Measurement 175 (2021) 109166

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Remaining useful life prediction of roller bearings based on improved 1D-CNN and simple recurrent unit

Dechen Yao [a,b,*], Boyang Li [a,b], Hengchang Liu [a,b], Jianwei Yang [a,b], Limin Jia [c]

[a] School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[b] Beijing Key Laboratory of Performance Guarantee on Urban Rail Transit Vehicles, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[c] State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Keywords:
Roller bearing
One dimensional convolution
Simple recurrent unit
Remaining useful life

ABSTRACT

To overcome the shortcomings of traditional roller bearing remaining useful life prediction methods, which mainly focus on prediction accuracy improvement and ignore labor cost and time, the present work proposed a novel prediction method by combining an improved one-dimensional convolution neural network (1D-CNN) and a simple recurrent unit (SRU) network. For feature extraction, the proposed method uses the ability of the 1D-CNN to extract signal features. Moreover, use the global maximum pooling layer to replace the fully connected layer. In the prediction part, a parallel-input SRU network was established by reconstructing the serial operation mode of a traditional recurring neural network (RNN). Finally, experiments were carried out using the XJTU-SY dataset to verify. Results revealed that on the premise of ensuring prediction accuracy, the 1D-CNN-SRU method could reduce manual intervention and time cost to a certain extent and provide an intelligent method for roller bearing remaining useful life prediction.

1. Introduction

Roller bearings are one of the most important parts of rotating machinery and directly affect production efficiency. It is reported that faults caused by roller bearings account for 45–55% of the most common rotating machinery problems [1]. Therefore, it is necessary to evaluate the reliability of roller bearings and predict their remaining useful life to maintain the safe operation of rotating machinery and bring certain economic benefits [2,3].

Therefore, a significant number of investigations have been carried out on roller bearing remaining useful life prediction based on feature extraction and prediction model construction. For feature extraction, traditional methods rely on manual experience and a large number of signal processing methods to extract feature sensitive to bearing degradation from the time domain, the frequency domain, and the time–frequency domain [4] or construct new statistical features by theoretical derivation and establish feature datasets after screening. The spectral kurtosis data driven, the construction of the health index screening algorithm are the methods for feature extraction and selection [5,6]. The isometric mapping algorithm (ISOMAP) and the principal component analysis (PCA) are mainly used to reduce feature dimensions [7,8]. Some researchers [9–11] have also constructed new statistical features, such as construction of feature index based on Gaussian mixture model (GMM) and Jensen-Rényi divergence (JRD), the feature index is constructed by the permutation entropy (PE) calculation method of tunable Q factor wavelet transform (TQWT), because the proposed PE calculation method includes an oscillation based procedure, the feature index is termed as oscillation based PE (OBPE), and the construction of a new index of MAC2-PSD based on moving average cross-correlation of the power spectral density of a vibration signal. For MAC2-PSD, PSD represents power spectral density, and MAC2 represents the moving average cross correlation. In addition, some methods divide the whole component to find out the key monitoring position [12], or introduce energy operator into fault monitoring to achieve accurate diagnostic [13]. These new diagnostic strategies also provide great help for the development of prognostics and health management (PHM).

In recent years, the application of deep learning technology in system state prognostics and health management also has been widely concerned [14]. Therefore, to improve prediction accuracy, some deep neural networks are proposed to replace traditional machine learning algorithms. Some researchers [15–17] have applied long short-term

* Corresponding author at: Beijing University of Civil Engineering and Architecture, School of Mechanical-Electronic and Vehicle Engineering, Beijing, 100044, China.
E-mail address: yaodechen@bucea.edu.cn (D. Yao).

https://doi.org/10.1016/j.measurement.2021.109166
Received 25 December 2020; Received in revised form 24 January 2021; Accepted 5 February 2021
Available online 12 February 2021

memory (LSTM) for equipment and parts remaining useful life prediction based on the powerful processing ability of a recurring neural network (RNN) for time series data. As a variant network of LSTM, the gated recurrent unit (GRU) network is also widely used in roller bearing life prediction [18,19]. Moreover, to further improve prediction accuracy, some mechanisms are often embedded in RNNs. For example, the construction of a bidirectional network to capture future moment information [20–22], the introduction of an attention mechanism to enhance the performance of important moments [23–25], ant colony algorithm and Bayesian algorithm are introduced to optimize the cell structure and network parameters [26,27].

However, the above methods only improve the back-end prediction model. For feature extraction, these methods still rely on traditional manual extraction processes to select sensitive features and construct datasets. For the degradation process of a roller bearing, sensitive features and feature changing rules under different working conditions are distinct. The performance abilities of a single feature and multi-features are also different. Therefore, the selection of different feature indexes and different numbers of features greatly affects prediction results. In fact, affected by actual engineering needs, operating conditions of rotating machinery become worse, making vibration signals nonlinear and non-stationary. Hence, the complexity of feature extraction from the time domain, the frequency domain, and the time–frequency domain has increased and the degree of intelligence has become low, feature extraction needs more manual experience and more complex signal processing methods, the labor cost and time consumption will be further increased. Therefore, some researchers have tried to use the automatic feature extraction ability of a convolution neural network (CNN) instead of manual extraction to reduce signal processing and realize system or parts remaining useful life prediction [28–30]. Although CNNs can replace manual feature extraction processes, they cannot process sequence data and cannot capture the dependence between sequence information; thus, they are not suitable for time series prediction problem.

The use of RNNs instead of traditional machine learning models as prediction algorithms can improve prediction accuracy to a certain extent; however, LSTM and GRU still have great limitations in calculation form.

Therefore, the present work proposed an improved one-dimensional (1D) CNN-SRU method for roller bearing remaining useful life prediction. For feature extraction, the 1D-CNN model was used instead of manual extraction and the global maximum pooling layer was introduced into CNN to simplify the model structure. For prediction, a parallel SRU structure was constructed to replace traditional RNNs. After experimental verification, it was found that on the premise of ensure the accuracy of the prediction, the method proposed could reduce the manual intervention, reduce the labor cost and time consumption to a certain extent.

The remaining paper is organized as follows: the second section introduces the theoretical background. The technical process of the proposed method is depicted in detail in the third section. The fourth section presents related experiments to verify the effectiveness of this method. Finally, conclusions are given in the fifth section.

2. Theoretical background

2.1. Feature extraction by 1D-CNN

Convolutional neural network (CNN) can deeply mine data through its powerful feature extraction ability. A CNN is generally two-dimensional and used to process image data [31]. A convolution kernel moves smoothly on an image to extract pixel features and realize image recognition and classification. In the field of roller bearing remaining useful life prediction, input data are mainly one-dimensional; thus, these features cannot be extracted by a two-dimensional convolution kernel. Therefore, in the feature extraction stage, according to the characteristics of sequence data, a 1D-CNN model is used to capture sensitive information and realize the feature extraction of vibration signals. Compared with traditional feature extraction methods, CNN can automatically extract data features without too much manual intervention and professional knowledge requirements [32]. The sequence data extraction process of the 1D-CNN model is presented in Fig. 1.

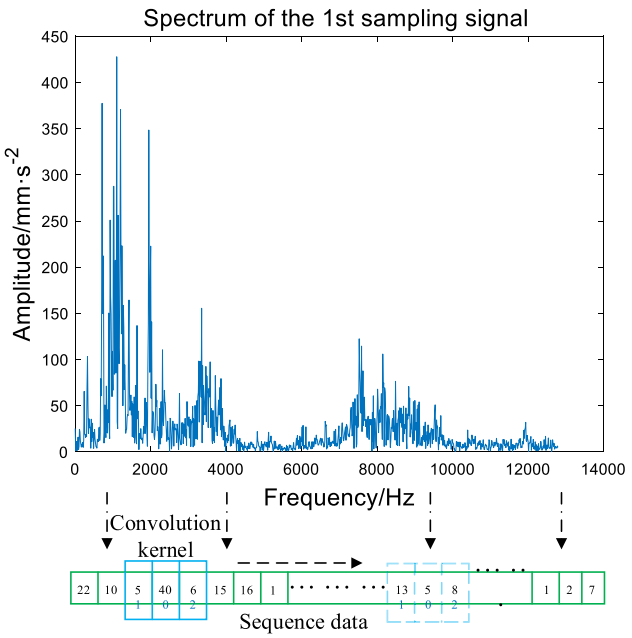

Fig. 1. Feature extraction sequence of 1D-CNN.

