interpretability, without additional signal processing steps. After tuning the hyperparameters of the network, the key segments of the bearing's time-domain vibration signals can be optimally extracted, the network performance is much better than traditional deep learning methods, and the classification accuracy can reach 95.18% and 99.85% on the two public bearing datasets from the Xi'an Jiaotong University and the University of Ottawa.

Keywords: intelligent fault diagnosis; transformer; time-varying rotational speed; feature visualization

check for updates

Citation: Chi, F.; Yang, X.; Shao, S.; Zhang, Q. Bearing Fault Diagnosis for Time-Varying System Using Vibration–Speed Fusion Network Based on Self-Attention and Sparse Feature Extraction. *Machines* **2022**, *10*, 948. https://doi.org/10.3390/machines10100948

Academic Editors: Hongrui Cao, Jianping Xuan and Yongqiang Liu

Received: 12 September 2022
Accepted: 17 October 2022
Published: 18 October 2022

1. Introduction

Rolling bearings are widely used in aerospace engineering, industrial manufacturing, military equipment, and other fields [1]. However, as bearings sustain a lot of thermal and compressive stresses when supporting the rotating shaft and other components for high-speed rotation, they are prone to wear and tear, and fail to meet the requirements, or even break down and cause the transmission system to collapse [2,3]. Therefore, it is necessary to monitor the health status of bearings during operation to improve the efficiency of machinery operation and reduce equipment failure [4]. A prognostics and health management system for key components such as bearings has been established at NASA to achieve the goal of a low failure rate of complex systems through real-time monitoring [5]. This high real-time, highly interactive health monitoring approach replaces traditional methods such as spectrum analysis [6], acoustic emission [7], and thermal imaging [8].

Under the actual working conditions, the bearing speed is prone to change, and the internal parts of the bearing will produce uneven force changes under variable speed, which leads to bearing failure under variable working conditions. At the same time, due to the influence of speed change, the effect of the time–frequency spectrum analysis method based on bearing vibration signals in the extraction of early, weak fault characteristics will be seriously reduced, resulting in a higher difficulty of early fault diagnosis under variable working conditions than that of constant working conditions [9]. However, the prerequisite of most current bearing fault diagnosis methods is constant speed, which means that in order to ensure the stability of monitoring, the existing bearing condition monitoring system needs to be further optimized and improved to adapt to more complex working conditions. In the study of bearing fault feature extraction under variable speed, the order spectrum analysis method based on angular domain resampling is similar to the traditional spectrum analysis method, mainly by interpolating the collected rotational speed phase signals to obtain a clearer order spectrum and enhance the extraction effect of bearing fault signal features under variable speed conditions [10]. However, the effect of this method suffers in the case of a strong rotational speed variation and suffers from the same drawbacks as traditional spectrum analysis methods, both of which rely on a manual selection of features. As shown in Figure 1, the spectra of the constant-speed bearing (model: 6205-2RS JEM SKF) and the variable-speed bearing (model: NSK6203) were obtained by a fast Fourier transform (FFT) during the test period, from which it can be seen that a serious "spectrum blurring" phenomenon occurred in the case of the variable speed.

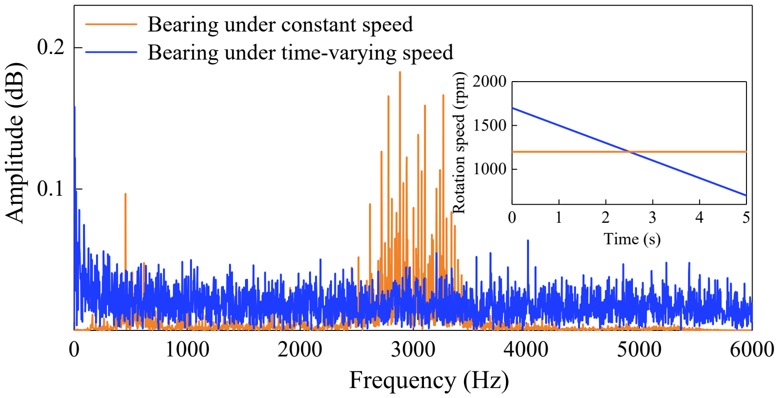


Figure 1. Spectrum blurring phenomenon at time-varying speed.

At present, deep learning is increasingly used in various fields because of its own excellent performance, especially in the field of signal analysis, and its excellent feature extraction capability can effectively overcome the shortcomings of traditional methods [11–15]. However, the bearing vibration features drift under variable speed, and the generalizability of the conventional model based on deep learning is significantly reduced. To solve this problem, there are two main research directions: one is to improve the network or enhance the bearing signal preprocessing step. For example, Zhang et al. [16] improved the diagnosis of a convolutional neural network (CNN) under variable operating conditions by designing a two-dimensional multiscale convolutional cascade network and generating two-dimensional image data from one-dimensional vibration signals. Hasan et al. [17] converted the acoustic emission signal to acoustic spectral imaging and then improved the overall robustness of the model for classification under variable speed by sharing parameters with a CNN. Zhao et al. [18] used a multiscale convolutional residual network to convert one-dimensional signals into three-dimensional images using a vector compression method for network training, which enhanced the signal feature extraction capability for bearings under variable operating conditions. The second research direction is to use transfer learning to enhance the cross-domain diagnostic capability. For example, He et al. [19] used a Morlet wavelet to improve the depth self-coding model to enhance the fault classification of bearings under new operating conditions. Han et al. [20] used data from known operating conditions to pretrain the CNN model and adapted this CNN model to fault diagnosis under unknown operating conditions by fine-tuning the weights. An et al. [21] adopted a framework of domain adaptation with multicore maximum mean discrepancies to make the features of different domains approach each other in the reconstructed kernel Hilbert space, which improved the stability and accuracy of the results. Both of the above research directions have proposed feasible solutions for bearing fault diagnosis under variable operating conditions, but there are still the following problems: first, different signal preprocessing methods cannot fairly compare the performance between models and cannot achieve end-to-end fault diagnosis; second, there are still limitations in the effect of cross-domain classification under time-varying speed, which can only achieve cross-domain diagnosis within the specified range of operating condition changes and cannot adaptively enhance the generalization. Therefore, to address the above problems, from the high-dimensional fusion of bearing vibration signal features and bearing speed pulse signal features, this paper proposes a vibration–speed data fusion network based on an SAE and a transformer (VSF-ST), which makes full use of the feature extraction ability of both networks for nonstationary signals to achieve end-to-end adaptive fault diagnosis classification under time-varying speed without additional signal processing, and provides a new idea to solve the problem under variable speed conditions. The main contributions of this manuscript are as follows:

- A data embedding layer is designed according to the characteristics of bearing vibration signals under time-varying speed, and the multihead attention mechanism of a transformer is fully used to extract bearing fault features under time-varying speed, so as to achieve high-dimensional feature fusion with speed pulse signals. Through adjusting parameters, the effectiveness of the fusion method proposed in this paper for bearing fault diagnosis under time-varying speed is proved in comparison with other deep learning models.
- The degree of influence of the speed fluctuation on the model classification is quantitatively analyzed, and the effectiveness of the model for bearing fault classification under the actual variable speed is verified by experimental data.
- From the perspective of model interpretability, the mechanism and principle of each module of the fusion model in extracting bearing time-domain signal features are analyzed, and the potential application of the deep learning model in multisensor signal fusion is explored.

2. Vibration–Speed Data Fusion Network

This section introduces the background and basic module of the vibration–speed data fusion network proposed in this manuscript, including the transformer with multihead self-attention, the SAE, and its network framework as shown in Figure 2. Bearing vibration signals under time-varying speed not only have local impact characteristics but also have mutual dependence among them. Therefore, this section adopts the multihead self-attention mechanism in the transformer model [22], which is more efficient to deal with time series problems and extract vibration features. According to the signal characteristics of bearing vibration in the time domain, this section designs an embedding layer, feature extraction layer, and classification layer for one-dimensional sequential data adapted to model training to achieve the end-to-end intelligent diagnosis of bearing faults.

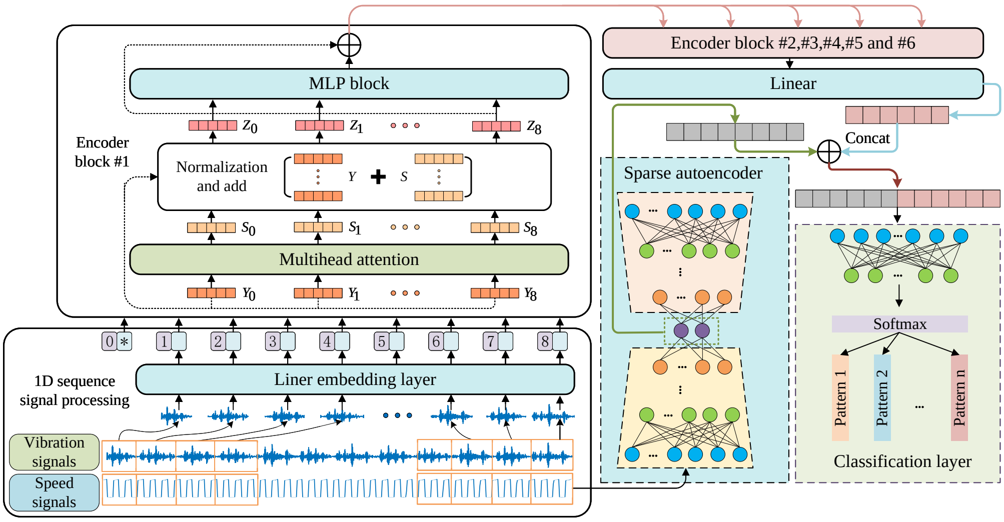


Figure 2. The framework of VSF-ST.

