Research Article

Early Fault Detection Method of Rolling Bearing Based on MCNN and GRU Network with an Attention Mechanism

Xiaochen Zhang, Yiwen Cong, Zhe Yuan ⓘ, Tian Zhang, and Xiaotian Bai

School of Mechanical Engineering, Shenyang Jianzhu University, Shenyang, Liaoning 110168, China

Correspondence should be addressed to Zhe Yuan; yuanzhe@sjzu.edu.cn

Aiming at the problem of early fault diagnosis of rolling bearing, an early fault detection method of rolling bearing based on a multiscale convolutional neural network and gated recurrent unit network with attention mechanism (MCNN-AGRU) is proposed. This method first inputs multiple time scales rolling bearing vibration signals into the convolutional neural network to train the model through multiscale data processing and then adds the gated recurrent unit network with an attention mechanism to make the model predictive. Finally, the reconstruction error between the actual value and the predicted value is used to detect the early fault. The training data of this method is only normal data. The early fault detection in the operating condition monitoring and performance degradation assessment of the rolling bearing is effectively solved. It uses a multiscale data processing method to make the features extracted by CNN more robust and uses a GRU network with an attention mechanism to make the predictive ability of this method not affected by the length of the data. Experimental results show that the MCNN-AGRU rolling bearing early fault diagnosis method proposed in this paper can effectively detect the early fault of the rolling bearing and can effectively identify the type of rolling bearing fault.

1. Introduction

As one of the key parts in rotating machinery, rolling bearing mainly plays a role in undertaking stress and transferring load in the system. Because of its long-term operation under high-speed, high-load working conditions, the rolling bearing has become the most easily damaged part of mechanical equipment [1]. Once the rolling bearing is damaged, it will have a very serious impact on the mechanical equipment, so it is of great significance to study rolling element bearings failure mechanisms.

The typical life curve of the rolling bearings is shown in Figure 1. There are four stages: (1) running-in stage, (2) normal operation stage, (3) early weak fault occurrence and healing stage, and (4) severe fault stage. The early faults are too weak to detect and once an early failure occurs, and the rolling bearings will deteriorate rapidly after a short period of the "healing stage"; it will lead to serious consequences. If the fault can be detected and remedied at an early stage, that

would avoid bigger safety problems and reduce losses. Therefore, the early fault detection of rolling bearings is very important [2–5]. And there are two problems to be faced in the detection of early faults: (1) The early faults are too weak to detect, and it is more difficult to extract the features. (2) There are less early fault data, which is not enough to train the model.

The current methods for diagnosing rolling bearing faults can be roughly divided into two categories [6]. The first category is model-based fault diagnosis methods, which mainly uses expert knowledge to analyse the fault frequency [7–9] or establish a degradation model [10, 11] to isolate early faults. However, this method relies on the subjective choice of people and the accuracy of the model; it requires high experience. The model is only designed for specific fields, which limits the scope of application to a certain extent. The second category is fault diagnosis methods based on data [12–14]. As the field of fault diagnosis enters the era of "big data," a series of data-based methods have emerged.
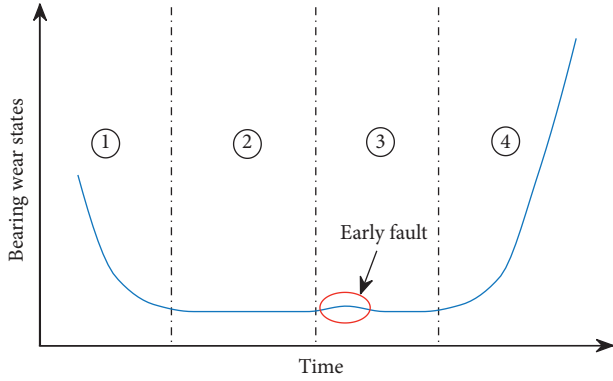
FIGURE 1: Typical life curve of rolling bearing.

The data-based method relies on the neural network to extract the features of the data on its own, eliminating the artificial subjectivity and dependence on the human experience, which is more in line with the monitoring of today's large-scale industry.

There are some statistical methods based to isolate early faults, mainly including principal component analysis (PCA) [6], autoregressive models [15], a neural network with multiple hidden layers [16, 17], support vector machine (SVM) [18, 19], K nearest neighbour (KNN) algorithm [20], and local outlier factor (LOF) algorithm [21, 22]. But these methods rely on well-selected features to classify faults.

In recent years, with the rapid development and wide application of deep learning; it has become the focus of fault diagnosis. There are some typical networks such as deep neural network (DNN) [23], deep belief network (DBN) [24], autoencoder (AE) [25], convolutional neural network (CNN) [26], and recurrent neural network (RNN) [27]. Although the accuracy of DBM and DNN is improved compared with shallow artificial neural networks, there is still the problem of artificial extraction of time series features, ignoring the characteristics of data timing. AE belongs to unsupervised learning, which is mainly used for data dimensionality reduction or feature extraction. It usually needs to be applied to the field of load forecasting after combining with other models. CNN is a neural network with convolution calculation and depth structure. Convolution and pooling are used to extract data features, which reduces the error caused by artificial feature extraction. It is widely used in image, voice, and other fields. However, it is difficult for a single CNN network to extract the weak features of early faults, so a multiscale convolution neural network is introduced to extract more comprehensive features. But it is difficult for an only MCNN model to learn the timing dynamics of retained data. RNN introduces the cyclic structure into the network so that it can model the dynamic time series data better than other neural networks [28]. Gated recurrent unit (GRU) is a special RNN. GRU and long short-term memory network (LSTM) [29–31] are solving the problem of gradient disappearance in RNN. They can consider the long-term and short-term dependence in time series more completely. Compared with LSTM, GRU has a faster convergence speed and no difference in accuracy. However,

when the input time series is long, RNN series networks such as LSTM and GRU are prone to lose sequence information and it is difficult to model the structure information between data, which affects the accuracy of the model [32]. The attention mechanism is a resource allocation mechanism, which can assign different weights to input features so that the features containing important information will not disappear with the increase of step size, highlight the influence of more important information, and make the model easier to learn the long-distance interdependence in the sequence [33].

However, although there are data-based early fault detection methods, early fault detection still faces the following challenges: (1) how to extract comprehensive and robust features from early fault signals; (2) consider the timing characteristics of bearing vibration signals to detect anomaly; (3) and when the data input length is too long, there is a problem of missing information.

Because of the above problems, this paper proposes an early fault diagnosis method of MCNN-AGRU. This method uses MCNN to extract the features of the rolling bearing at different time scales and filter out certain noise in the multiscale calculation to obtain more robust and comprehensive features of the bearing. The GRU network with an attention mechanism can learn the long-term dependence characteristics of the data, and the features containing important information will not disappear with the increase of the step size, thereby highlighting the influence of more important information, making the model easier to learn the long-term sequence. The interdependence of distance [34] solves the problem of information loss caused by too long data. Finally, a large amount of normal operating data of rolling bearings is used to construct a predictive model of the normal operating state of rolling bearings. The model can learn the distribution of normal data through training and use the learned prediction value and the reconstruction error of the true value to measure the operating state of the rolling bearing and perform early alarm.