The first sampling condition of a data group in the XJTU-SY dataset was taken as an example (Sequence values in Fig. 1 do not represent actual spectrum values. Section 4.1 of this paper will introduce the XJTU-SY data set and related bearing information in detail. At the same time, except for Section 4.3.2, the bearing data used in the demonstration, data collection and test verification of other contents in this paper are provided by XJTU-SY data set). Affected by the radial load, the information reflected by the horizontal vibration signal is more obvious [33]. In addition, for feature extraction, spectrum information can effectively reduce the sequence length and enhance the information representation [34]. Therefore, fast Fourier transformation (FFT) was applied to the horizontal vibration signal. The sampling point of $N = 2048$ reduced the original signal from 32,768 data points to 2048 points in the spectrum. Moreover, as the spectrum of the real signal was symmetrical, only the positive part of the spectrum, the points at 0 Hz, and the central symmetry point (the first 1025 points) were taken to effectively reduce the length of the sequence, avoid the stacking of the convolution layer, and reduce the amount of calculation. The sequence data were then inputted into the 1D-CNN model, and the feature data were extracted by the smooth movement of the one-dimensional convolution kernel from left to right.

Further, by taking the current moment in Fig. 1 as an example, the convolution operation process could be expressed as "feature sequence $\otimes$ convolution kernel," where "$\otimes$" represents convolution operation symbol. Therefore, after the convolution operation, the number of extracted features was 17.

The spectrum information of bearings in different stages during their life cycles are also different, especially for the bearings in their middle and later stages of degradation, this difference becomes most obvious. Taking the same dataset as an example, the spectra of two groups of sampling data in the middle and later stages of degradation were randomly selected (Fig. 2) and compared with the spectrum information at the initial sampling stage in Fig. 1, and an obvious difference in the detected amplitudes was noticed. Based on the difference of data at

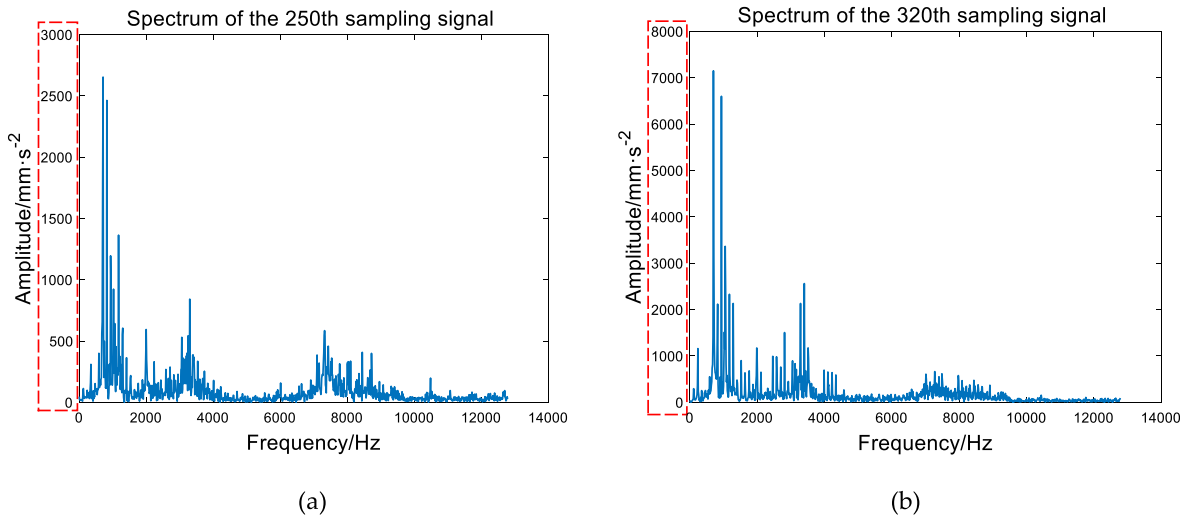

Fig. 2. Sampling spectra at different degradation stages: (a) 250th sampling and (b) 320th sampling.

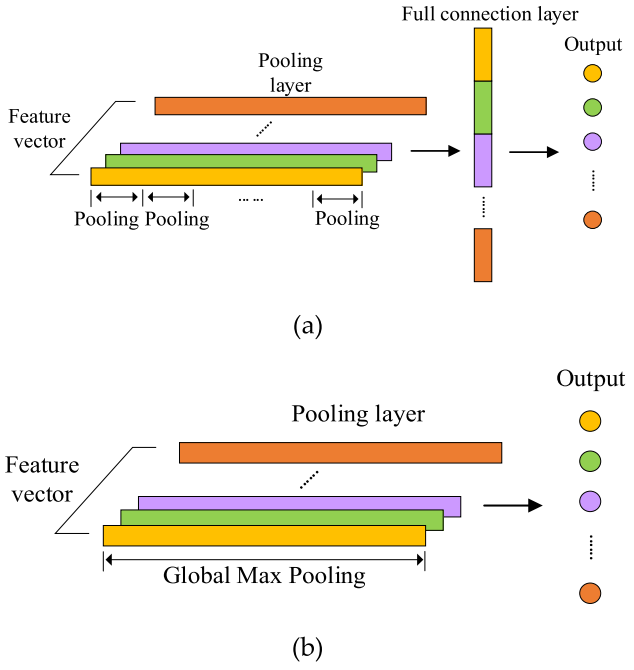

Fig. 3. Feature extraction: (a) Fully connected layer extraction and (b) GMP extraction.

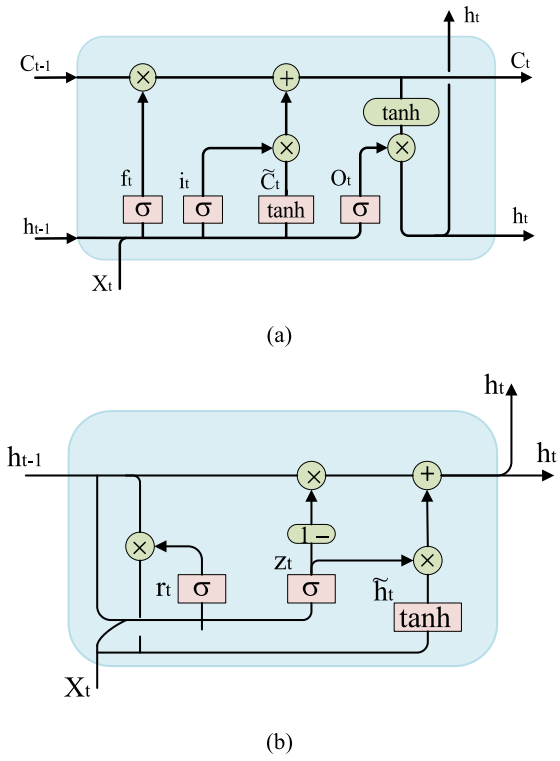

Fig. 4. Internal structures of (a) LSTM and (b) GRU models.

different moments, some features can be extracted manually to realize the prediction. Therefore, the 1D-CNN model could also extract different feature information from the sequence data at different times by the smooth movement of the convolution kernel and then realize the prediction based on the extracted features.

2.2. Improved 1D-CNN structure

A traditional CNN network relies on the output feature of the fully connected layer; therefore, the construction of the fully connected layer leads to parameter redundancy, overfitting, and calculation amount increment. To avoid a large number of parameters, simplify the model structure and improve operation efficiency, the global maximum pooling (GMP) is constructed by combining the pooling layer function and fully connected layer function of a CNN. Global aims at the entire vector rather than local pooling. As the difference of features reflects the degradation state of a bearing, the most representative feature of the max pooling output vector is generally selected (Fig. 3).