2.1. One-Dimensional Sequence Embedding Layer

In order to segment a long sequence of vibration signals under variable speed and improve the efficiency of network calculations, it is necessary to establish a 1D data-embedding layer. The segment was divided into $X = \left[x_s^1, x_s^2, x_s^3, \ldots, x_s^N\right] \in R^{N\times S}$. Among them, X is a subsequence of the collected data, N is the number of subsequences being segmented, and S is the length of the segmented fragment. A different segmentation means different values of N and S. In order to ensure that the dimensionality of the segmented data did not change after the embedding layer, a linear layer with weight W was set in this paper, which was equivalent to projecting X as a D-dimensional vector:

$$X' = \left[x_0, \left[x_s^1, x_s^2, x_s^3, \ldots, x_s^N\right] \cdot W\right], W \in R^{S\times D} \tag{1}$$

where x_0 is a D-dimensional vector, also known as a classification token, which plays a role similar to global pooling and aims to enhance the generalization performance [23]. The position of the data is unchangeable in a physical sense after splitting, otherwise the sequence would be chaotic, so a one-dimensional position encoding $E_{pos} \in R^{(S+1)\times D}$ that can be learned by the network was added in this paper to extract the relative positions of the vibration signals. Both relative position coding and absolute position coding can enhance the generalization performance, and the difference between these two effects is very small [24], so a random initialized weight matrix was adopted as the position coding in this paper. The output of the linear embedding layer can be presented as the following formula:

$$Y = \left[x_0; \left[x_s^1, x_s^2, x_s^3, \ldots, x_s^N\right] \cdot W\right] + E_{pos} \in R^{(S+1)\times D} \tag{2}$$

2.2. Encoder Block Based on Multihead Attention

In deep learning, convolution blocks are mostly used to extract local features, which are not suitable for all temporal dependence models. At the same time, under the high-sampling sensor, the convolution operation is easily disturbed by noises and redundant data in vibration signals. The multihead self-attention mechanism in the transformer can effectively replace the convolution operation, and at the same time has an efficient filtering ability in processing the highly redundant information [25]. In Figure 3, the modules of the multihead self-attention (MSA) mechanism used in this paper are shown.

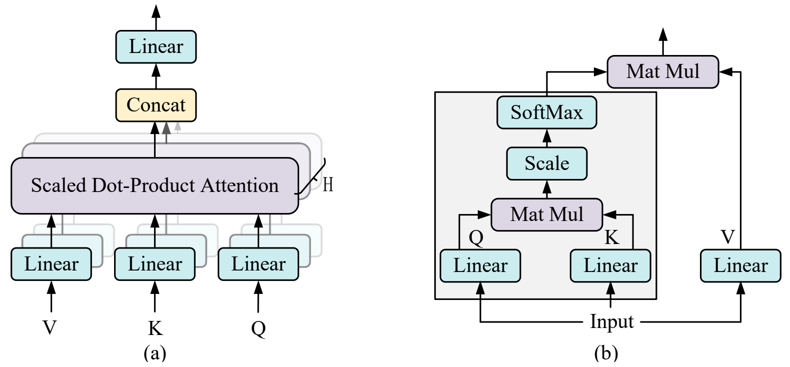


Figure 3. (**a**) Multihead self-attention module. (**b**) Scaled dot-product attention.

In Figure 3, Q, K, and V, denote the query vector matrix, key vector matrix, and value vector matrix generated by the linear mapping of the output vector Y of the embedding layer, respectively. The correlation operation between Q and K was divided by a scaling factor $\sqrt{d_k}$. The value obtained was normalized by the softmax function and finally multiplied by V to obtain the output of self-attention. The specific calculation was as follows:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)\text{V} \tag{3}$$

The adopted MSA module contained *h* self-attention blocks. This method can measure the effective features in the data from *h* dimensions, avoiding the coincidence caused by single-headed self-attention. At the same time, matrix stitching and splitting were used in the multihead operation to realize the parallel operation of the MSA module, which reduced the complexity of the model operation. As many stacked fully connected layers were adopted in the transformer, the Gaussian error linear unit (GeLU) activation function was used in the feedforward neural network of the encoder block in this manuscript for rapid convergence, which effectively avoided the problem of gradient disappearance after using multiple encoder blocks. The GeLU activation function is as follows:

$$\text{GeLU}(x) = 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + 0.044715x^3\right)\right]\right) \tag{4}$$

2.3. Sparse Autoencoder

The SAE with symmetric network structure has a strong high-dimensional feature extraction capability and unsupervised learning capability, which is more suitable for extracting the characteristics of speed pulse signal. Its sparsity is mainly based on the added sparse penalty factors so that the hidden layer of the network is in a state of high inhibition and low activation. In this way, the features automatically extracted from samples are high-dimensional and sparse, which consume less memory cost and are more generalizable than the features extracted by the general nonlinear mapping. The sparsity is mainly reflected in the fact that when the output of the hidden layer is -1 through the activation function tanh, the node is in the inhibitory state. This is achieved by adding a sparse penalty factor:

$$\text{KL}(\rho||\widehat{\rho}_j) = \rho \log \frac{\rho}{\widehat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\widehat{\rho}_j} \tag{5}$$

where KL is the scatter calculation, a measure of the difference in distribution between two variables with ρ as the mean and one with $\widehat{\rho}_j$ as the mean. ρ is the sparsity parameter, where $\rho = 0.05$, which means that the average activation of neurons in the hidden layer is 0.05. The practical meaning of KL is that when the value of $\widehat{\rho}_j$ is close to ρ, the value of KL is the smallest, and when the value differs greatly, the value of KL tends to infinity. So adding the KL term to the loss function can well reflect the sparsity.

2.4. Vibration–Speed Data Fusion Network Training Process

In this manuscript, an adaptive fusion network model based on VSF-ST for bearing time-varying operating conditions was developed for the purpose of deeply extracting the nonstationary and nonlinear features from signals under variable speed. From the perspective of a multisource data network fusion, this paper used a transformer network with a high parallel efficiency and information fusion efficiency and used an unsupervised SAE network with a high generalization capability to perform high-dimensional feature fusion on the collected bearing vibration signals and rotational speed pulse signals, respectively, so as to realize the adaptive diagnosis of bearing faults under time-varying speed operating conditions. As shown in Table 1, the corresponding data were used in the model.

Table 1. Data usage of VSF-ST adaptive network in operating conditions.

Dataset	Pretraining Phase	Fusion Training Phase	Testing Phase
Vibration signal (signal 1)	×	√	√
Speed signal (signal 2)	√	√	√
Label	×	√	×

The flow chart based on the VSF-ST adaptive model is shown in Figure 4, and the specific steps are as follows:

- Step 1: Install the vibration sensor and speed sensor at the motor drive end; the collected bearing vibration amplitude signal $\{X_i\}_{i=1}^{d}$ is marked as signal 1 and the direct tacho signal (typical square wave) $\{S_i\}_{i=1}^{d}$ is marked as signal 2, where d is the number of samples. In dividing the data, both sensors are sampled at the same frequency, so the samples intercepted by using a fixed-length sliding window for both signal 1 and signal 2 are the working signals in the same state. Meanwhile, the same random number seed is set for both signal 1 and signal 2 when disrupting the samples to ensure that the same operating condition features are extracted during the network fusion training. To augment the dataset, the samples are intercepted in the sliding window so that there is partial overlapping between adjacent samples.
- Step 2: After dividing the data set for two channels, the tacho pulse signal samples from signal 2 are first fed into the SAE network for pretraining. The samples at this stage do not contain labels because the unsupervised learning capability of the SAE can automatically extract the deep representations of the tacho pulses. After completing a set number of iterations, the SAE network has learned the potential spatial representations of the tacho pulse signals during the compression-recovery of the data, and finally all the hidden layer parameters that can represent these potential spatial representations are frozen for subsequent network fusion training.
- Step 3: The vibration samples from signal 1 are fed into the transformer network constructed in this paper based on 1D sequential signals, while the tacho pulse signal samples from signal 2 are fed into the SAE network with the hidden layer parameters frozen. Note that the hidden layer here is only taken as the SAE encoder layer, because the compressed features obtained after the coding layer are more helpful for the network fusion training. After signal 1 and signal 2 have completed forward propagation in their respective channels, the high-dimensional feature outputs (B, X^d) and (B, S^d) are obtained in the last linear layer of the two networks, respectively, where B is the number of batches of data and d is the vector dimension. These two sets of vectors are spliced to output vector $(B, cat(X^d, S^d))$. Finally, the cross-entropy calculation is done in the classification layer and the backpropagation of errors is performed. In the iterative training of the network, only the parameters of the input layer of the SAE network are updated. At the end of the training, results are finally obtained by inputting the testing dataset.