The main contributions of this paper are as follows: (1) proposing an early fault diagnosis method that only needs to use the normal bearing data to train the model, which solves the problem of less early bearing fault data; (2) using multiscale data processing methods to make the features extracted by CNN more robust; (3) and using GRU network with an attention mechanism to make model predictive ability independent of the length of the data.

The main structure of the paper is as follows: the second part introduces the basic theoretical knowledge. The third part proposes the MCNN-AGRU method. The fourth part verifies the performance of the scheme through simulation. Finally, the conclusion is in the fifth section.

2. Fundamental Theory

2.1. Multiscale Data Processing. At present, most feature extraction methods directly use the raw vibration signals of rolling bearings to input into the neural network, but the features extracted from single time scale data are not comprehensive [35], so multiscale data processing methods

are used. As shown in Figure 2, this paper uses a multiscale data processing layer to process the original vibration signal and obtains vibration signals of multiple time scales by moving without overlapping windows and calculating the arithmetic average. The data preprocessed in this way can filter out high-frequency disturbance and random noise to a certain extent. The specific operation of this method is as follows: a segment of vibration signal $X = \{x_1, \ldots, x_n, \ldots, x_N\}$ is given, N is the length of the original input data, s is the number of multiscale processing scales, and x_n is the nth vibration value of the original signal. If the multiscale output signal is assumed to be $y_{s,j}$, the calculation process of the multiscale processed data is shown in

$$y_{s,j} = \frac{1}{s} \sum_{n=(j-1)s+1}^{js} x_n, \quad 1 \le j \le \frac{N}{s}. \tag{1}$$

The data length after multiscale data processing is N/s. The range of s selected in this paper is 1~4.

2.2. Convolutional Neural Network.
The convolutional neural network is a multilevel neural network, including filtering level and classification level. Among them, the filtering stage is used to extract the features of the input signal, the classification stage classifies the learned features, and the two-stage network parameters are obtained through joint training [36]. The filter stage includes a convolutional layer and a pooling layer, and uses an activation function to perform nonlinear operations on it. The convolution layer uses the convolution kernel to perform convolution operations on the local area of the input signal and generate corresponding features. The most important feature of the convolutional layer is weight sharing; that is, the same convolution kernel will traverse the input once with a fixed step. Weight sharing reduces the network parameters of the convolutional layer and avoids overfitting caused by too many parameters. The main purpose of the pooling layer is to reduce the parameters of the neural network and extract the features obtained by the convolutional layer twice. The one-dimensional convolution process is shown in Figure 3. The convolution kernel moves the input signal according to the step length to extract the features, and then the obtained features are pooled to obtain more advanced features.

2.3. Gated Recurrent Unit Network.
Gated recurrent unit network (GRU) is a variant of the recurrent neural network (RNN). RNN is a type of recurrent neural network that takes sequence data as input, recursively in the evolution direction of the sequence, and all recurrent units are connected in a chain [37]. As shown in Figure 4, the GRU network consists of an update gate and a reset gate. The main function of the update gate is to control the extent to which the state information from the previous moment is brought into the current state. The larger the value of the update gate, the more state information from the previous moment is brought in [38]. The main function of the reset gate is to determine the degree of discarding previous information.

The smaller the value, the more information is ignored. The GRU expression is as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]), \tag{2}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]), \tag{3}$$

$$\widetilde{h}_t = \tanh(W \cdot [r_t{}^* h_{t-1}, x_t]), \tag{4}$$

$$h_t = (1 - z_t)^* h_{t-1} + z_t{}^* \widetilde{h}_t. \tag{5}$$

In the above formula, σ represents the Sigmoid activation function. The parameters in the formula are W, W_z, and W_r.

2.4. Attention Mechanism.
The attention mechanism is a resource allocation mechanism that simulates the attention of the human brain. At a certain moment, the human brain will focus its attention on the areas that need to be focused, reducing or even ignoring the attention to other areas to get more attention. Needing to pay attention to the details of information and suppressing other useless information, its core idea is to change the attention to information ingeniously and reasonably, ignore irrelevant information and amplify the required information. The attention mechanism allocates sufficient attention to key information through probability allocation, highlights the impact of important information, and improves the accuracy of the model. The structure of the attention mechanism is shown in Figure 5. Among them, x_t(t∈[1, n]) represents the input of the GRU network, h_t(t∈[1, n]) corresponds to the hidden layer output of each input through GRU, α_t(t∈[1, n]) is the attention probability distribution value of the attention mechanism to the GRU hidden layer output, and y is the GRU output value of the attention mechanism introduced.

2.5. Support Vector Data Description.
Support vector data description (SVDD) is a single-valued classification algorithm, which can distinguish target samples from nontarget samples. At present, the SVDD algorithm is mainly used for abnormal state detection and fault identification that only define the normal working state space to judge whether the working state is normal or not. Given a training sample $\{x_i \in R^d, i = 1, 2, \ldots, N\}$, the goal of SVDD is to determine a hyperspherical body that can surround all training samples with a minimum volume. Assuming that a and R are the center and radius of the hypersphere, respectively, the SVDD optimization problem can be expressed as follows:

$$\min R^2 + C \sum \xi_i, \tag{6}$$

$$\|\varphi(x_i) - a\|^2 \le R^2 + \xi_i, \quad \xi_i \ge 0, \forall_i. \tag{7}$$

C is the constant used to control the degree of punishment for misdivided samples. ξ_i is the relaxation factor. $\varphi(x_i)$ is the mapping from sample space to feature space.
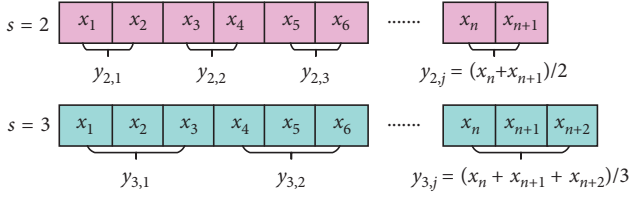
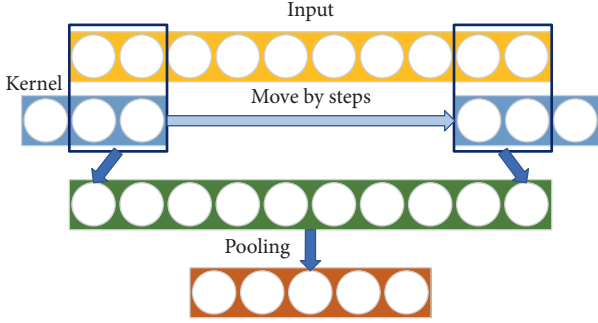FIGURE 2: Multiscale processing.