2.3. SRU model

Similar to a CNN, a recurrent neural network (RNN) is an important algorithm in the field of deep learning and has excellent performances in natural language processing [35], text translation, and sequence prediction. However, as the input of the current time is the output of the previous time, this fixed serial input–output structure limits the training speed of a traditional RNN, such as LSTM, and increases the amount of calculation [36]; thereby, it is difficult to train a RNN as fast and efficient as a CNN. As a variant structure of LSTM [37], a GRU network optimizes the “three gates” structure of LSTM to a “two gates” structure,

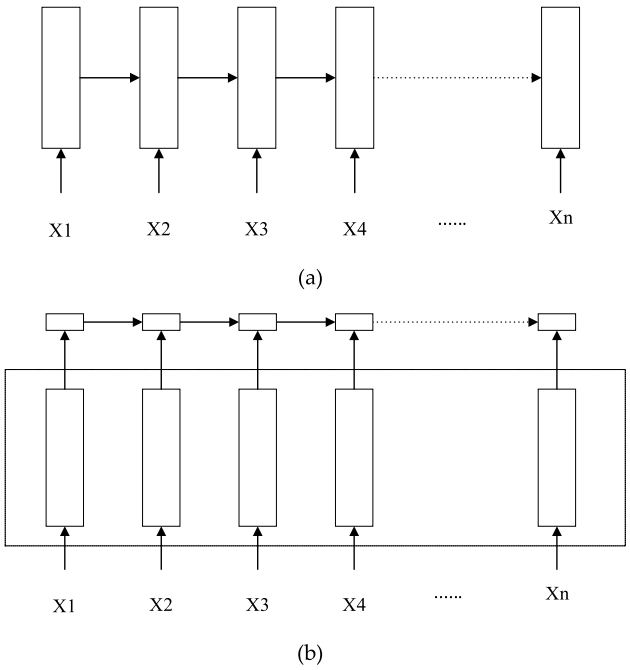

Fig. 5. Information transmission processes of (a) traditional RNN and (b) SRU models.

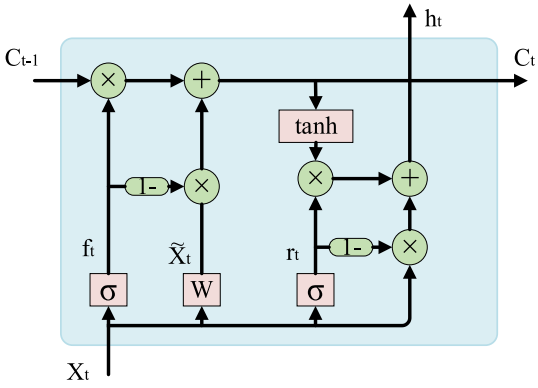

Fig. 6. Internal structure of SRU.

reduces the number of parameters by one-third, lessens the training time and calculation amount to a certain extent, and improve operational efficiency. However, this structure still does not fundamentally solve the problem of the serial operation mode of a RNN. Fig. 4 presents a comparison of LSTM and GRU models.

It is clear from Fig. 4 that both LSTM and GRU rely on the output h_{t-1} of the previous time. To solve this problem, Lei et al. [38] proposed the SRU model, which is a special variant network of an RNN. The unique internal structure of SRU can process each input step independently from other inputs, eliminate time dependency in operation, and turn a RNN into a network that can be parallelized to a certain extent without computing h_{t-1}. In the present work, a SRU network was applied to predict the remaining useful life of roller bearings. The information transmission processes of traditional RNN and SRU models are displayed in Fig. 5, and the internal structure of SRU is presented in Fig. 6.

In Fig. 6, the "+" symbol indicates the addition of two terms, the "×" symbol implies the multiplication of two terms, *W*is the weight matrix, and σ and *tanh* represent the sigmoid activation function and the hyperbolic tangent activation function, respectively, and are used to realize nonlinear transformation of input data.

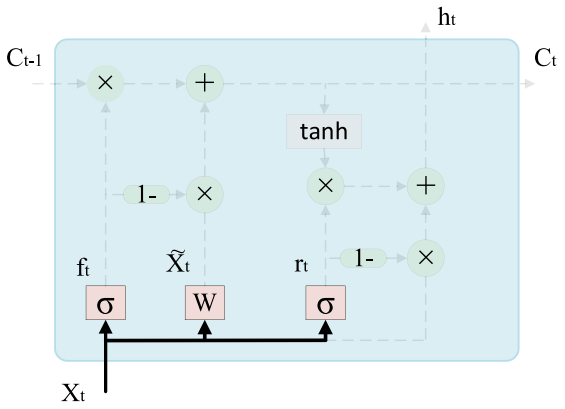

Fig. 7. Information flow direction (1) of SRU.

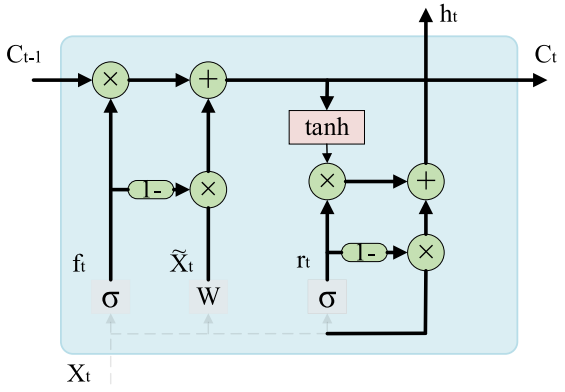

Fig. 8. Information flow direction (2) of SRU.

$$S(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2}$$

During calculations, SRU accepts the input X_t at the current time. In this step, SRU only calculates $X_t$without relying on the output at the previous time. The information flow direction of SRU is displayed in Fig. 7.

$$f_t = \sigma\left(\mathrm{W_f X_t} + \mathrm{b_f}\right) \tag{3}$$

where the forgetting gate f_t indicates the degree of forgetting historical information. After the linear transformation of X_t, f_t outputs a value between 0 and 1 based on the sigmoid activation function. This step determines the degree of forgetting old information.

$$\widetilde{X}_t = W \cdot X_t \tag{4}$$

In Eq. (4), X_t is multiplied by the weight matrix W to obtain the information $\widetilde{X}_t$ ready to be supplemented to a new cell state C_t.

$$r_t = \sigma(\mathrm{W_r X_t} + \mathrm{b_r}) \tag{5}$$

In Eq. (5), similar to the forgetting gate f_t, the reset gate r_t is also a sigmoid gate and used to control the degree of information update.

After the previous operation, the information continues to be transmitted in SRU, and the new cell state C_t and the output h_t at the current time are finally determined. The information flow direction of SRU is displayed in Fig. 8.