As shown in Table 2, the VSF-ST adaptive framework parameters were configured. Among them, the SAE and transformer contained a large number of fully connected layers, which increased the volume of the network model to a certain extent. In order to speed up the network training, GeLU was used as the main activation function. Compared to ReLU, GeLU is a more mainstream activation function in transformer models. While ReLU has a strong sparsity and lower complexity, it is more likely to lead to neuron values in attention-based mechanisms computing the "necrotic" state [26], which causes some features to be impaired in transmission. Therefore, it is not suitable for network structures with fully connected layers. The classification layer adopted a simple three-layer fully connected network structure, so as to avoid overfitting. From the change of data dimension in the linear embedding layer, we can see that the classification token was equivalent to adding an extra dimension to the original data dimension. In the final encoding block, the output of the classification token was equivalent to the features extracted by the whole transformer network, so the classification token acted as a global pooling.

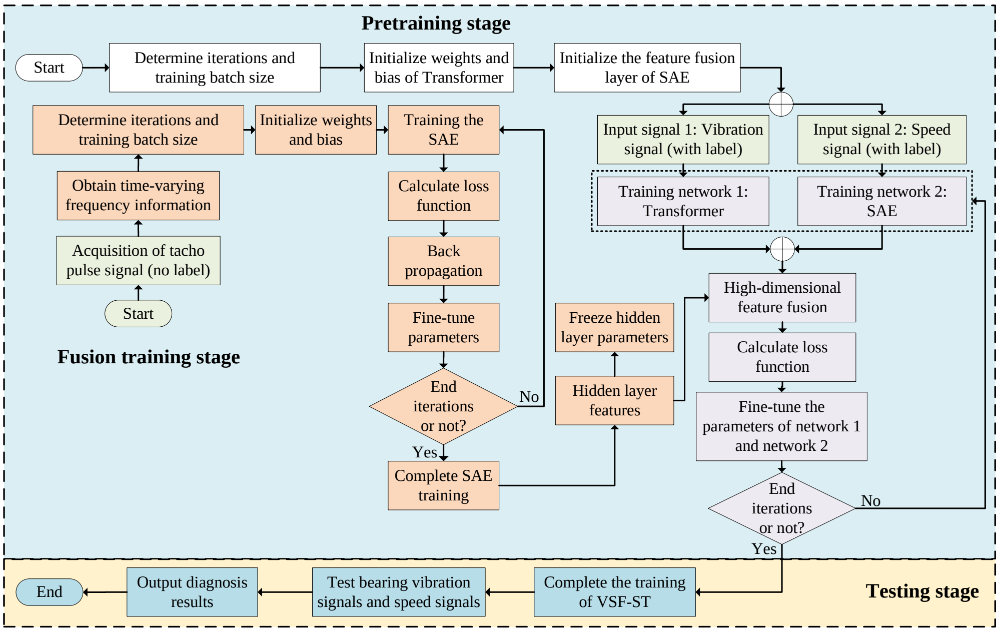

Pretraining stage
Start
Determine iterations and training batch size
Initialize weights and bias of Transformer
Initialize the feature fusion layer of SAE
Determine iterations and training batch size
Initialize weights and bias
Training the SAE
Input signal 1: Vibration signal (with label)
Input signal 2: Speed signal (with label)
Obtain time-varying frequency information
Calculate loss function
Training network 1: Transformer
Training network 2: SAE
Acquisition of tacho pulse signal (no label)
Back propagation
Start
High-dimensional feature fusion
Fine-tune parameters
Freeze hidden layer parameters
Calculate loss function
Fusion training stage
End iterations or not?
No
Hidden layer features
Fine-tune the parameters of network 1 and network 2
Yes
Complete SAE training
End iterations or not?
No
Yes
End
Output diagnosis results
Test bearing vibration signals and speed signals
Complete the training of VSF-ST
Testing stage

Figure 4. Flowchart of proposed VSF-ST model for time-varying speed bearing fault diagnosis.

Table 2. Network structure parameters of the proposed VSF-ST model.

SAE Architecture	Network Structure Parameters	Input Size	Output Size
Encoder–decoder	$Linear \begin{bmatrix} 2048-1024-512 \\ 128 \\ 512-1024-2048 \end{bmatrix} (GeLU)$	(1, 2048)	(1, 128)
Transformer architecture			
Linear embedding layer	Class token (1, 64) Position code (33, 64)	(1, 2048)	(33, 64)
Encoder block × 8	$\begin{bmatrix} \text{MSA } Linear\ (64, 192) \\ \text{MLP} - Block \begin{matrix} Linear\ (64-256 \\ -64)\ (GeLU, LN) \end{matrix} \end{bmatrix} \times 6$	(33, 64)	(33, 64)
Linear	$Linear\ (64, 128)$	(1, 64)	(1, 128)
Classification layer	$Linear\ (256-128-n)\ (Softmax)$	(1, 256)	(1, n)

3. Dataset Introduction

In this paper, two open-source datasets were used to verify the effectiveness of the VSF-ST model for bearing fault diagnosis under variable speed. Two datasets were obtained from the laboratories of Xi'an Jiaotong University and the University of Ottawa, Canada, which simulate two operating conditions, "start-run-brake" and time-varying speed of the bearing, respectively, to verify the effectiveness of the model more adequately in the experiments. These two datasets are mainly affected by the change of speed, while the influence of load is relatively small. The details of the division method and the setup in the experiments for these two datasets in this paper are described below.

3.1. SQV Bearing Dataset from Xi'an Jiaotong University

As shown in Figure 5a, the bearing NSK6203 was artificially damaged by a fault simulation test bench for the outer and inner rings, and the bearing vibration acceleration signal and bearing speed pulse signal were collected by the data acquisition instrument during the whole "start-run-brake" cycle, which was adopted from the SQ (Spectra Quest) fault simulation test bench, hereinafter referred to as the SQV dataset [27,28]. As shown in Figure 5b, according to the degree of damage, a total of three fault levels were identified, with numbers 1, 2, and 3 indicating mild, moderate, and severe damage, respectively, and their fault details are shown in Table 3.

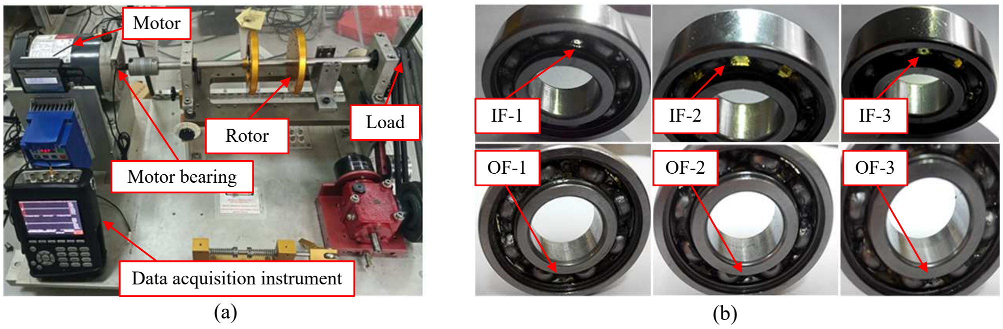


Figure 5. (**a**) SQV dataset acquisition test bench. (**b**) Fault simulation bearings.

Table 3. Division of SQV dataset.

Data Label	Fault Location	Damage Level Label	Damage Area (mm^2)	Damage Depth (mm)
IF-1	Inner race	1	4	0.5
IF-2	Inner race	2	8	2
IF-3	Inner race	3	12	4
OF-1	Outer race	1	4	0.5
OF-2	Outer race	2	8	2
OF-3	Outer race	3	12	4
NC-1	Health	None	None	None

For all health and fault bearings, the sampling time was 15 s and the sampling frequency was 25,600 Hz. Throughout the sampling period, the data set recorded the signals generated throughout the process of starting the bearing at 0 rpm, accelerating to a smooth speed of 3000 rpm, and finally decelerating to a speed of 0. As shown in Figure 6, the speed variation trend of the bearing was fitted by the speed pulse signal. Since the knob was operated artificially to control the rotational speed, the instantaneous values of rotational speed change were not guaranteed to be identical between the data sets, but the overall trend of rotational speed change was the same.