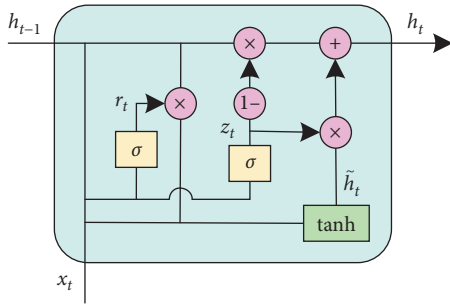

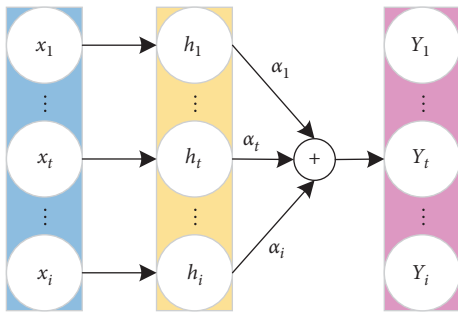FIGURE 3: One-dimensional convolution.

FIGURE 4: GRU structure.

FIGURE 5: Attention mechanism structure.

The Lagrange operator is used to solve the above optimization problem, and the following dual form can be obtained:

$$\max L = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \varphi(x_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j \varphi(x_i) \varphi(x_j), \quad (8)$$

where α_i is the Lagrange multiplier, $0 \leq \alpha_i \leq C(\sum_i \alpha_i = 1)$.

In order to improve the adaptability of the algorithm, the Gaussian kernel function $K(x_i, x_j)$ is introduced to replace the inner product operation on $\varphi(x_i)$ to improve the generalization ability of SVDD. The Gaussian kernel function is as follows:

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) = e^{-\|x_i - x_j\|^2/\sigma}, \quad (9)$$

where σ is the Gaussian kernel parameter, which has a great impact on the detection performance of SVDD.

To solve the above maximum optimization problem, the solution set $\{\alpha_i\}$ can be obtained; then, the center and minimum radius of the sphere can be obtained by the following formula:

$$a = \sum_i \alpha_i \varphi(x_i), \quad (10)$$

$$R^2 = K(x_k, x_k) - 2\sum_{i=1}^{n} \alpha_i K(x_i, x_k) + \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j), \quad (11)$$

where x_k is an arbitrary support vector.

For test sample Z, its thresholding algorithm is

$$f(z) = \|z - a\|^2 = 2\sum_{i=1}^{n} \alpha_i K(x_i, z) - \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j). \quad (12)$$

When $f(z) \geq R^2$, the sample is the normal sample; otherwise, it is the abnormal sample.

3. MCNN-AGRU Early Fault Detection Method

Most fault diagnosis methods based on deep learning are learning and classifying the serious faults, but there are a few methods for the early fault of bearings. The MCNN-AGRU method proposed in this paper solves this problem. MCNN can extract data features of different scales to increase the number of data sets and filter out part of the noise in the process to extract more robust features. The GRU network with the attention mechanism can solve the problem of information loss and the difficulty of taking into account the relationship between data and information when a single GRU network inputs data with too long sequence. Therefore, the MCNN-AGRU early fault detection method proposed in this paper is improved compared with the previous methods in feature extraction and timing processing. The experiment proves that the early fault of the bearing can be detected accurately and quickly.

3.1. MCNN-AGRU Fundamental.
The structure of the MCNN-AGRU model proposed in this paper is shown in Figure 6, which is mainly divided into three parts: the multiscale input layer, multiscale feature extraction layer, and prediction layer. First, the original vibration data is transformed into data of four time scales after multiscale preprocessing, as shown in Figure 7. Then, input the data of these four scales into the CNN network to extract the features, finally concatenate the features extracted from the data of the four scales to obtain the comprehensive feature, and
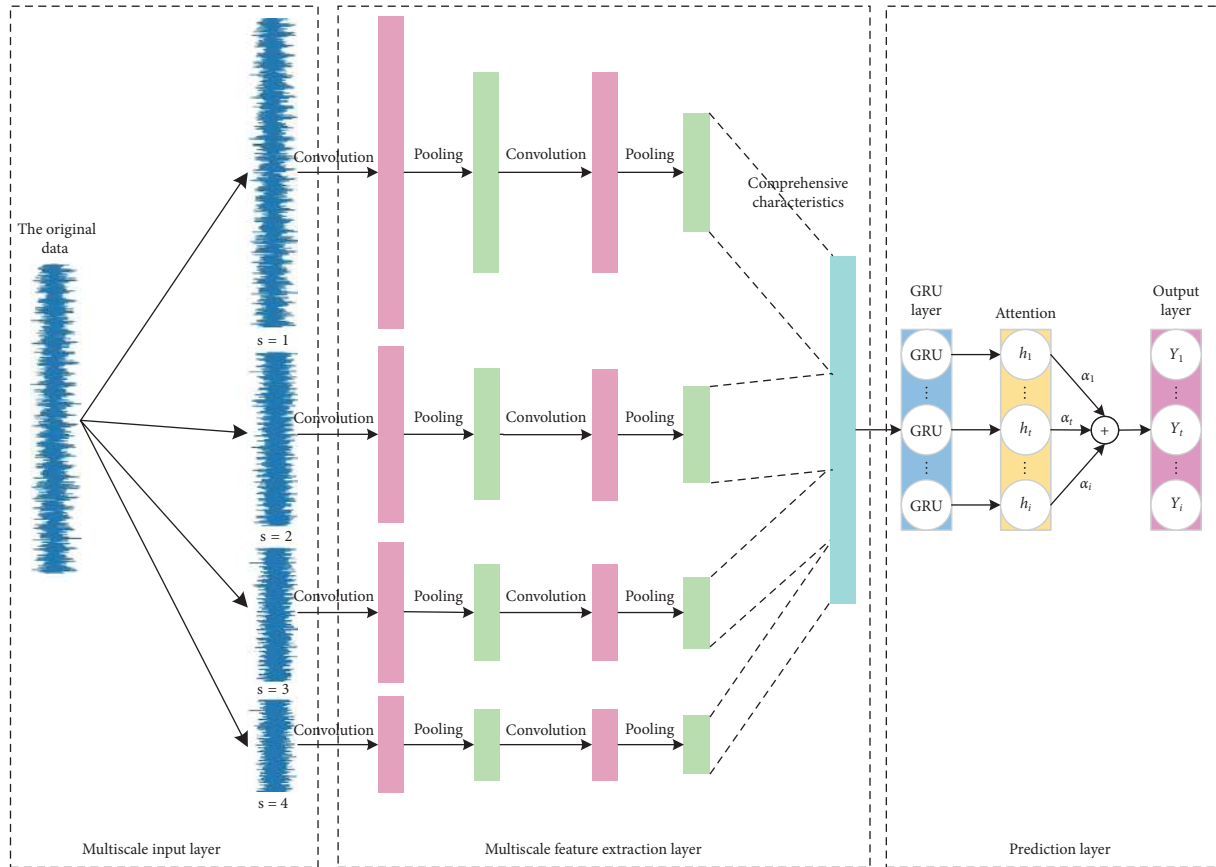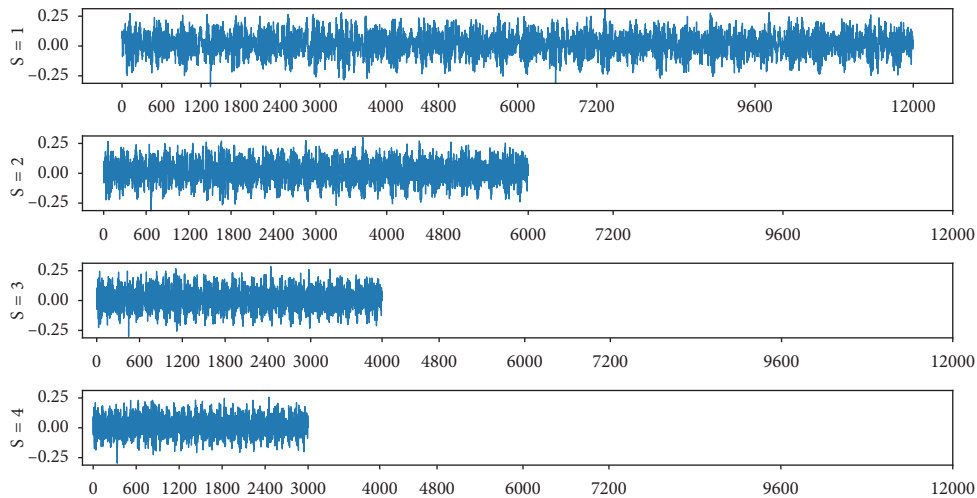
Figure 6: MCNN-AGRU model structure.