$$C_t = f_t \times c_{t-1} + (1-f_t) \times \widetilde{X}_t \tag{6}$$

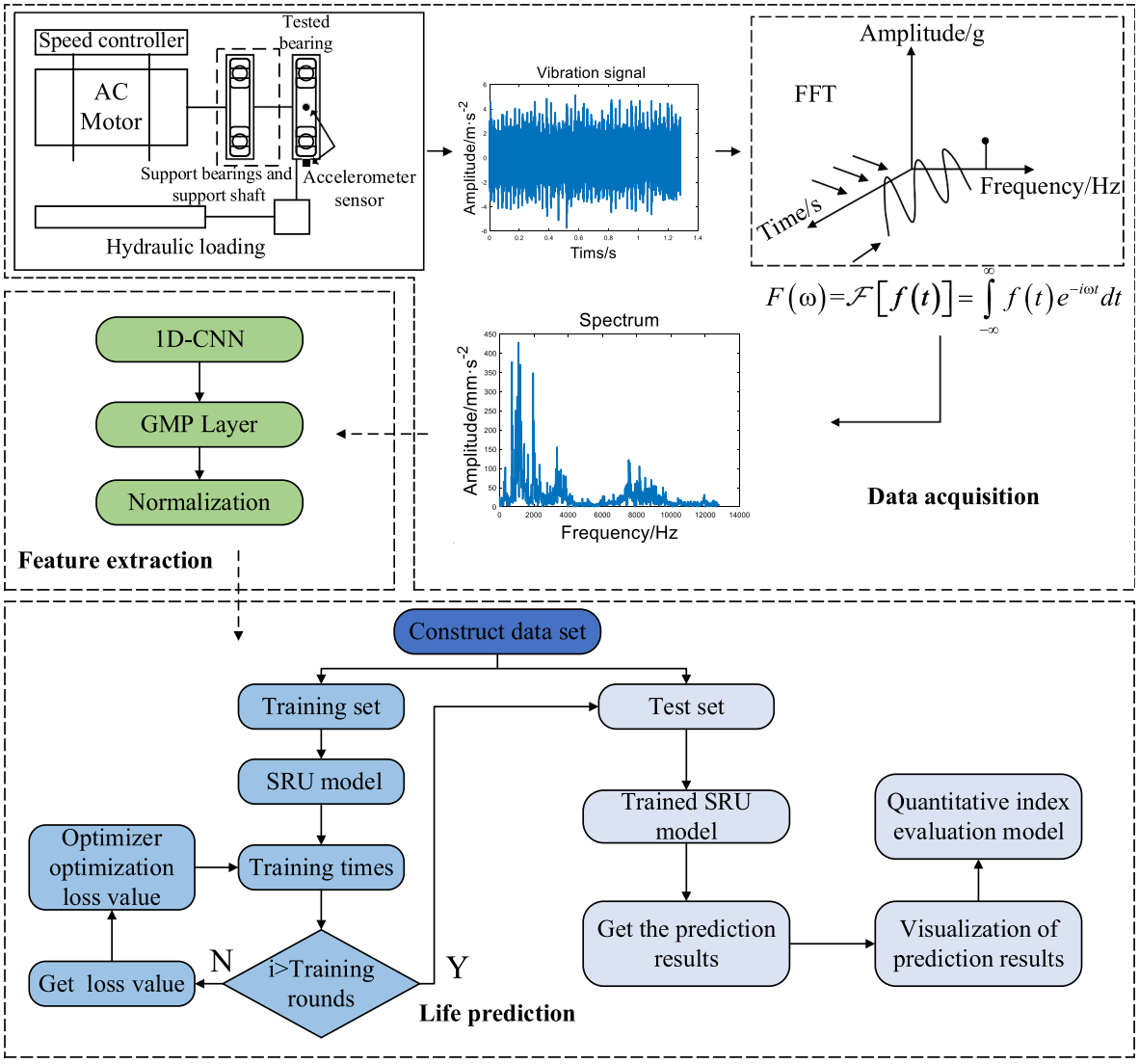

Fig. 9. Technology flowchart.

In Eq. (6), the cell state C_t is updated. In the first half, the forgetting gate f_t is multiplied by the old cell state C_{t-1} to determine how much information in the old state is forgotten. In the second half, the non-forgotten part is multiplied by the current input information $\widetilde{X}_t$ after linear transformation to determine how much new information is retained. Finally, the retained new information is added to the old cell state to form a new cell state C_t.

$$h_t = r_t \times tanh(c_t) + (1 - r_t) \times X_t \tag{7}$$

In Eq. (7), the output h_t at the current time is determined. First, after the new cell state C_t is activated by the activation function tanh, it is multiplied by the reset gate r_t to obtain what information is outputted from the new cell state C_t at the current time. The remaining information ready to be outputted is obtained by multiplying the part outside the reset gate r_t with the input X_t at the current time. The output h_t at this time is obtained by adding the two parts.

Therefore, it is discernible that SRU only relies on the update of the cell state C_t and the input X_t at the current time and does not rely on the hidden layer state h_{t-1} of the previous time; thus, a parallel and efficient information training process is realized to a certain extent.

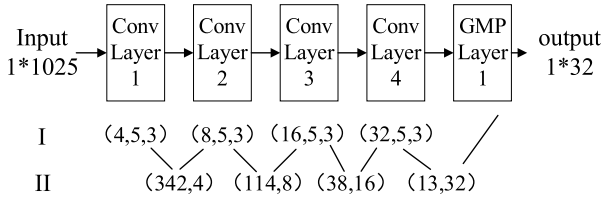

Fig. 10. 1D-CNN structure.

3. Proposed methodology

3.1. Technology flowchart

First, FFT was applied to the original signal to reduce the amount of original data and enhance information representation to make the information easier to be processed. Second, an improved 1D-CNN network was built for feature extraction. The extracted features were normalized, and the input dataset was constructed according to the degradation process. A SRU network was then built as the prediction model. Finally, the training set was used for model training by adjusting internal parameters and optimizing the model structure, and the test set was used to test the model effect. The aforesaid technical flowchart is presented in Fig. 9.

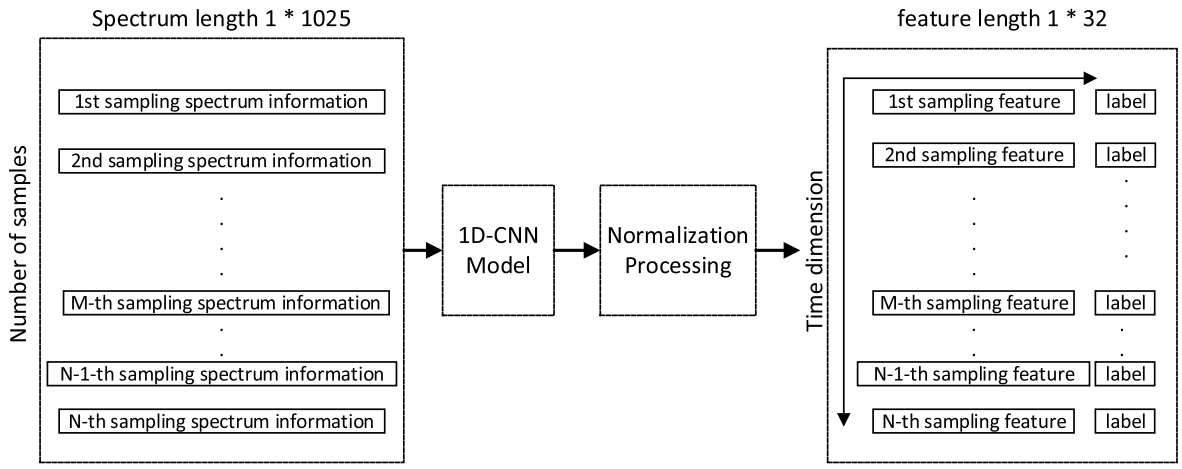

Fig. 11. Dataset construction.

3.2. Construction of 1D-CNN model

A five-layered one-dimensional convolution neural network was constructed, and its structure is presented in Fig. 10. The first four layers were convolution layers used to extract sequence features, and the last layer was the global maximum pooling layer used to output features.