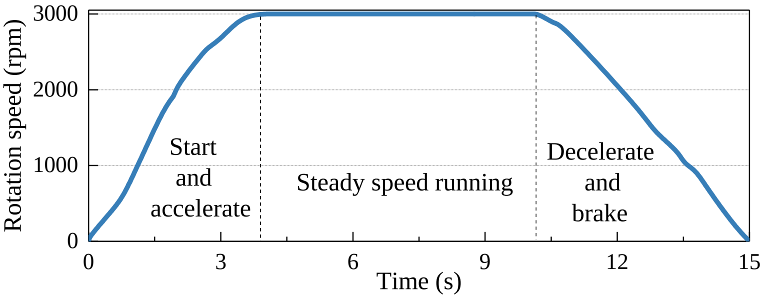


Figure 6. Graph of bearing speed variation.

3.2. Bearing Dataset from the University of Ottawa

The dataset from the University of Ottawa Laboratory [29] contains different fault conditions at four types of time-varying speeds. As shown in Table 4, the operating speed in this dataset changed all the time during the sampling time, and the variation was divided into four cases: acceleration (A), deceleration (D), acceleration and then deceleration (A-D), and deceleration and then acceleration (D-A). In Figure 7, the variation of the operating speed under the four operating conditions are shown, and the speed is indicated by the bearing working frequency. The bearing health status was divided into three cases: normal clear, inner race fault, and outer race fault, which are indicated by the labels NC, IF, and OF. Each sample was collected for a total of 10 s with a sampling frequency of 200 kHz, and three repetitions of the experiment were conducted to collect data under each operating condition.

Table 4. Division of the University of Ottawa dataset.

Operating Condition Label	Variation of Bearing Speed	Fault Label
A	Acceleration	NC & IF & OF
D	Deceleration	NC & IF & OF
A-D	Acceleration and then deceleration	NC & IF & OF
D-A	Deceleration and then acceleration	NC & IF & OF

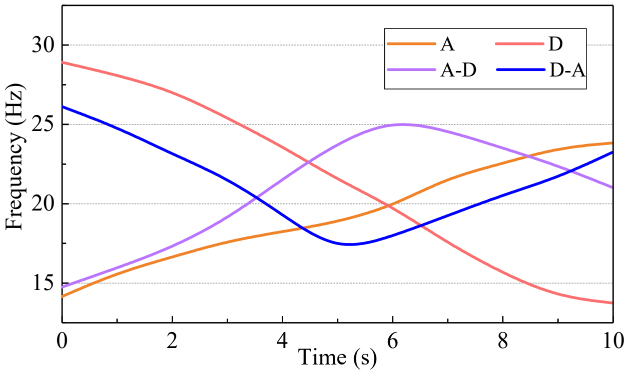


Figure 7. Rotational frequency variation in the University of Ottawa dataset for the time-varying state of the bearing speed.

3.3. Data Processing

These two datasets used in this paper were sampled at different times, 10 s and 15 s, respectively. If each divided sample was guaranteed to contain information about the data within at least one rotation cycle of the bearing, only 100 to 500 samples subsets could be obtained for each type of health condition, so the two open-source datasets in this paper used data augmentation as shown in Figure 8 to increase the training sample set and testing sample set.

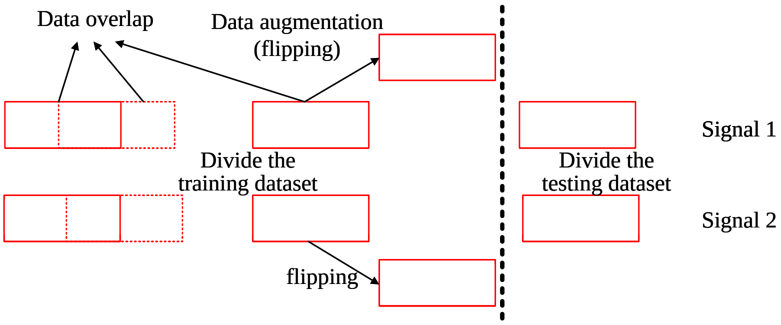
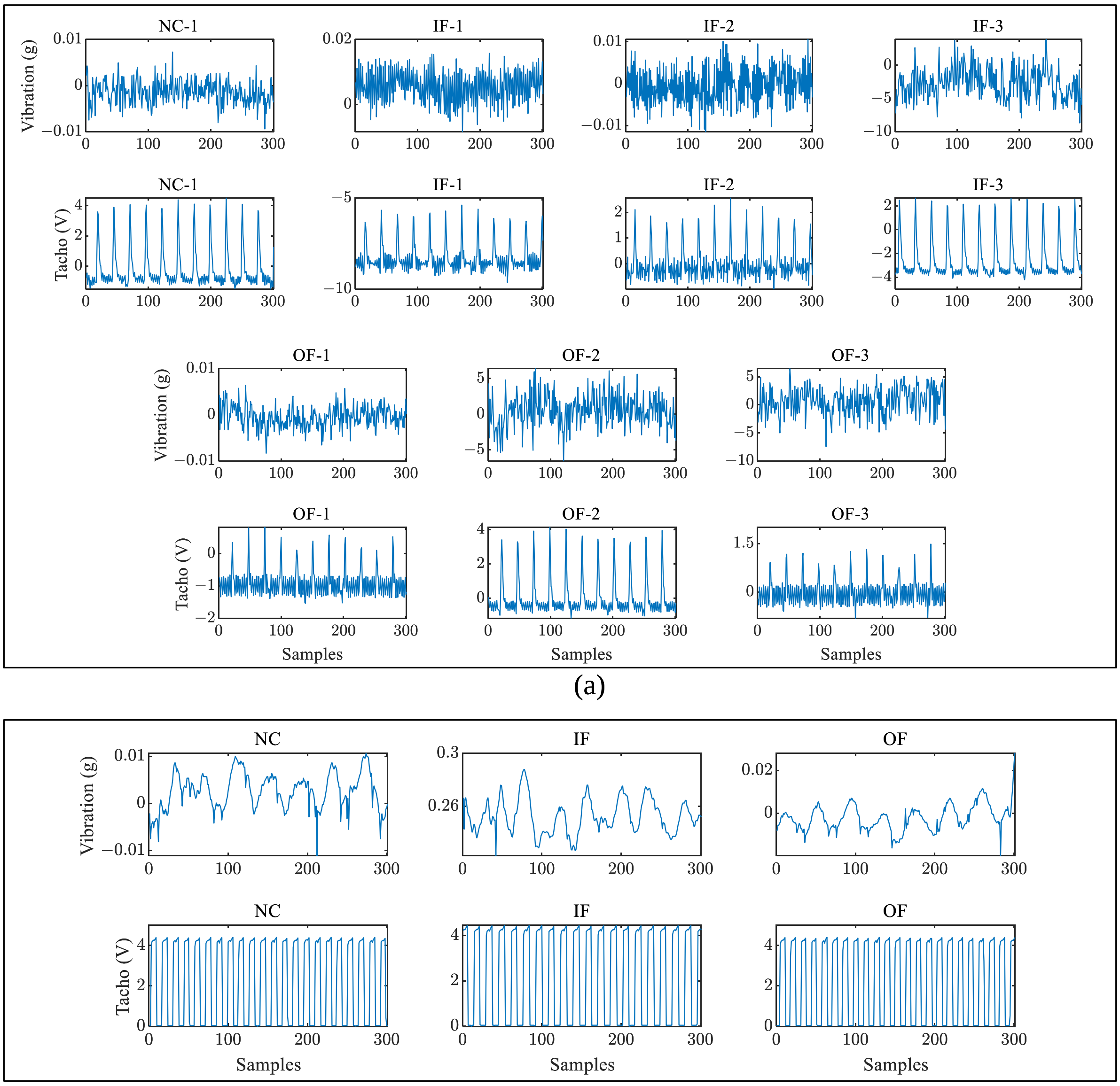


Figure 8. Method of bearing data augmentation for variable speed condition.

The bearing vibration signal (signal 1) and the bearing speed pulse signal (signal 2) were acquired from the data acquisition, respectively. Since signal 1 and signal 2 were acquired synchronously at the same sampling frequency, a sliding window with the same length was used for subsample interception. The interception time for each subsample was 1 s. Then, the first 80% of the data of the subsample were enhanced by data overlapping acquisition and data flipping, and they were divided into a training set, and the next 20% of the data were divided into a testing set in the same way. Among them, 50% of the sample length was overlapped. This ensured that there was no overlap of data between the training and testing sets, and also allowed the training data to be collected for the bearings at each speed stage. The bearing vibration signal amplitude increased during acceleration and decreased during acceleration. Similar to the image translation invariance in image data augmentation, the data characteristics of acceleration and deceleration mapped to each other could be simulated to some extent by flipping the bearing sample data. As shown in Figure 9, the image of part of the sample data after data augmentation is shown.