Figure 7: Multiscale processing of data.

then input it into the GRU network with the attention mechanism through the fully connected layer. The output is obtained after weighting.

The algorithm flow chart of the model is shown in Figure 8, including offline modelling and online monitoring. The offline modelling phase uses historical normal data to train MCNN and GRU network with an attention mechanism. When training the GRU network, GRU becomes a sequence generator, and the network outputs a prediction sequence with the same dimensions as the input data. During online monitoring, real-time data is input to the MCNN-AGRU model, and the forecast data at the next moment is output. The reconstruction error of the forecast data and the real data at the next moment is judged by the
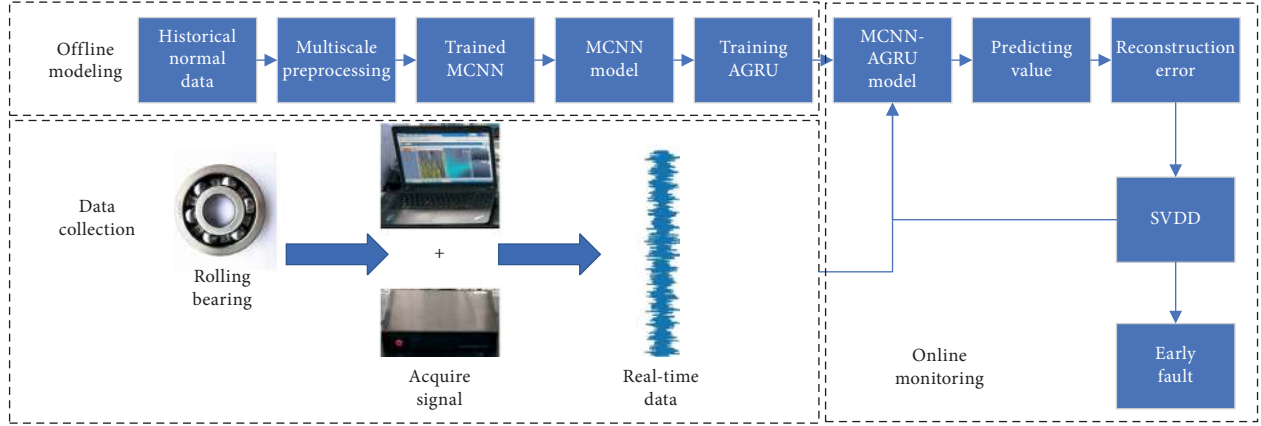
FIGURE 8: MCNN-AGRU early failure detection process.

support vector data description (SVDD) to determine the current operating status of the rolling bearing. If the result is normal, continue monitoring; otherwise, perform an alarm.

Each layer in the model is described as follows:

(1) Multiscale input layer: The input layer processes the original data through multiple scales to obtain four different scale inputs and inputs them into four different convolutional neural networks. The original data is $X = \{x_1, \ldots, x_n, \ldots, x_N\}$.

(2) Multiscale feature extraction layer: In this layer, two pairs of convolutional pooling layers are used for feature extraction for data of each scale, and the extracted features are connected in series to form a comprehensive feature. The input of the first convolutional layer is a signal of length $L = N/s$, and a convolution kernel of length m is selected to move on the data to extract features. Therefore, the output z_i of the i node in the feature graph is

$$z_i = \sigma\left(\mathbf{w}^{\mathrm{T}} y_{i:\, i+m-1} + b\right), \tag{13}$$

(i) where $\mathbf{w}^{\mathrm{T}}$ represents the weight matrix, b is the bias, $y_{i:\, i+m-1}$ represents the subsignal of length m starting from the i-th period in the original data y, and σ represents the activation function. ReLU activation is used here. The function can prevent the gradient from disappearing and speed up the function convergence. Sliding the convolution kernel from the beginning to the end, the j-th feature can be seen as

$$\mathbf{z}_j = [z_1, z_2, z_3, \ldots, z_{L-m+1}]. \tag{14}$$

(ii) After that, the pooling layer is used to further extract the features obtained by the convolutional layer, the max-pooling with a pooling length of p is adopted for calculating the local max value over the input feature map, and the k features are combined to obtain

$$\mathbf{h}_k = [h_1, h_2, h_3, \ldots, h_{L-m/p+1}], \tag{15}$$

$$h_j = \max_{(j-1)p+1 \leq i \leq jp}\{z_i\}. \tag{16}$$

(iii) The features after the pooling layer are expressed as $\mathbf{h}'_k$, and then the features are connected in series to get

$$\mathbf{q}^{(s)} = [\mathbf{h}'_1, \mathbf{h}'_2, \mathbf{h}'_3, \ldots, \mathbf{h}'_{K_2}]. \tag{17}$$

(iv) Finally, the features obtained from the four scales are connected in series to obtain comprehensive features:

$$\mathbf{q} = [\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \mathbf{q}^{(3)}, \mathbf{q}^{(4)}]. \tag{18}$$

(3) Prediction layer: The prediction layer is composed of the GRU layer, attention mechanism layer, and output layer. The GRU layer learns the feature vectors extracted by the multiscale feature extraction layer. By building a single-layer GRU structure, the proposed features are fully learned to capture its internal changing laws. The output of the GRU layer is denoted as H, and the output at step t is expressed as

$$h_t = \mathrm{GRU}\left(\mathbf{H}_{C,t-1}, \mathbf{H}_{C,t}\right), \quad t \in [1, i]. \tag{19}$$

The input of the attention mechanism layer is the output vector H that has been activated by the GRU network layer. The probability corresponding to different feature vectors is calculated according to the weight distribution principle, and the better weight parameter matrix is continuously updated and iterated. The calculation formula of the weight coefficient of the attention mechanism layer can be expressed as

$$e_t = u \tanh(wh_t + b), \tag{20}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{t} e_j}, \tag{21}$$

$$s_t = \sum_{t=1}^{i} \alpha_t h, \tag{22}$$

where e_t represents the attention probability distribution value determined by the output vector h_t of the GRU network layer at time t; u and w are weight coefficients; b is the bias coefficient; and the output of the attention layer at time t is represented by s_t. Finally, the input of the output layer is the output of the attention mechanism layer. The output layer calculates the output with a prediction step length of m through the fully connected layer. The prediction formula can be expressed as

$$y_t = \sigma(w_o s_t + b_o). \tag{23}$$

Among them, y_t represents the predicted output value at time t; w_o is the weight matrix; and b_o is the deviation vector. The activation function σ is Sigmoid.