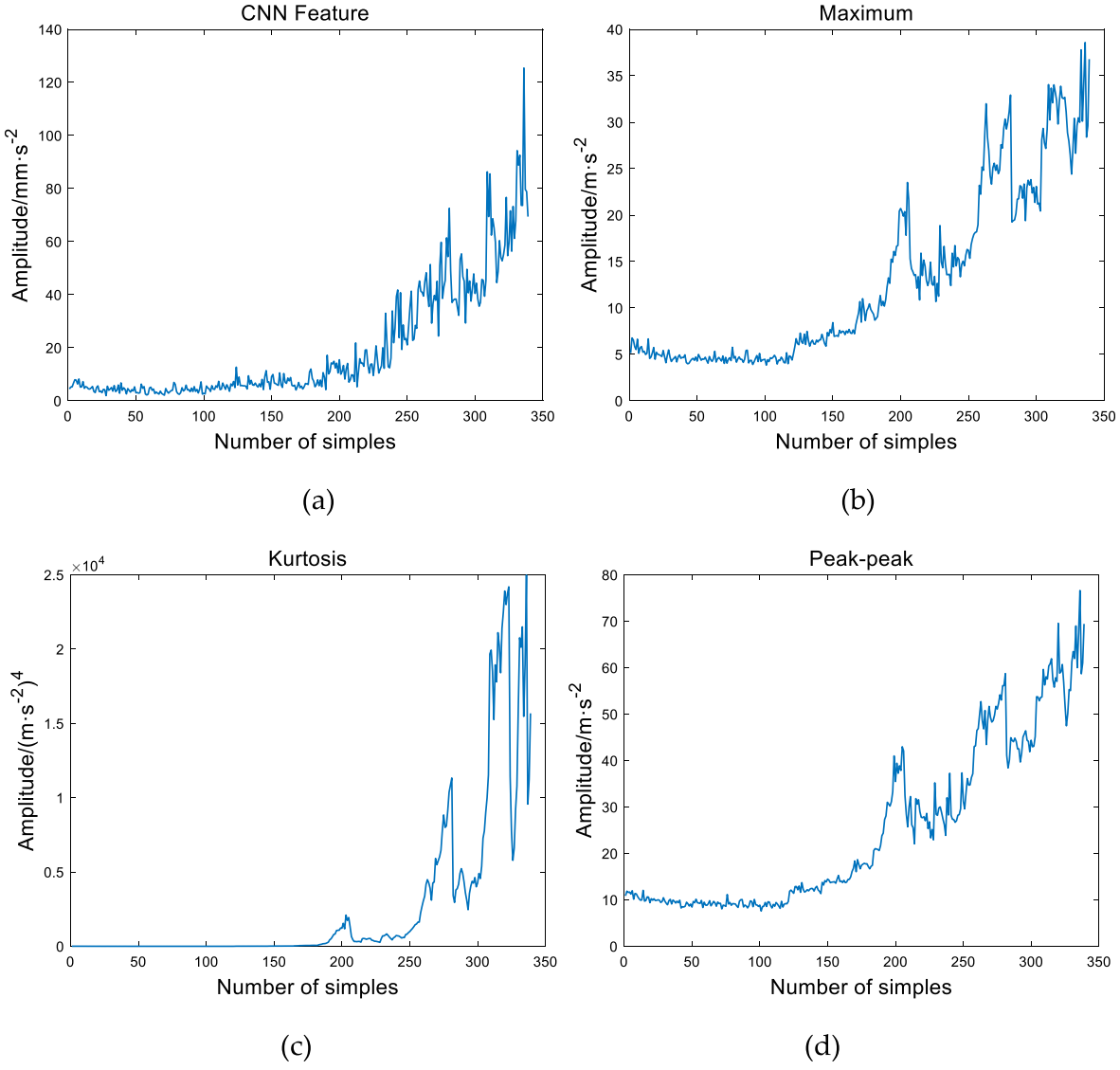

Fig. 12. Different feature indexes (a) CNN Feature (b) Maximum (c) Kurtosis (d) Peak-peak.

Row I represents the relevant parameters of convolution kernels in each layer. For example, in the sequence (4, 5, 3), 4 represents the number of convolution kernels in this layer, 5 represents the size of convolution kernels, and 3 represents the step size of convolution kernels in each move on this sequence. To ensure the extraction of boundary information, all convolution layers were filled with the "same" method.

The activation function was the ReLU function [39], and it was formulated as

$$f(x) = max(0, x) \tag{8}$$

Row II represents the size of the sequence after the convolution calculation of each layer. The calculation formula can be expressed as

$$L_1 = \lceil \frac{L}{s} \rceil \tag{9}$$

where L represents the length of the input sequence, s represents the step size, represents the rounding up, and L_1 is the output result.

3.3. Normalization

Data normalization is one of the common methods of data processing [40], and it reduces data to the range of [0–1]. Normalization is often used in some evaluation index processing methods to remove the restriction of units on data and facilitate the comparison or weighing between different indicators. The advantages of data normalization are presented below.

(1) The convergence speed and prediction accuracy of a model can be improved to some extent by data normalization.
(2) In some neural networks, normalization can effectively prevent gradient explosion.

The calculation formula for data normalization can be expressed as

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{10}$$

where x_{min} and x_{max}, respectively, represent the minimum and maximum values of feature data. After the normalization of feature data, the input dataset was constructed.

3.4. Feature extraction and data set construction

After converting the original data into spectrum information, the 1D-CNN model was used for feature extraction. After feature extraction and normalization, the input dataset was constructed according to the vertical time sequence and the horizontal feature number. The length of the vertical time series was determined by the number of samples, and 32 feature vectors of length 1 were outputted from the GMP layer and arranged in the horizontal direction. Each data took the current time life as the corresponding label. The dataset construction process is exhibited in Fig. 11.

The SRU model used for prediction adopted supervised learning. Therefore, normalization was also used to construct the label data for the actual remaining life, and the life label was reduced to the interval of [0, 1], where 1 represents the bearing at the beginning of degradation and 0 represents the bearing failure. The aim of supervised learning is to find the mapping relationship between an independent variable X and a dependent variable Y. After repeated training, a model learns this mapping relationship autonomously. When the model receives the new independent variable X, it tries to give the corresponding Y value to achieve prediction. Generally, the state of the system or mechanical parts at the next time of prediction should be determined according to the sampling interval of the first N groups of data used. The data sets used in this paper are all sampled in minutes, so the prediction results should also be in minutes.

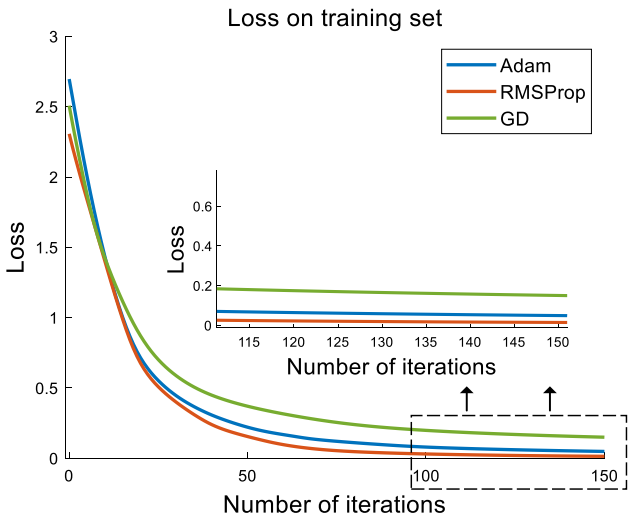

Fig. 13. Iteration processes of different optimizers.

3.5. Extraction features of 1D- CNN

Different feature indexes are displayed in Fig. 12. Fig. 12(b), (c), and (d) represent the maximum, the kurtosis, and the peak–peak, respectively. The features extracted from the 1D-CNN model are presented in Fig. 12(a). Among them, in the interpretation of physics, the maximum and peak-peak are dimensional indexes, and the unit is the same as the sampling value unit. The maximum represents the largest number in a section of signal, the expression is $X_{max} = max(x_i)$. The peak-peak represents the difference between the maximum and minimum in a section of signal, the expression is $X_{\mathrm{pp}} = max(x_i) - min(x_i)$. Kurtosis itself is a dimensionless index, it is a measure of the sharpness or flatness of probability density distribution, the expression is $K = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\sigma^4}$. However, for $\beta = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{x_i} - \bar{\mathrm{x}})^4$, the amplitude is processed to the fourth power, the high amplitude will be further highlighted, while the low amplitude is relatively suppressed, this feature will show trend or monotony. The extraction of effective features is the basis of reflecting the degradation state of bearings [41]. In traditional prediction, some feature indexes represented by (b), (c) and (d) in Fig. 12 are often used to reflect the bearing degradation state. For figure (a), the features extracted by CNN also have certain tendency and monotonicity. So, it can also be used as prediction features.

3.6. Optimizer selection

An optimizer is used to reduce the loss function value of a model and find the optimal solution to obtain suitable parameters of the model. Fig. 13 exhibits the iterative processes of loss values of different types of optimizers during training. It is noticeable that the performance of the traditional GD optimizer was poor at the beginning and end of iteration; hence, the loss value could not converge. As one of the most popular optimizers in neural networks, the Adam optimizer is widely used because of its strong adaptive adjustment ability. However, in the current experiment, the effect of the Adam optimizer was slightly worse than that of the RMSProp optimizer; thus, the RMSProp optimizer was selected for training. The RMSProp optimizer could correct the excessive swing to a certain extent and make the entire network converge faster.

4. Experimental verification

In the current section, the prediction accuracies of the proposed method and other models are compared, at the same time, the stability of the feature indexes and the training time of different models were

Table 1
LDK UER204 bearing parameters.