(a)

(b)

Figure 9. Part of the dataset via augmentation: (**a**) SQV dataset; (**b**) University of Ottawa dataset.

4. Experimental Results and Discussion

In this section, the optimal structural parameters of the VSF-ST network model for fault diagnosis under variable speed, as well as the performance comparison between the proposed model and other network models are experimentally investigated, including a visual interpretation of the VSF-ST model in the process of classification. According to the sampling differences between the SQV dataset and the University of Ottawa dataset, the data batches were set to 32 and 64, respectively, and the length of each data sample was 2048 and 8096. The data were directly fed into the network model for training without any preprocessing. The gradient descent method was a stochastic gradient descent with a learning rate of 0.001, a momentum factor of 0.9, and a weight decay factor of 5×10^{-5}. The deep learning environment used consisted of Pytorch 1.8.1, CUDA 10.1, and Python 3.8.

4.1. Effect of Network Model Parameters on Diagnostic Results

The VSF-ST model proposed in this paper had a large number of network parameters, and the network parameters such as the number of encoder blocks, the number of head and the length of the data segmentation in the transformer network, which extracted the key vibration characteristics under variable speed conditions, had an impact on the network performance. A box plot of the fault diagnosis effect of the VSF-ST model on the SQV dataset is shown in Figure 10, where each set of experiments was repeated 20 times. In Table 5, the extent to which changes in the network hyperparameters affected the results is listed using multiple metrics. L represents the length of a single sample, and L/N represents the subsamples of vibration signals that were further divided into N blocks for processing in the process of MSA calculation. A reference group was set up for the experiment, and the remaining comparison experimental groups were used to change the network hyperparameter values by means of control variables. It can be seen from Table 5 that the larger the number N of values for the segmentation of the sample signal, the higher the accuracy, but at the same time the computational complexity increased sharply with the increase of N. From $N = 1$, $L/N = 2048$ to $N = 256$, $L/N = 8$, the accuracy increased from 91.45% to 95.18%, but the computational complexity increased from 4.21 M (FLOPs) to 406.93M (FLOPs), the complexity increased by nearly 97 times, and the number of parameters only decreased by 18%. One of the reasons for the decrease in the parameter amount was that the dimensionality of the segmented data and the dimensionality of the position encoding were reduced. For the transformer, it was the number of heads and the number of encoder blocks that caused the change in the number of parameters. The multihead self-attention mechanism was the key of the transformer model, where the accuracy reached the best value when the number of multiheads was $h = 4$. It can be seen from Equation (3) that the increase in the number of heads also caused a sharp increase in computational complexity, as well as a significant increase in the number of network parameters. The accuracy basically reached a peak of 95.18% when the number of encoder blocks was increased to $depth = 6$, after which a further increase in network depth did not cause overfitting and the accuracy was always maintained at the peak. As mentioned in Section 2, the transformer network contains a large number of residual connections and dropout methods, which, together with the setting of the weight decay coefficients during gradient descent, can prevent overfitting phenomena. Therefore, under the comprehensive consideration of accuracy and operation cost, the network parameters were: $N = 128$, $L/N = 16$, $h = 4$, and $depth = 6$.

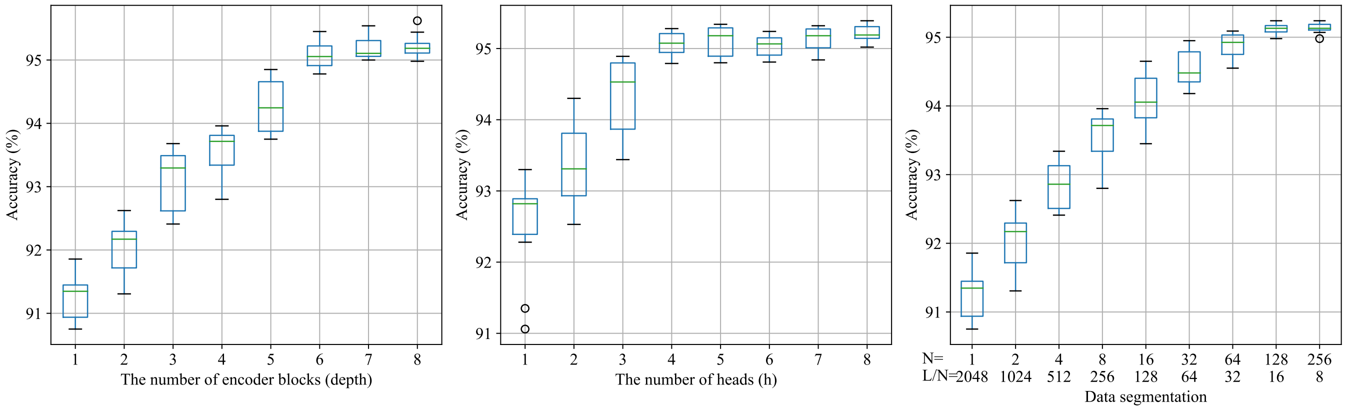


Figure 10. Box plot with different network structure parameters.

Table 5. Influence of the VSF-ST structure hyperparameters.

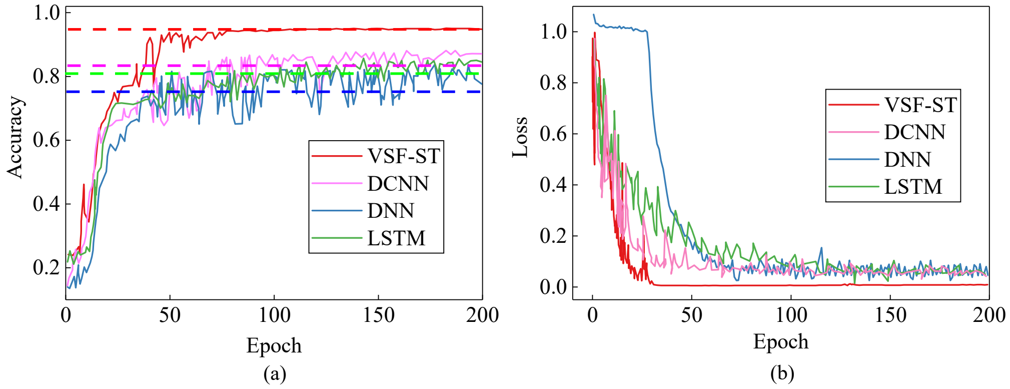
Label	Hyperparameters				Accuracy	FLOPs	Parameters Number
	N	*L/N*	*h*	*Depth*			
Baseline	256	8	4	6	95.18%	406.93 M	1.59 M
c	128	16			95.18%	205.27 M	1.59 M
	64	32			94.85%	104.85 M	1.59 M
	32	64			94.47%	53.72 M	1.61 M
	16	128			94.21%	28.12 M	1.62 M
	8	256			93.78%	15.69 M	1.64 M
	4	512			92.81%	9.56 M	1.73 M
	2	1024			92.30%	5.56 M	1.86 M
	1	2048			91.45%	4.21 M	1.94 M
b			1		92.74%	152.61 M	0.71 M
			2		93.52%	207.41 M	0.81 M
			3		94.67%	268.16 M	1.48 M
			8		95.14%	812.36 M	1.78 M
a				1	91.53%	69.37 M	0.31 M
				2	92.24%	139.61 M	0.54 M
				3	93.41%	275.83 M	0.72 M
				8	95.19%	561.27 M	2.38 M

4.2. Comparisons with Other Network Models

To demonstrate the advancement of the VSF-ST model for fault diagnosis in variable speed conditions of bearings, three models were added for comparison, which were a deep convolutional neural network (DCNN), a long short-term memory (LSTM) network, and a fully connected deep neural network (DNN). As shown in Table 6, the hyperparameter settings of the three networks are shown. To ensure the fairness of the comparison experiments, all three networks used the same gradient descent algorithm and the same relevant parameter settings as the VSF-ST network. The first comparison was the classification accuracy and loss value based on the SQV dataset. From Figure 11, it can be seen that all four models basically stabilized by the 50th iteration, and the proposed model in this paper had a higher classification accuracy and classification stability than the other three network models. The maximum accuracy of the other models only reached about 85%, and the accuracy was still fluctuating after 50 iterations, especially the DNN, which had the largest fluctuation range. In the loss-value descent curve, the VSF-ST could quickly extract the local and global features of bearing signals and complete the update of network parameters by relying on its attention mechanism. For the DNN, which also contained a large number of fully connected layers, it took nearly 30 iterations for the loss value to start decreasing in the absence of an attention mechanism.

Table 6. Hyperparameters of the networks in the comparison experimental group.

Model	Model Hyperparameters
DCNN	Conv1d; kernel numbers: 16, 32, 64, 128; kernel size: 1 × 15, 1 × 3, 1 × 3, 1 × 3 2nd layer maxpooling (2, 2), 4th layer maxpooling (4); linear: 128 × 4, 256, 128
LSTM	[LSTM (64, 64, tanh), dropout (0.1)] × 5; linear: 128, 128
DNN	Linear: 1024, 256, 256, 256, 128, 128; dropout (0.1)



(a)

(b)

Figure 11. Comparison experiments based on SQV dataset: (**a**) accuracy curve; (**b**) loss curve.