The reconstruction error is calculated as follows:

$$E = \frac{y_t - x_t}{x_t}. \tag{24}$$

Only normal data is used to train MCNN-AGRU to make the model have the ability to predict the normal behaviour of the system along the time axis model in the process of detecting the early fault of the bearing. When the online data is input into the model, the model can predict the value of the next time of the data and calculate the reconstruction error with the actual value of the next time. The bearing has different reconstruction error when different faults occur. For example, when the system is normal, the reconstruction error is very small. When the system is abnormal, the reconstruction error will increase obviously. More importantly, the reconstruction errors of different types of early faults are also different. Therefore, we have reason to believe that the running state of the system can be judged by the reconstruction error. Various types of vibration signals and normal vibration signals are input into MCNN-AGRU to get the reconstruction error, and abnormal reconstruction error and normal reconstruction error are used to train SVDD to indicate the running state of the system.

4. Experimental Results and Analysis

This section verifies the accuracy and feasibility of the proposed MCNN-AGRU method through two sets of experiments on the self-built mechanical failure comprehensive simulation experiment platform and a full life cycle data set from the intelligent maintenance system (IMS) of the University of Cincinnati [39, 40].

4.1. MCNN-AGRU Fault Classification Experiment.
This part mainly uses experiments to verify the accuracy of the model's classification. The data set was acquired from the self-built mechanical failure comprehensive simulation experiment platform. This test stand consists of a motor, a rotor, a principle axis, a vibration sensor, and different kinds of rolling bearings (shown in Figure 9). The fault data set consists of four categories: normal state (N), the inner ring failure (IRF), the outer ring failure (ORF), and the rolling elements failure (REF). For the same fault, the degree is 0.2 mm, and the motor speeds is 1800 RPM. Digital data was collected at 12,000 samples per second.

This data set is used to evaluate the fault diagnosis performance of the algorithm. It contains four operating states common to the rolling bearing (N, IRF, ORF, REF). Each state has 120000 points, of which 80000 is selected for training data, 20000 for validation set, and 20000 for test set, and then test data of four operating states are entered into the trained model and judge the state of the system by the reconstructed error. The detailed information is listed in Table 1. For the proposed method, all structural hyperparameters are shown in Table 2.

Figure 10 shows the test results of the model. The black dots indicate the normal state, the green triangle indicates the rolling element failure, the blue square indicates the inner ring failure, and the pink cross indicates the outer ring failure. (1) The reconstruction error of the normal state fluctuates less than 2.5, and the reconstruction error of the abnormal state (rolling element failure, outer ring failure, and inner ring failure) is 2.5 to 20. It can be seen that there is a clear difference between the reconstruction error of the normal state and the abnormal state, which means that the model can distinguish the normal state from the abnormal state very well and has a good abnormality detection ability. (2) The reconstruction error range of rolling element failure is 2.5 to 5, the reconstruction error of inner ring failure fluctuates about 7.5, and the reconstruction error of outer ring failure ranges from 12 to 20. It can be seen that the model can distinguish three different types of faults well, indicating that the model has good fault classification capabilities.

4.2. MCNN-AGRU Fault Prediction Experiment.
This experiment is mainly used to verify the fault prediction ability of the model. To verify the performance of the model extended to the early state recognition of the rolling bearing, it is first necessary to analyse the operating characteristics of the rolling bearing throughout its life cycle. This article uses the full life cycle data of bearings from the Intelligent Maintenance Center of the University of Cincinnati for analysis. As shown in Figure 11, the bearing test bench carries four bearings on a shaft, which is driven by an AC motor. The speed is maintained at 2000 r/min. A radial load of 6000 lbs is applied to the shaft and bearing through a spring mechanism to accelerate bearing aging. The oil circulation system can measure the flow and temperature of lubricating oil. Besides, the electromagnet installed in the
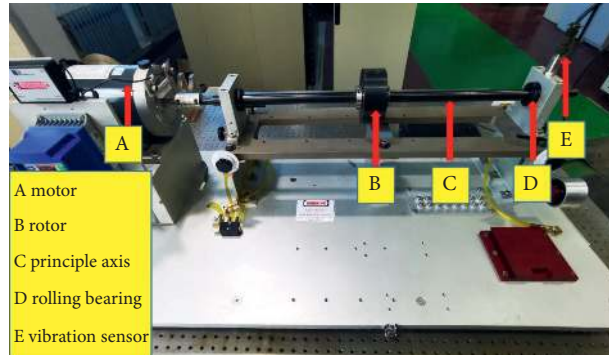
FIGURE 9: Comprehensive mechanical failure test bench.

TABLE 1: Rolling bearing failure test data.

Fault Location	Fault diameter (mm)	Number of points
Normal	0	120000
Outer ring failure (ORF)	0.2	120000
Rolling element failure (REF)	0.2	120000
Inner ring failure (IRF)	0.2	120000

oil return pipe will collect debris in the oil to prove the performance degradation of the bearing system. When the accumulated debris attached to the electromagnet exceeds a certain level, the system will stop running. A vibration acceleration sensor is installed on each bearing box. The data sampling rate is 20 kHz, sampling once every ten minutes. And there are 20480 points in each sample.

This paper chooses the data of experiment C in the IMS full life cycle experiment as the training set of this model to train the model. This experiment started on April 8th and ended on April 18th. After the accelerated aging test with applied load, the outer ring failure occurred on the 3# bearing. The data contains the vibration acceleration signals of the 3# bearing from normal operation to the occurrence of outer ring failure and contains 1399 samples in total. The data sampling rate is 20 kHz and each vibration signal snapshot length contains 20480 points. The first 800 samples are the healthy running data of 3# bearing. Select the first 500 samples of the sample file as the training data of the model, and the last 300 samples are validation data. The last 599 samples are used to test the performance of the model, and the last 599 samples contain the degradation process data of the 3# bearing. To ensure that the data input model has a certain physical meaning, 600 sampling points for roughly one revolution by calculating the sampling frequency and motor speed are obtained. Therefore, the data is rearranged and the data is input into the model for training and testing according to the cycle.