Parameter name	Numerical value	Parameter name	Numerical value
Inner race diameter	29.30 mm	Ball diameter	7.92 mm
Outer race diameter	39.80 mm	Number of balls	8
Bearing mean diameter	34.55 mm	Contact angle	0°
Load rating (dynamic)	12.82kN	Load rating (static)	6.65kN

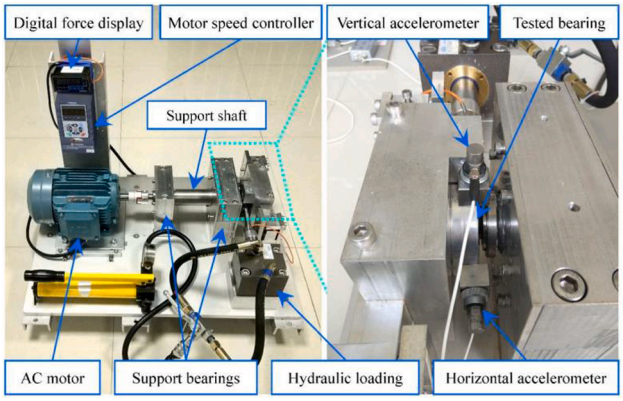

Fig. 14. Experimental platform (provided by XJTU-SY data set [42]).

evaluated.

4.1. Dataset introduction

The XJTU-SY dataset [42] was used for experimental verification. It is an open-source bearing life dataset from the Design Science and Basic Research Institute of Xi'an Jiaotong University. It contained 15 bearing life data under three different working conditions. The bearing designation used in the test is LDK UER204. The relevant parameters are shown in Table 1.

The experimental platform used to obtain XJTU-SY data set is shown in Fig. 14. The device consists of AC motor, motor speed controller, rotating shaft, supporting bearing, hydraulic loading system and test bearing. The test platform can adjust the radial force and speed. The radial force is generated by the hydraulic loading system and acts on the bearing seat of the test bearing and the speed is set and adjusted by the speed controller of AC motor.

The data under the second working condition were selected for the current experiment, under this condition —speed = 2250 rpm, additional vertical load = 11 kN. The sensor type selected in the test is PCB 352C33, sampling frequency = 25.6 kHz, sampling time = 1.28 s, sampling points = 32768, and sampling interval = 1 min. Each sampling data is saved as a csv file independently, each csv file contains two columns of data, the first column is the horizontal vibration signal, and the second column is the vertical vibration signal, as described in Section 2.1, the horizontal vibration signal is selected for all tests. The data of bearing2_5 and bearing2_2 were selected as the training set, the data of bearing2_4 were selected as the test set, take bearing2_5 as an example, 2 represents the second working conditions and 5 represents the fifth group data under the working conditions, other bearing data are named in the same way. There are 339 samples in bearing2_5, 161 samples in bearing2_2 and 42 samples in bearing2_4. The horizontal vibration signals in the whole life cycle of the three groups of data selected for the test are shown in Fig. 15.

Through the horizontal vibration signal, it can be seen that the degradation process of the three bearings has a certain similarity. They all run smoothly in the initial stage, and enter the degradation stage after a certain period of time. The whole degradation process is relatively smooth, and the vibration signal increases gradually from small to large. In general, the three bearings are a relatively slow degradation process, and there is no sudden failure and other abnormal conditions.

Two groups were randomly selected to represent the degradation trajectory of the bearing using the features extracted by CNN, as shown in Fig. 16. It can be seen from the images that the degradation trajectories of the three groups of bearings have certain similarity and

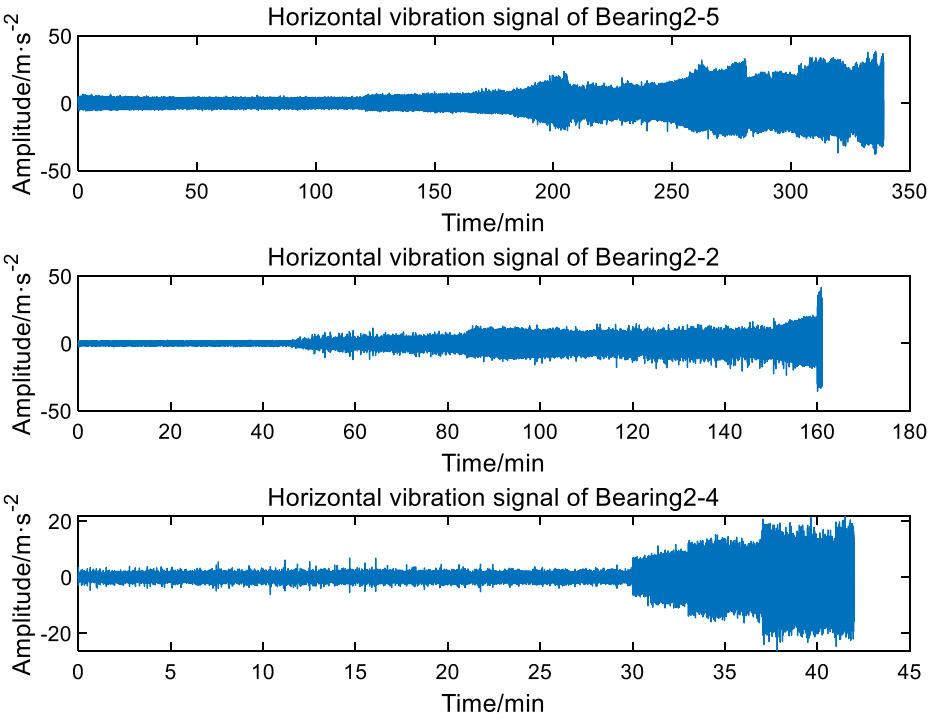

Fig. 15. Horizontal vibration signals of different bearings.

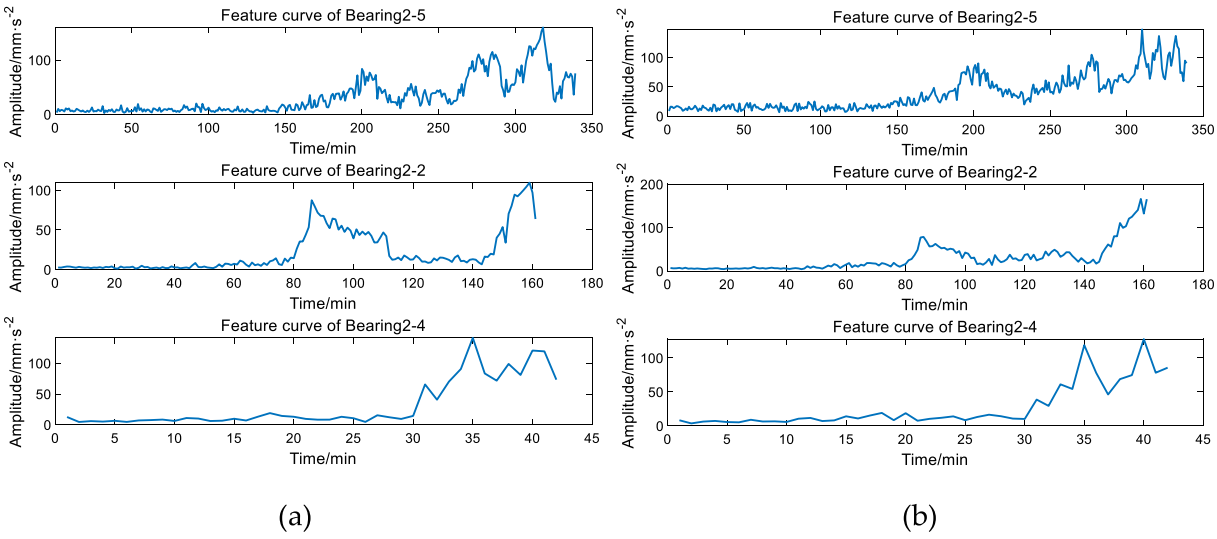

Fig. 16. Degradation curve (a) data group 1 (b) data group 2.