As shown in Figure 12, the same four models were used for comparison on the University of Ottawa dataset. From the figure, it can be seen that the VSF-ST model could achieve almost 100% classification on the University of Ottawa dataset, and the standard deviation was also the lowest among the 20 repetitions. Although the CNN and LSTM networks could occasionally reach the upper accuracy limit of nearly 98%, their models had poor robustness and generalization when facing the classification problem under variable speed conditions. This was because the random initialization of the network and the random arrangement of the training samples had a certain impact on the feature extraction effect, resulting in a large fluctuation of the accuracy in multiple experiments.

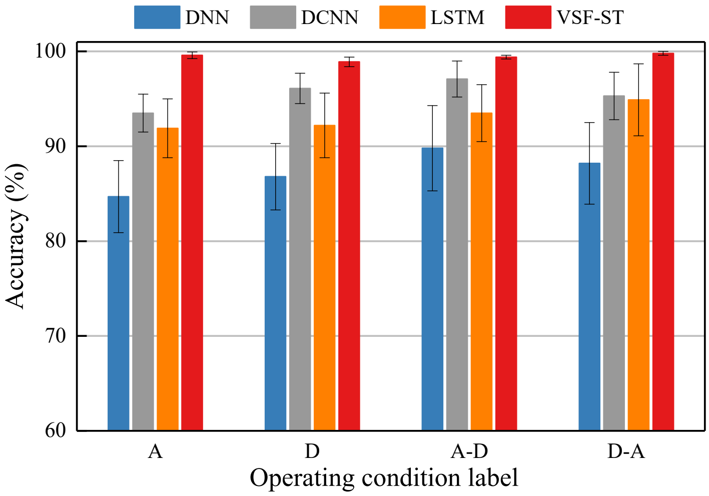


Figure 12. Comparison experiments based on the University of Ottawa dataset.

4.3. Analysis of the Influence of Rotational Speed Fluctuation

Rotational speed fluctuation was the main factor affecting the classification accuracy of the model. In order to further explore the influence of rotational speed on the generalization of the model, the cross-domain diagnostic classification experiments of four models on different variable speed operating conditions were set up, and the results are shown in Figure 13, where A → D indicates that the working condition in the training set is acceleration and the working condition in the testing set is deceleration. In the cross-domain classification, VSF-ST performed the worst generalization in group A → D with only 95% classification accuracy, but in group D → A , the result was close to 100%. This was probably because the deceleration frequency range in the University of Ottawa data set was from 37.5 Hz to 13.5 Hz and the acceleration frequency range was from 14.7 Hz to 24.2 Hz, so the testing set in group A → D had a higher frequency range than the training set. This phenomenon was also reflected in the other three models. In the cross-domain diagnostic classification task, the generalization performance of VSF-ST outperformed the other models in each task, due to the efficient fusion of vibration signals and rotational speed pulse signals in VSF-ST, which made the fault feature extraction under variable speed conditions more stable.

In contrast, the DCNN, LSTM, and DNN all showed performance degradation in the cross-domain diagnostic classification task and could not perform the best performance of the network in the face of signals at unknown rotational speeds. From the comparison experiments between the SQV and the University of Ottawa datasets, it can be seen that VSF-ST showed the superiority of the model compared with traditional deep learning algorithms DCNN, LSTM and DNN in the whole process of "start-run-brake" and the time-varying speed of the bearing, and also showed the generalization performance of the model in the cross-domain diagnostic classification task.

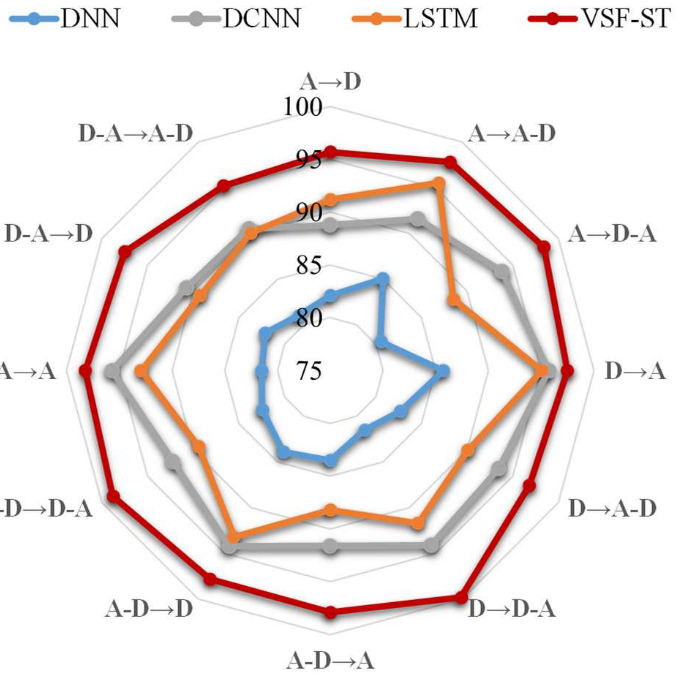
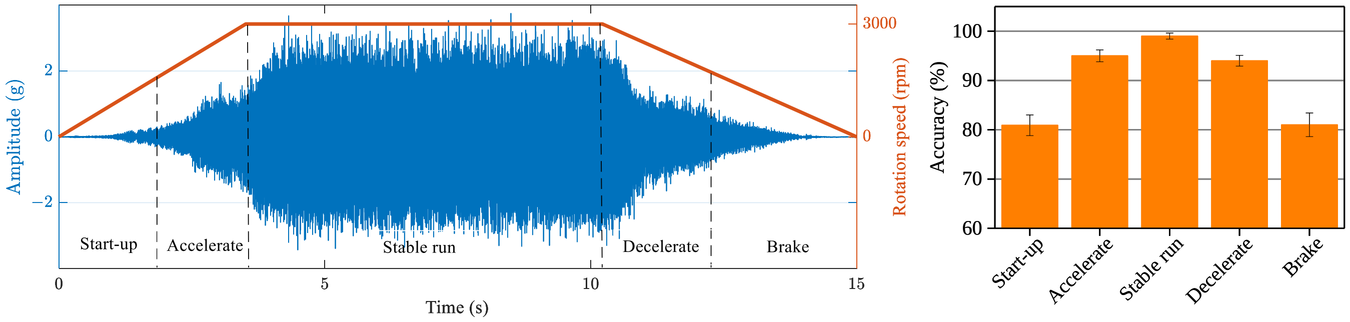


Figure 13. Comparison experiment of variable-speed cross-domain diagnosis classification based on the University of Ottawa dataset.

From Figure 14a, it can be seen that the amplitude of the bearing vibration signal was proportional to the speed, so the SQV data were further divided into five stages according to the size of the speed interval: start-up, accelerating, stable running, decelerating, and braking. The results of using VSF-ST to do fault classification for each of these five phases are shown in Figure 14b. In the stable operation stage, the accuracy basically reached 100%; in the acceleration and deceleration stages, the accuracy dropped to about 95%; and in the start-up and braking stages, the accuracy could only reach about 80%. This indicated that it was most challenging to accurately complete the health status assessment at the early stage of bearing start-up, but when the signal monitoring time was appropriately extended, a good level of diagnostic classification could be achieved using the model in this paper.



(a) (b)

Figure 14. Fault diagnosis based on SQV dataset by speed interval: (**a**) division of speed period; (**b**) the diagnosis results of each time period.

4.4. Visual Interpretation of the Model

To further investigate the process of effective feature mining in VSF-ST for fault classification under variable speed, a visual analysis of the operation mechanism of the attention mechanism in the VSF-ST network is presented. The bearing vibration signal is the main basis for the VSF-ST model to do fault classification of bearings at variable speed, and the speed signal plays the role of high-dimensional fusion with the vibration signal in the model. As shown in Figure 15, a set of data from the testing set was forward-propagated once through the trained VSF-ST model to obtain the product feature maps of the Q and K matrices of the six-layer encoder block, where the horizontal represents the individual multihead self-attention feature matrix maps and the vertical represents the multihead self-attention feature matrix maps of each layer. From Figure 15, it can be seen that the first layer encoder block was still mainly in a kind of initialization state, and each segment of the bearing vibration signal was given a certain range of initial values. From Equation (3), it can be found that the product of the Q and K matrices was a measure of correlation, and this value was multiplied with the V matrix and then the scaling dot product completed the key calculation of the attention mechanism. This process has begun to reflect in the figure of the second layer's head 3 and head 4 feature matrix. Under the action of the multihead self-attention mechanism, the bearing vibration signals were gradually assigned weights to some key segments through a layer-by-layer feature mining of the six-layer encoder blocks, while each head also extracted some important features from different dimensions.