Figure 12 shows the performance of 3# bearing data in Experiment C. It uses normal data to train the model so that the model can learn the data changes of the rolling bearing in the health condition and use the reconstruction error between the actual value and the predicted value to measure the running state of the bearing. It can be seen from the partial enlargement of Figure 13 that the model first showed abnormal condition in the 8250th cycle. Then, the vibration signal returned to normal. This consists of the failure process of rolling bearings. When an early fault occurs to the outer ring and the rolling bearing is running, the weak defects in the outer ring will be smoothed by the continuous moving of the rolling elements. This abnormality will gradually diminish, so there will be short-term data similar to normal conditions. The rolling bearing with early fault will continue to run, and these two states will alternate. But the duration of the two states is getting shorter, and the amplitude of each abnormal signal will gradually increase.

To verify the stability and advancement of the MCNN-AGRU method proposed in this paper, this method is compared with several other fault detection methods.

As shown in Figures 14 and 15, it is clear that MCNN-AGRU can describe the development of the rolling bearing's damage. It is very sensitive to initial anomalies than other methods through Figures 14 and 15. For Kurtosis, it is not sensitive enough to abnormal changes in the signal and about 6650 revolutions slower than the method proposed in this paper. When Kurtosis adds the MCNN, its detection ability is enhanced, but its ability to predict the next running state of the bearing is reduced. For the RMS, it has a certain response to the early fault, but it is not obvious, and it cannot accurately predict the next running state of the bearing. When the RMS adds the MCNN, both detection ability and the ability which predicts the next running state are reduced. Compared with RMS, the MCNN-AGRU proposed in this paper is obviously larger than that in amplitude. It means that when both methods detect early faults, the MCNN-AGRU's response to early faults is more sensitive and obvious, while RMS is easily masked by noise. In conclusion, the MCNN-AGRU extracted data features are more stable and more sensitive to early faults.

TABLE 2: Model hyperparameters.

Hyperparameters	Value
Kernel number	8
Multiscale number	4
CNN layers	2
CNN optimizer	Adam
CNN activation function	ReLU
AGRU hidden layer	1
AGRU optimizer	Adam
AGRU activation function	Sigmoid
Gaussian kernel parameter σ	25

- Normal
- × Outer ring failure
- ■ Inner ring failure
- ▼ Rolling element failure
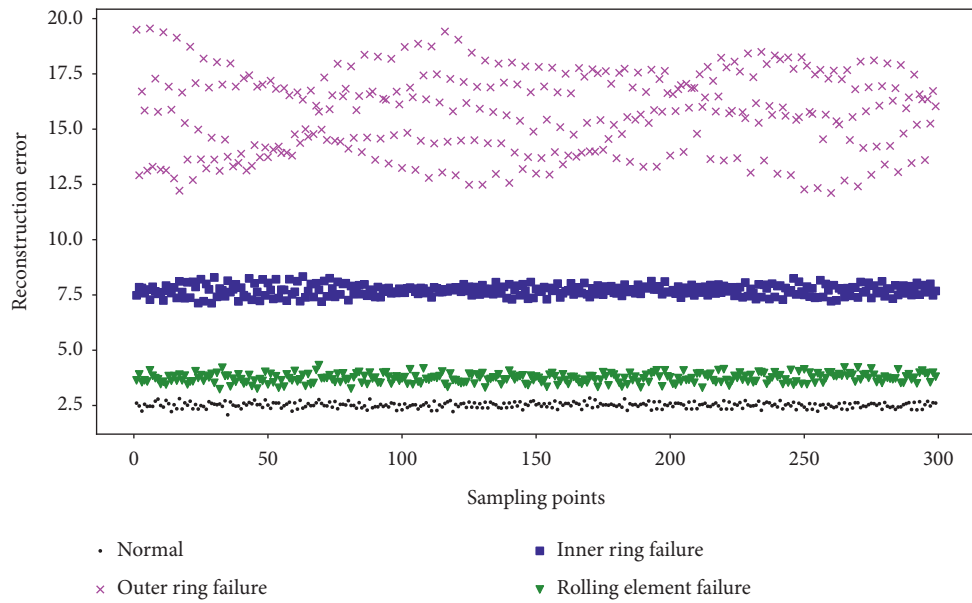
FIGURE 10: Diagram of model diagnosis results in four states.
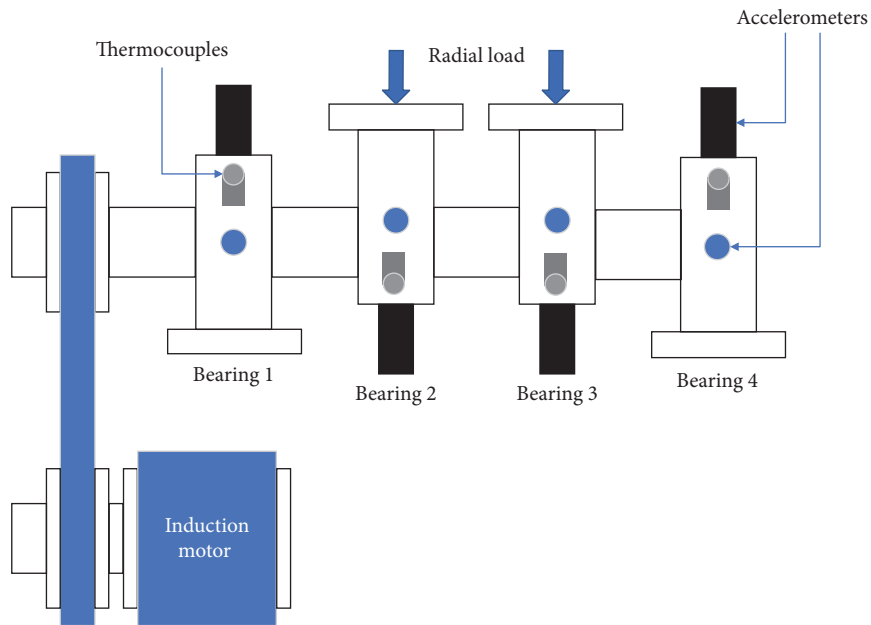
FIGURE 11: IMS full life cycle test bench.
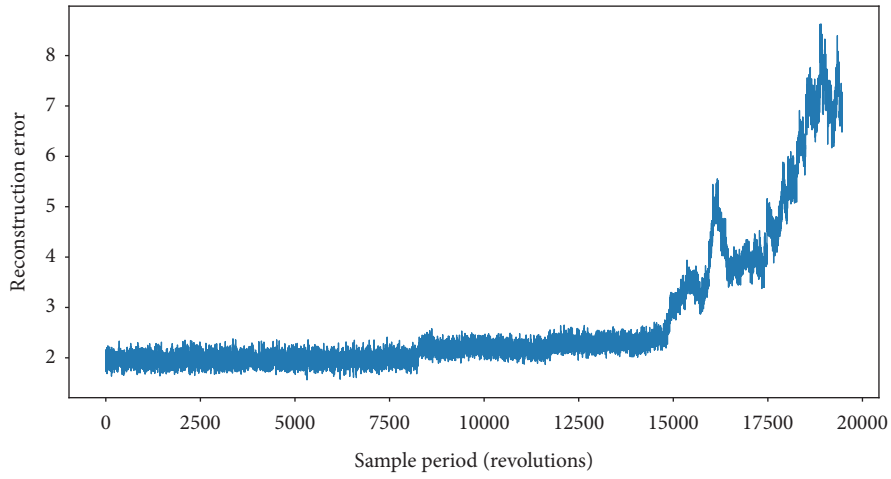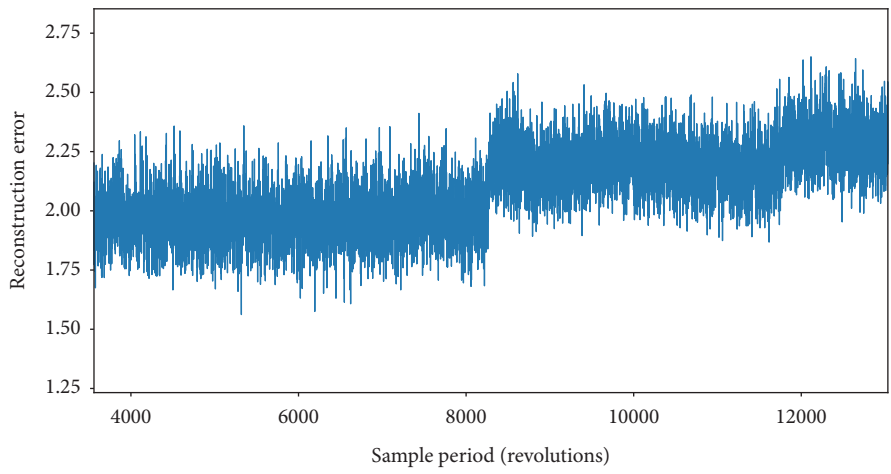
FIGURE 12: IMS full life cycle curve.