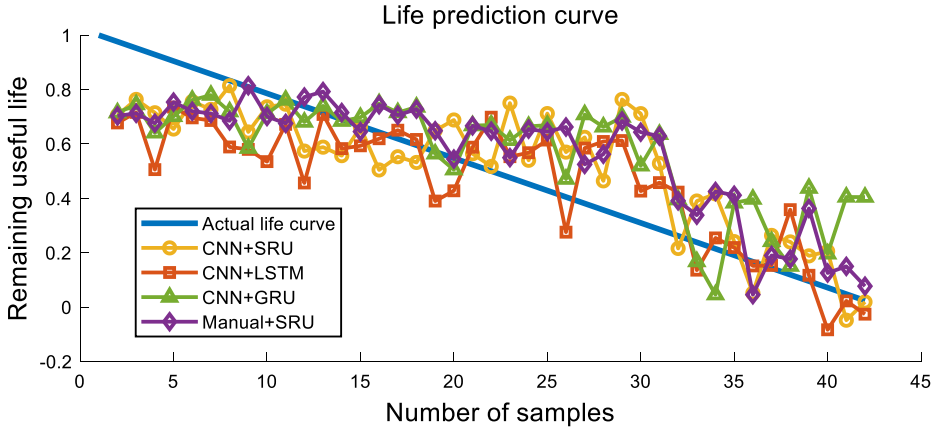

Fig. 17. Prediction effects of different models under Case 1.

monotonicity, and the variation range of feature amplitude is also similar. Therefore, according to this similarity, the SRU model is established by using the data of the whole life cycle of the training bearing, and then the nonlinear mapping relationship between the feature data of the test bearing and its life is established. That is, the life range reflected by the bearing with similar feature change law also has certain similarity.

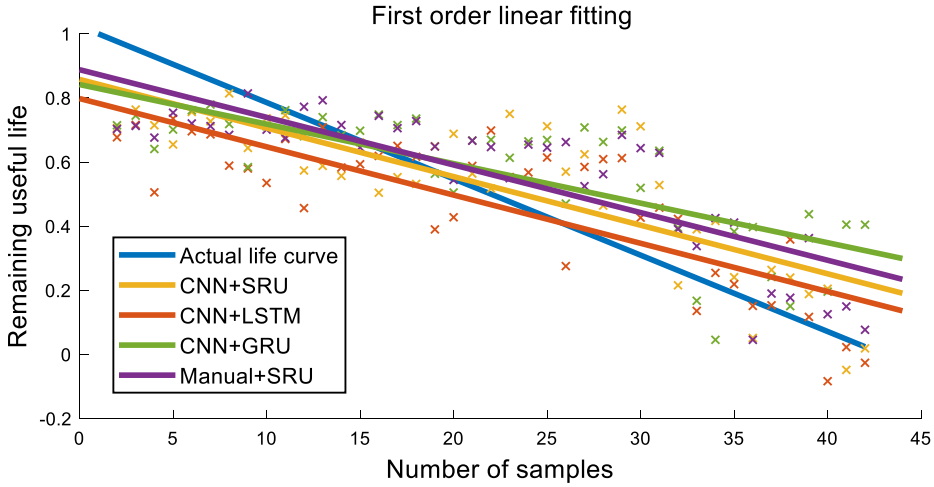

Fig. 18. First-order linear fitting curves of different models under Case 1.

Table 2
Evaluation indexes of different models under Case 1.

	CNN + SRU	CNN + LSTM	CNN + GRU	Manual + SRU
MAE	0.1416	0.1449	0.1591	0.1345
RMSE	0.1699	0.1778	0.1906	0.1637
Time	109.21 s	123.97 s	122.34 s	–

4.2. Metrics and evaluation

All samples were preprocessed by the above-mentioned method. After extracting the sample features, the input dataset was constructed and inputted into the model. To determine prediction accuracy, the mean absolute error (MAE) and the root mean square error (RMSE) were selected for quantitative evaluation [43]. The training time was calculated as the measurement standard.

$$MAE = (x,h) = \frac{1}{m}\sum_{i=1}^{m}\left|h\left(x^{(i)}\right) - y^{(i)}\right| \tag{11}$$

$$RMSE = (x,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right)^2} \tag{12}$$

The SRU model was designed based on the TensorFlow open-source deep learning framework. It had Intel (R) Core (TM) i5-9300h CPU (main frequency = 2.4 GHz), a Win10 64-bit operating system, and an NVIDIA GTX1660ti video card. CUDA 10.0 and cudnn7.3 were installed on the host to accelerate the operation of the card.

4.2.1. Case 1- prediction results of single training sets

In this group of experiments, using bearing2-5 as the training set, bearing2-4 as the test set. The maximum value, the kurtosis, the peak-peak value, and other manual features were extracted as contrast features, and LSTM and GRU were selected as comparison models. The prediction results, first-order fitting curves, and errors of different models are presented in Fig. 17, Fig. 18, and Table 2, respectively.

4.2.2. Case 2- prediction results of multiple training sets

In this group of experiments, in addition to the data of bearing2_5, the data of bearing2_2 were also added to the training set to form a new training set and bearing2_4 was still used for the test dataset. The new training dataset contained 500 samples, and the dataset construction method was consistent with that described in the previous section. The prediction results, first-order fitting curve, and errors of different models are presented in Fig. 19, Fig. 20, and Table 3, respectively.

4.3. Feature stability evaluation

4.3.1. Different combination features on the same dataset

In Case1 and Case2, different manual features are randomly com-

Table 3
Evaluation indexes of different models under Case 2.

	CNN + SRU	CNN + LSTM	CNN + GRU	Manual + SRU
MAE	0.1399	0.1405	0.1574	0.1324
RMSE	0.1726	0.1715	0.1841	0.1606
Time	149.89 s	171.41 s	170.06 s	–

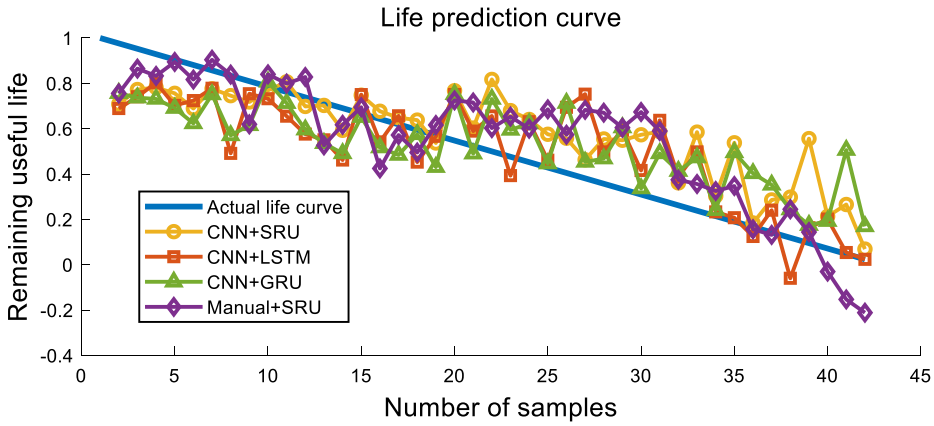

Fig. 19. Prediction effects of different models under Case 2.

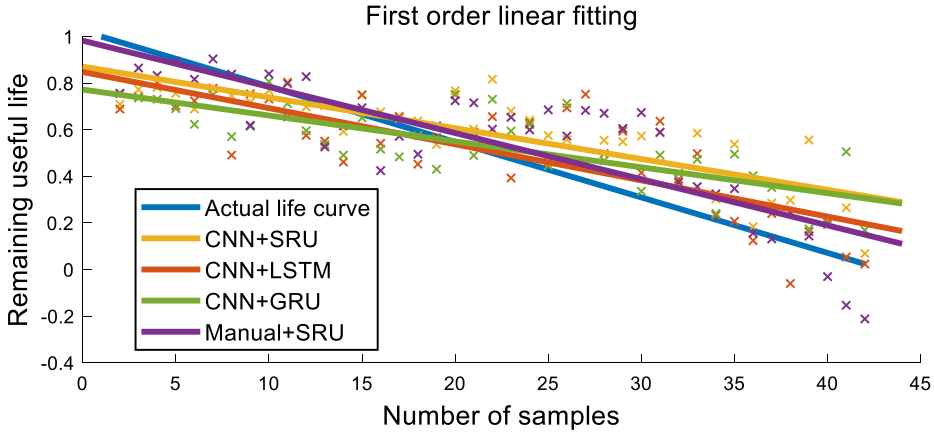

Fig. 20. First-order linear fitting curves of different models under Case 2.