As the VSF-ST input was a one-dimensional vibration signal, the vibration sequence could be restored by one-dimensional splicing after each sample was segmented with position encoding and local features were extracted. The output signal of the first layer and the output signal of the sixth layer were compared and analyzed as follows: first, the testing set data were propagated forward and the output matrix of the sixth layer and the first layer was reduced to a one-dimensional sequence, and then 0.5 times the maximum value of the amplitude change was set as the change threshold, and finally the red dots marked the points that exceeded the change threshold. Results are shown in Figure 16. From the figure, it can be seen that there were many points labeled with normal bearing vibration data, and fewer points labeled with various types of fault bearing, indicating that the features extracted in the VSF-ST network were different for normal and fault bearings. Each layer of the network did not extract some characteristic points or characteristic fragments on normal bearings, while the network focused on fault bearings to extract some characteristic vibration shock fragments formed due to the fault points. This phenomenon is difficult to be reflected in other deep learning algorithms, so model interpretability is also a major advantage of VSF-ST.

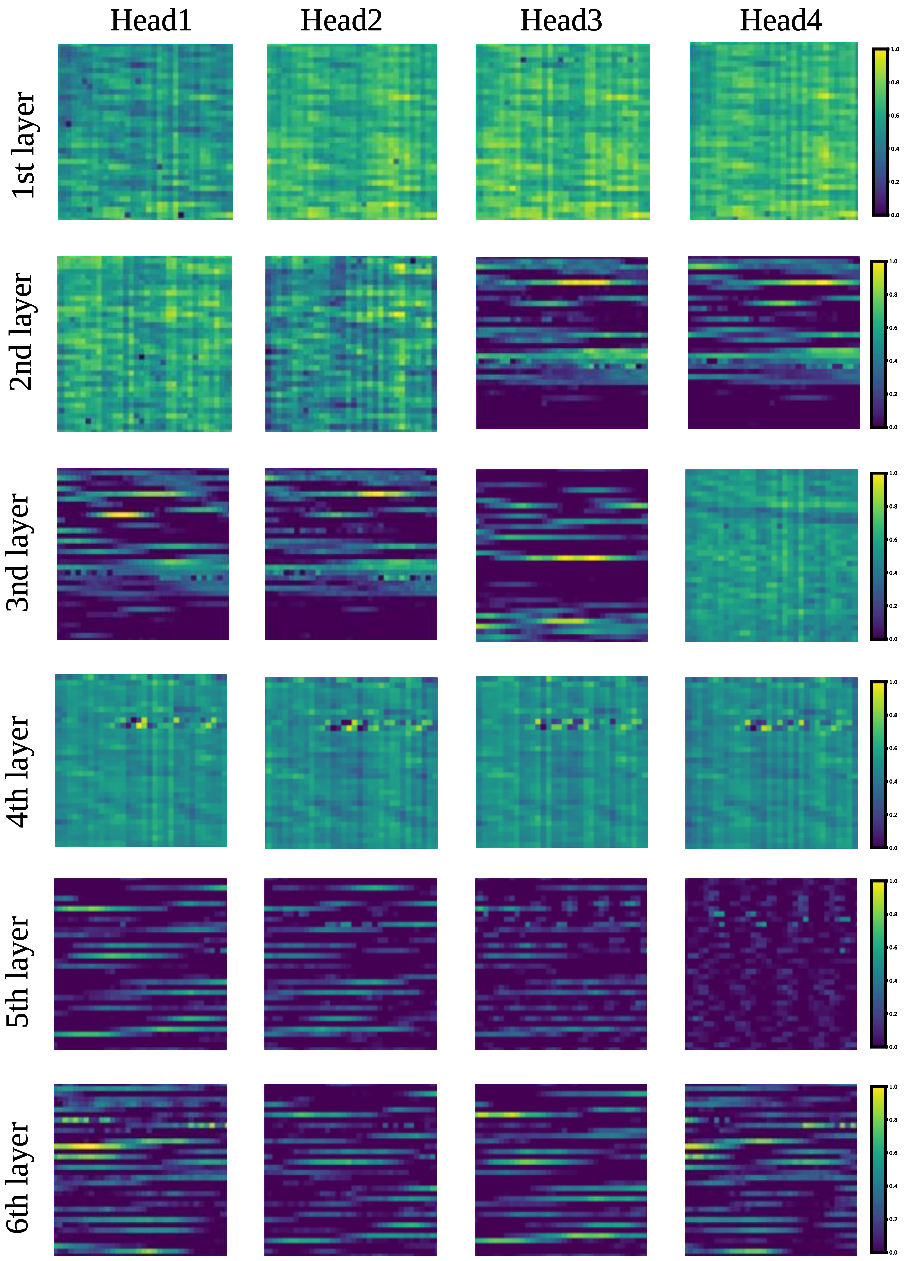


Figure 15. Visualization of Q and K feature matrix in the multihead self-attention mechanism.

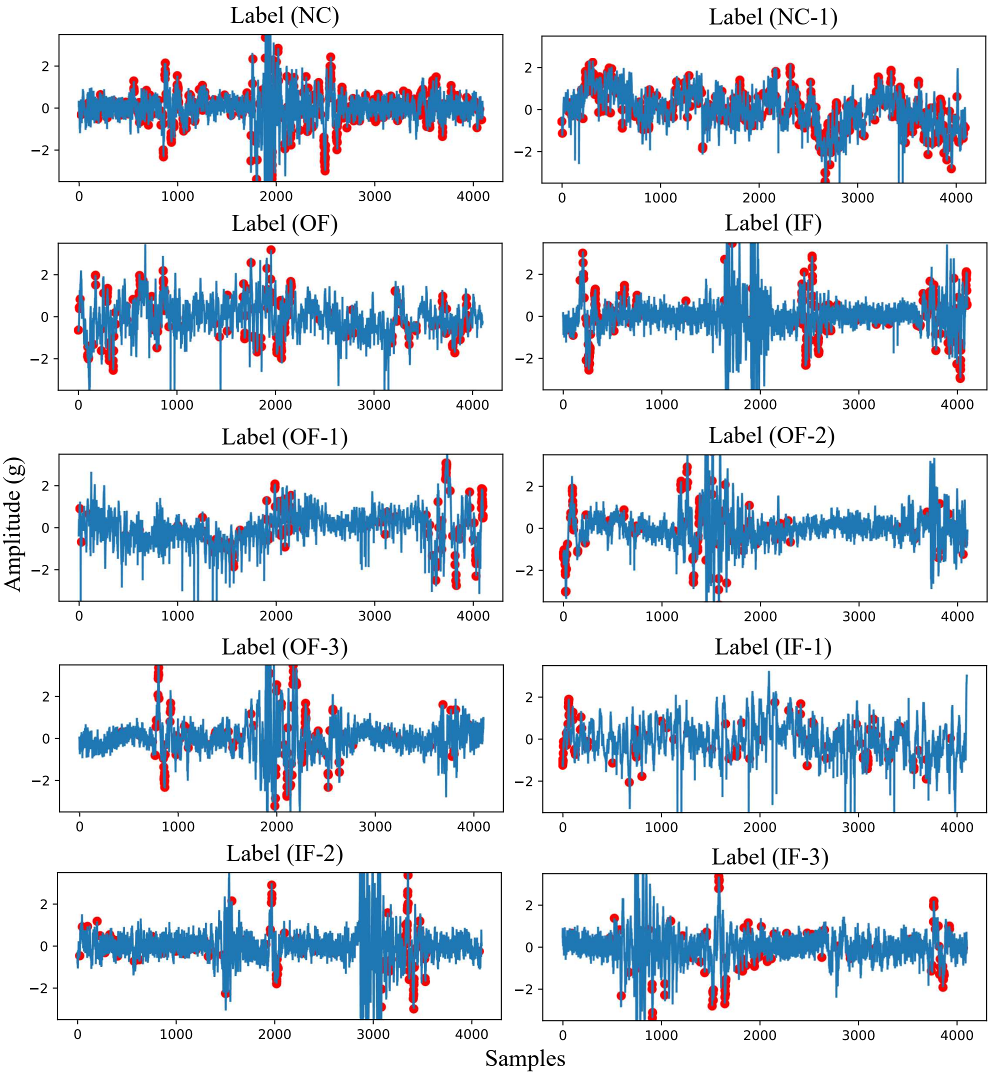


Figure 16. Visualization of the amplitude changes of the bearing vibration signal.