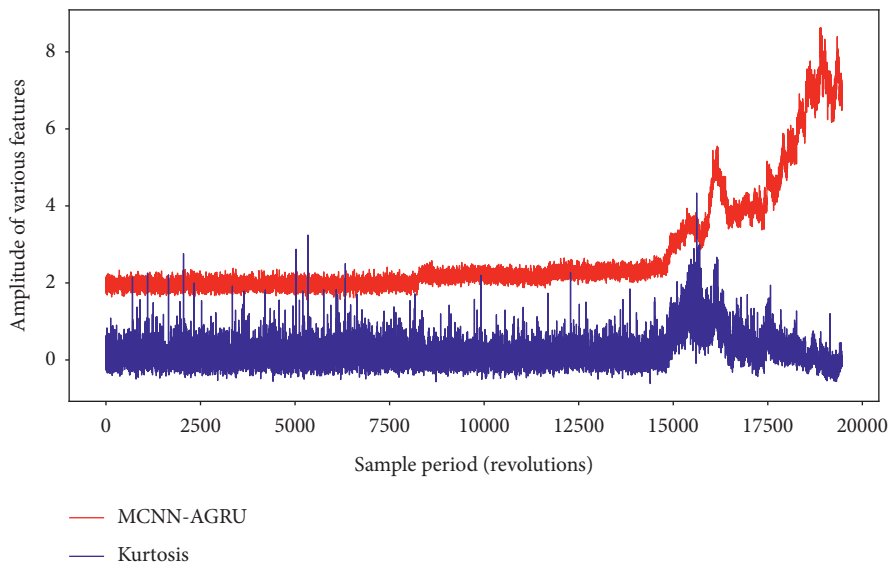FIGURE 13: Local enlargement between sample periods 4000 and 12000.

—— MCNN-AGRU

—— Kurtosis

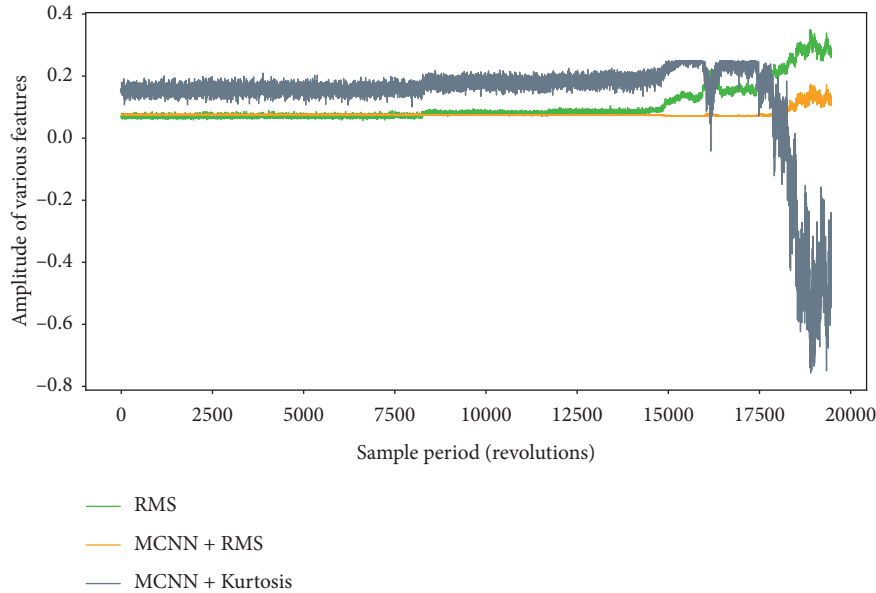FIGURE 14: Comparison of MCNN-AGRU and Kurtosis.

FIGURE 15: Comparison of RMS, MCNN + RMS, and MCNN + Kurtosis.

5. Conclusions

The early fault diagnosis method of MCNN-AGRU rolling bearing proposed in this paper integrates the multiscale feature extraction of the signal and the GRU network considering the timing characteristics of the data to achieve end-to-end bearing fault diagnosis.

(1) Compared with the traditional diagnosis method, the MCNN-AGRU method reduces the dependence on prior knowledge and experience, making the bearing fault diagnosis more intelligent.

(2) Extracting features of fault signals through MCNN can well extract the comprehensive features of the data, highlighting fault feature information, and then the GRU network with attention mechanism will process the sequence characteristics of the book sequence, which is similar to traditional shallow neural networks that can retain the timing correlation of the input features so that the diagnosis results are more accurate.

(3) Under different failure levels, through comparative analysis with Kurtosis, RMS, MCNN + Kurtosis, and MCNN + RMS methods, the MCNN-AGRU method is superior to other methods in the ability of early fault detection and the ability to predict the next running state running of the rolling bearing. It is proved that the method in this paper has high accuracy and good robustness.

(4) Since the experimental data in this article is collected in a laboratory environment, there is a certain difference between it and the actual production environment. At the same time, the determination of the current network model structure largely depends on experience, and different parameters have a greater impact on the recognition effect of the network. In future work, we will continue to study the model structure setting strategy for rolling bearing fault diagnosis.