5. Conclusions

In order to solve the problem of insufficient generalization under time-varying speed by the intelligent fault diagnosis method, a VSF-ST adaptive network for variable speed conditions based on the dual-channel signal processing of bearing vibration signal and bearing speed signal was proposed in this paper, and the effectiveness of the model was verified on the SQV dataset and the University of Ottawa dataset. By exploring the appropriate hyperparameters, the visualization and interpretability of the model feature extraction process, and the comparison experiments with other network models, the following conclusions are drawn:

- Under the effects of the VSF-ST model's position coding and multihead attention mechanism, a small and suitable segmentation length could maximize the performance of the model for bearing vibration signals under variable speed conditions. Increasing the number of heads and the number of layers of coding blocks could also increase the performance of the model, but under the comprehensive consideration of the computational complexity and the number of parameters, four heads and six layers of coding blocks could achieve the requirement of an accurate classification under variable speed conditions, with an average classification accuracy of 95.18% in the SQV dataset and 99.85% in the University of Ottawa dataset.

- The advantages and effectiveness of the proposed model in fault diagnosis were verified by experiments under the bearing "start-run-brake" and bearing time-varying speed conditions.
- Without signal processing, the VSF-ST network could directly learn useful vibration shock fragments in the fault signal by a layer-by-layer feature extraction, so as to classify the fault bearings from the health bearings and the fault bearings from each other. The fusion of high-dimensional features of bearing speed signals while extracting features of bearing vibration signals could further enhance the adaptivity of the network under variable speed conditions.

In future research, the fusion method and fusion mechanism of bearing signal features under variable speed conditions will be further investigated to enhance the interpretability of the model. In addition, the generalizability of the model will be further improved by fusing the bearing signal feature extraction method with a deep learning algorithm in the case of a limited sample set of variable speed conditions.

Author Contributions: Conceptualization, F.C. and X.Y.; methodology, F.C.; software, F.C.; validation, S.S. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shaanxi Provincial Natural Science Foundation (No. 2022JQ-344) and the Fourteenth Five-Year Plan Advance Research Project (No. 50902060401).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

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

Abbreviations

The following abbreviations are used in this manuscript:

SAE	Sparse autoencoder
FFT	Fast Fourier transform
CNN	Convolutional neural network
VSF-ST	Vibration–speed data fusion network based on SAE and transformer
MSA	Multihead self-attention
GeLU	Gaussian error linear unit
KL	Scatter calculation
ReLU	Rectified linear unit
MLP	Multilayer perceptron
SQV	Spectra Quest varying speed dataset
A	Acceleration
D	Deceleration
A-D	Acceleration and then deceleration
D-A	Deceleration and then acceleration
DCNN	Deep convolutional neural network
LSTM	Long short-term memory
DNN	Deep neural network

References

1. Lei, Y.; Jia, F.; Lin, J.; Xing, S.; Ding, S.X. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. *IEEE Trans. Ind. Electron.* **2016**, *63*, 3137–3147. [CrossRef]
2. Lei, Y.G.; Lin, J.; He, Z.J.; Zuo, M.J. A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mech. Syst. Signal. Process.* **2013**, *35*, 108–126. doi: 10.1016/j.ymssp.2012.09.015. [CrossRef]
3. Randall, R.B.; Antoni, J. Rolling element bearing diagnostics-A tutorial. *Mech. Syst. Signal. Process.* **2011**, *25*, 485–520. doi: 10.1016/j.ymssp.2010.07.017. [CrossRef]

4. Lee, J.; Wu, F.; Zhao, W.; Ghaffari, M.; Liao, L.; Siegel, D. Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mech. Syst. Signal. Process.* **2014**, *42*, 314–334. [CrossRef]
5. Grieves, M.; Vickers, J., Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In *Transdisciplinary Perspectives on Complex Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 85–113.
6. Chen, J.; Pan, J.; Li, Z.; Zi, Y.; Chen, X. Generator bearing fault diagnosis for wind turbine via empirical wavelet transform using measured vibration signals. *Renew. Energy* **2016**, *89*, 80–92. [CrossRef]
7. Ade, P.A.R.; Aghanim, N.; Arnaud, M.; Ashdown, M.; Aumont, J.; Baccigalupi, C.; Banday, A.J.; Barreiro, R.B.; Bartlett, J.G.; Bartolo, N.; et al. Planck 2015 results XIII. Cosmological parameters. *Astron. Astrophys.* **2016**, *594*, A13. doi: 10.1051/0004-6361/201525830. [CrossRef]
8. Shao, H.D.; Xia, M.; Han, G.J.; Zhang, Y.; Wan, J.F. Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer convolutional neural network and thermal images. *IEEE Trans. Industr. Inform.* **2021**, *17*, 3488–3496. doi: 10.1109/tii.2020.3005965. [CrossRef]
9. Lu, S.L.; Yan, R.Q.; Liu, Y.B.; Wang, Q.J. Tacholess Speed Estimation in Order Tracking: A Review With Application to Rotating Machine Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 2315–2332. doi: 10.1109/tim.2019.2902806. [CrossRef]
10. Antoni, J. The infogram: Entropic evidence of the signature of repetitive transients. *Mech. Syst. Signal. Process.* **2016**, *74*, 73–94. doi: 10.1016/j.ymssp.2015.04.034. [CrossRef]
11. Cheng, Y.W.; Lin, M.X.; Wu, J.; Zhu, H.P.; Shao, X.Y. Intelligent fault diagnosis of rotating machinery based on continuous wavelet transform-local binary convolutional neural network. *Knowl. Based Syst.* **2021**, *216*, 106796. doi: 10.1016/j.knosys.2021.106796. [CrossRef]
12. Jin, Y.R.; Qin, C.J.; Huang, Y.X.; Liu, C.L. Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network. *Measurement* **2021**, *173*, 108500. doi: 10.1016/j.measurement.2020.108500. [CrossRef]
13. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal. Process.* **2021**, *151*, 107398. doi: 10.1016/j.ymssp.2020.107398. [CrossRef]
14. San Martin, G.; Droguett, E.L.; Meruane, V.; Moura, M.D. Deep variational auto-encoders: A promising tool for dimensionality reduction and ball bearing elements fault diagnosis. *Struct. Health. Monit.* **2019**, *18*, 1092–1128. doi: 10.1177/1475921718788299. [CrossRef]
15. Wang, X.; Mao, D.X.; Li, X.D. Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. *Measurement* **2021**, *173*, 108518. doi: 10.1016/j.measurement.2020.108518. [CrossRef]
16. Zhang, L.; Lv, Y.; Huang, W.; Yi, C. Bearing fault diagnosis under various operation conditions using synchrosqueezing transform and improved two-dimensional convolutional neural network. *Meas. Sci. Technol.* **2022**, *33*, 085002. [CrossRef]
17. Hasan, M.J.; Islam, M.M.; Kim, J.M. Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. *Measurement* **2019**, *138*, 620–631. [CrossRef]
18. Zhao, W.; Wang, Z.; Cai, W.; Zhang, Q.; Wang, J.; Du, W.; Yang, N.; He, X. Multiscale inverted residual convolutional neural network for intelligent diagnosis of bearings under variable load condition. *Measurement* **2022**, *188*, 110511. [CrossRef]
19. He, Z.; Shao, H.; Ding, Z.; Jiang, H.; Cheng, J. Modified deep autoencoder driven by multisource parameters for fault transfer prognosis of aeroengine. *IEEE Trans. Ind. Electron.* **2021**, *69*, 845–855.
20. Han, T.; Liu, C.; Yang, W.G.; Jiang, D.X. Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions. *ISA Trans.* **2019**, *93*, 341–353. doi: 10.1016/j.isatra.2019.03.017. [CrossRef]
21. An, Z.H.; Li, S.M.; Wang, J.R.; Xin, Y.; Xu, K. Generalization of deep neural network for bearing fault diagnosis under different working conditions using multiple kernel method. *Neurocomputing* **2019**, *352*, 42–53. doi: 10.1016/j.neucom.2019.04.010. [CrossRef]
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *NIPS* **2017**, *30* .
23. Jin, Y.; Hou, L.; Chen, Y. A Time Series Transformer based method for the rotating machinery fault diagnosis. *Neurocomputing* **2022**, *494*, 379–395. [CrossRef]
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
25. Cordonnier, J.B.; Loukas, A.; Jaggi, M. On the relationship between self-attention and convolutional layers. *arXiv* **2019**, arXiv:1911.03584.
26. Dubey, S.R.; Singh, S.K.; Chaudhuri, B.B. A Comprehensive Survey and Performance Analysis of Activation Functions in Deep Learning. *arXiv* **2021**, arXiv:2109.14545.
27. Liu, S.; Chen, J.; He, S.; Shi, Z.; Zhou, Z. Subspace Network with Shared Representation learning for intelligent fault diagnosis of machine under speed transient conditions with few samples. *ISA Trans.* **2021**, *128*, 531–544. [CrossRef]
28. Shi, Z.; Chen, J.; Zi, Y.; Zhou, Z. A novel multitask adversarial network via redundant lifting for multicomponent intelligent fault detection under sharp speed variation. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–10. [CrossRef]
29. Huang, H.; Baddour, N. Bearing vibration data collected under time-varying rotational speed conditions. *Data Brief.* **2018**, *21*, 1745–1749. [CrossRef]