Data Availability

This paper verifies the accuracy and feasibility of the proposed MCNN-AGRU method through two sets of experiments on the self-built mechanical failure comprehensive simulation experiment platform and a full life cycle data set from the intelligent maintenance system (IMS) of the University of Cincinnati [1-2]: [1] W. Gousseau, J. Antoni, F. Girardin, "Analysis of the Rolling Element Bearing data set of the Center for Intelligent Maintenance Systems of the University of Cincinnati." CM2016 2016 and [2] H. Qiu, J. Lee, J. Lin, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," Journal of Sound and Vibration, vol. 289(4), pp. 1066-1090, 2006.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] L. S. Qu, X. N. Zhang, and Y. D. Shen, *Theory and Methodology for Mechanical Condition Monitoring and Fault Diagnosis*, Xi'an Jiao Tong University Press, Xi'an, China, 2009.

[2] L. Cui, J. Huang, F. Zhang, and F. Chu, "HVSRMS localization formula and localization law: localization diagnosis of a ball bearing outer ring fault," *Mechanical Systems and Signal Processing*, vol. 120, pp. 608–629, 2019.

[3] I. El-Thalji and E. Jantunen, "A summary of fault modelling and predictive health monitoring of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 60-61, pp. 252–272, 2015.

[4] G. Prakash, S. Narasimhan, and M. D. Pandey, "A probabilistic approach to remaining useful life prediction of rolling element bearings," *Structural Health Monitoring*, vol. 18, no. 2, pp. 466–485, 2019.

[5] W. Guangbin, D. Moujun, H. Liangpei, and L. Long, "Prediction of bearing damage in wind turbines based on the quadratic root mean square of sub-band manifold," in *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 232, no. 18, pp. 3213–3223, 2018.

[6] S. Yin, S. X. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 11, pp. 6418–6428, 2014.

[7] H. Li, X. Lian, C. Guo, and P. Zhao, "Investigation on early fault classification for rolling element bearing based on the optimal frequency band determination," *Journal of Intelligent Manufacturing*, vol. 26, no. 1, pp. 189–198, 2015.

[8] A. Brkovic, D. Gajic, J. Gligorijevic, I. Savic-Gajic, O. Georgieva, and S. Di Gennaro, "Early fault detection and diagnosis in bearings for more efficient operation of rotating machinery," *Energy*, vol. 136, pp. 63–71, 2017.

[9] T. W. Rauber, F. de Assis Boldt, and F. M. Varejao, "Heterogeneous feature models and feature selection applied to bearing fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 1, pp. 637–646, 2015.

[10] D. U. Campos-Delgado and D. R. Espinoza-Trejo, "An observer-based diagnosis scheme for single and simultaneous open-switch faults in induction motor drives," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 2, pp. 671–679, 2011.

[11] I. Hwang, S. Kim, Y. Kim, and C. E. Seah, "A survey of fault detection, isolation, and reconfiguration methods," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 3, pp. 636–653, 2010.

[12] M. Gan and C. Wang, "Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 72-73, pp. 92–104, 2016.

[13] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mechanical Systems and Signal Processing*, vol. 72-73, pp. 303–315, 2016.

[14] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171–178, 2016.

[15] X. Jin, Y. Sun, Z. Que, Y. Wang, and T. W. S. Chow, "Anomaly detection and fault prognosis for bearings," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 9, pp. 2046–2054, 2016.

[16] J. Zarei, M. A. Tajeddini, and H. R. Karimi, "Vibration analysis for bearing fault detection and classification using an intelligent filter," *Mechatronics*, vol. 24, no. 2, pp. 151–157, 2014.

[17] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-d convolutional neural networks," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 11, pp. 7067–7075, 2016.

[18] P. Bangalore and L. B. Tjernberg, "An artificial neural network approach for early fault detection of gearbox bearings," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 980–987, 2017.

[19] J. Huang and X. Yan, "Related and independent variable fault detection based on KPCA and SVDD," *Journal of Process Control*, vol. 39, pp. 88–99, 2016.

[20] Q. P. He and J. Wang, "Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 20, no. 4, pp. 345–354, 2007.

[21] H. Ma, Y. Hu, and H. Shi, "Fault detection and identification based on the neighborhood standardized local outlier factor method," *Industrial & Engineering Chemistry Research*, vol. 52, no. 6, pp. 2389–2402, 2013.

[22] Z. Zhao, J. Yang, W. Lu, and X. Wang, "Application of local outlier factor method and back-propagation neural network for steel plates fault diagnosis," in *Proceedings of the 27th Chinese Control and Decision Conference (2015 CCDC)*, pp. 2416–2421, Qingdao, China, May 2015.

[23] F. Mohammad, K. B. Lee, and Y. C. Kim, *Short Term Load Forecasting Using Deep Neural Networks*, https://arxiv.org/pdf/1811.03242.pdf, 2018.

[24] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Y. ang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 3, pp. 2296–2305, 2017.

[25] C. Tong, J. Li, and C. Lang, "An efficient deep model for day-ahead electricity load forecasting with stacked denoising autoencoders," *Journal of Parallel and Distributed Computing*, vol. 117, pp. 267–273, 2018.

[26] Z. Hao, G. Liu, and H. Zhang, "Correlation filter-based visual tracking via adaptive weighted CNN features fusion," *IET Image Processing*, vol. 12, no. 8, pp. 1423–1431, 2018.

[27] Z. C. Lipton, "A critical review of recurrent neural networks for sequence learning," 2015, http://arxiv.org/abs/1506.00019.

[28] A. Graves, "Generating sequences with recurrent neural networks," 2014, https://arxiv.org/pdf/1308.0850.pdf.

[29] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks," *Applied Energy*, vol. 212, pp. 372–385, 2018.

[30] Y. F. Zhang, Q. Ai, and L. Lin, "A very short-term load forecasting method based on deep LSTM RNN at zone level," *Power System Technology*, vol. 43, no. 6, pp. 1884–1892, 2019.

[31] P. Li, S. He, and P. F. Han, "Short-term load forecasting of smart grid based on long-short-term memory recurrent neural networks in condition of real-time electricity price," *Power System Technology*, vol. 12, no. 12, pp. 4045–4052, 2018.

[32] Y. G. Cinar, H. Mirisaee, and P. Goswami, "Period-aware content attention rnns for time series forecasting with missing values," *Neurocomputing*, vol. 312, pp. 177–186, 2018.

[33] K. Ke, S. Hongbin, Z. Chengkang, and C. Brown, "Short-term electrical load forecasting method based on stacked auto-encoding and GRU neural network," *Evolutionary Intelligence*, vol. 12, no. 3, pp. 385–394, 2019.

[34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, https://arxiv.org/abs/1409.0473.

[35] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196–3207, 2019.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, pp. 367–415, MIT Press, Cambridge, UK, 2016.

[38] K. Cho, B. V.. Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: encoder-decoder approaches," 2014, https://arxiv.org/abs/1409.1259.

[39] W. Gousseau, J. Antoni, and F. Girardin, *Analysis of the Rolling Element Bearing Data Set of the Center for Intelligent Maintenance Systems of the*, University of Cincinnati, Cincinnati, OH, USA, 2016.

[40] H. Qiu, J. Lee, J. Lin, and G. Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," *Journal of Sound and Vibration*, vol. 289, no. 4-5, pp. 1066–1090, 2006.