Table 4
Standard deviations of different features.

	Case 1		Case 2	
	MAE-SD	RMSE-SD	MAE-SD	RMSE-SD
CNN + SRU	0.1176	0.1471	0.1137	0.1416
Manual + SRU	0.2977	0.3229	0.2883	0.3006

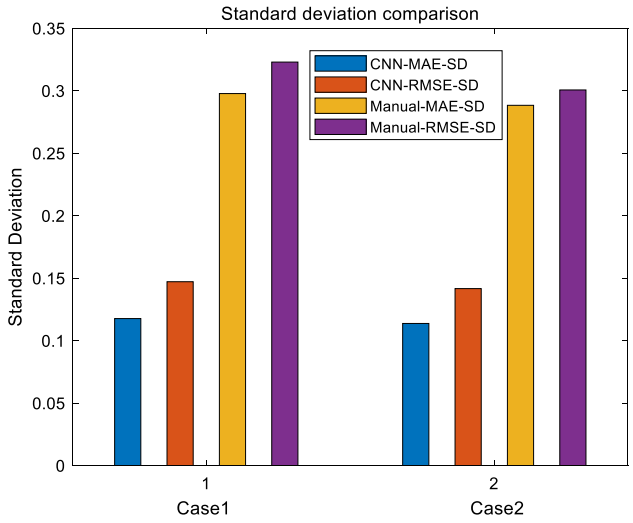

Fig. 21. Standard deviation comparison of different models.

bined to construct different data sets, at the same time, 1D-CNN is also used for feature extraction for many times. After many experiments, multiple MAE and RMSE are obtained. The standard deviations of multiple MAE and RMSE are calculated to evaluate the stability of feature indexes. The calculation formula of SD is expressed in Eq. (13), and the relevant data are presented in Table 4.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{13}$$

It is noticeable from Table 3 that the standard deviations of MAE and RMSE for CNN-extracted features were relatively small. Due to the randomness of network parameter initialization, the output features of the CNN model had certain differences; thus, the prediction results also had certain volatility. But it can be seen from the standard deviation that this is a relatively small range. However, the standard deviations of MAE and RMSE for manual features were relatively large, and the prediction effect was found to be the most unstable. It happened because the prediction effect of this method largely depended on the manually-extracted features. When the sensitivity of the selected features was high, the prediction effect of this method was relatively better. On the contrary, when the correlation of several features was low, the prediction effect became worse. In Table 1 and Table 2, when the prediction results using manual features had a better effect, the prediction results of the CNN-SRU model were compared with those of manual extraction, and MAE and RMSE both had errors of less than 10%. In Table 1, the errors of MAE and RMSE are 5.01% and 3.65%, respectively, whereas the corresponding values in Table 2 are 5.36% and 6.95%, respectively, which is an acceptable range in most case. In other cases, the prediction effect of the combined sets of manual features was inferior to that of the CNN-SRU model. The use of CNN instead of manual extraction could reduce manual intervention and labor cost to a large extent. In contrast, the prediction effect of CNN feature extraction was more stable. Fig. 21 depicts a comparison of the standard deviations of different features under Case 1 and Case 2.

Table 5
Error comparison of different methods.

	MAE	RMSE
CNN + SRU	0.03094	0.03751
Manual + SRU	0.04351	0.05267
Relative Error	28.89%	28.78%

4.3.2. *Same features on different datasets*

In Section 4.3.1, the prediction effects of different features on the same dataset are analyzed. In the current section, the prediction effects of the same feature index on different datasets are analyzed, several manual feature indexes which get better prediction results on XJTU-SY dataset are still used in this set of experiments. A part of the bearing data set from PHM2012 Data Challenge [44], jointly provided by IEEE and FEMTO-ST, was selected for a set of validation experiments. The bearing3_1 data under the working condition (3) of this dataset (speed = 1500 rpm, additional load = 5000 N, sampling frequency = 25.6 kHz, sampling time = 0.1 s, sampling interval = 10 s) were selected as the training set, and the first 100 sets of bearing3_3 samples were selected as prediction data. The remaining data construction method was consistent with the above-described approaches. The prediction effects and error

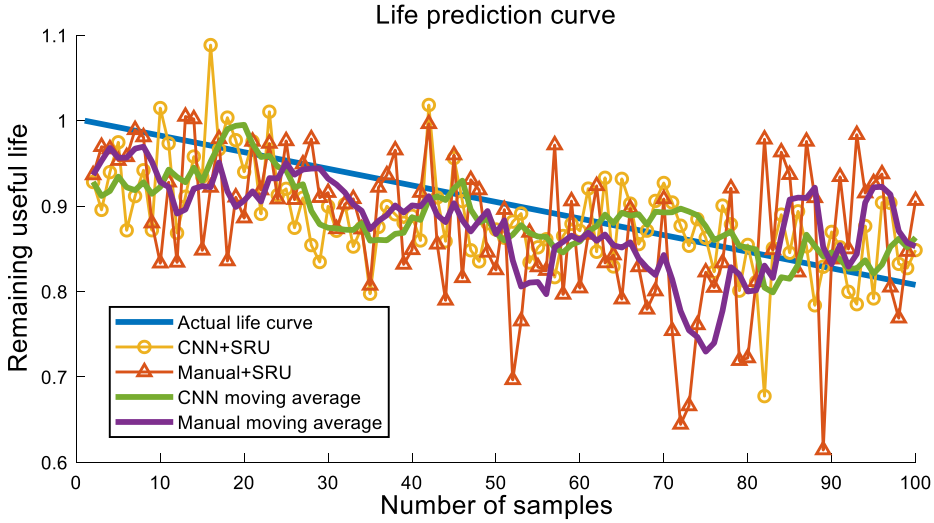

Fig. 22. Prediction effects of different methods.

Table 6
Training time comparison.

Training data	Number of hidden units	Step	SRU	LSTM	GRU
Case 1	12	20	65.76 s	72.63 s	73.81 s
		40	108.33 s	123.28 s	121.45 s
		60	148.56 s	173.50 s	169.40 s
	24	20	69.00 s	82.74 s	76.49 s
		40	108.80 s	145.09 s	133.03 s
		60	153.97 s	209.38 s	178.23 s
	36	20	69.47 s	101.36 s	81.89 s
		40	116.14 s	167.04 s	145.39 s
		60	169.67 s	243.27 s	207.13 s
Case 2	12	20	86.09 s	97.97 s	99.71 s
		40	147.98 s	172.30 s	169.84 s
		60	212.28 s	241.96 s	237.54 s
	24	20	87.30 s	119.84 s	104.60 s
		40	154.75 s	205.41 s	186.44 s
		60	216.04 s	283.93 s	269.26 s
	36	20	89.53 s	142.15 s	111.06 s
		40	153.42 s	237.14 s	206.91 s
		60	230.96 s	360.75 s	312.88 s

indexes of different models are presented in Fig. 22 and Table 5, respectively.

It is evident from Fig. 20 that the prediction effect of CNN-extracted features was better than that of manually-extracted features. The trend and volatility of the CNN curve were significantly better than those of the manual feature curves. This is because even when the same manual features were extracted, the generalization ability of these feature indicators was insufficient. In the PHM dataset, the sensitivity and monotonicity of these manual feature indicators were not obvious, resulting in a poor prediction effect. When the ability of these manual features is insufficient, to solve this problem, more complex signal processing methods are required to further screen or construct features, including raw data processing, feature selection and fusion, new feature index construction and so on. And these methods need more rich human experience and knowledge, at the same time, it further increases the cost and time consumption, too much human intervention reduces the intelligent degree of feature extraction. However, CNN feature extraction can avoid this problem to a certain extent. It ensures the validity and stability of features, meanwhile, increases the intelligence of feature extraction to a certain extent.

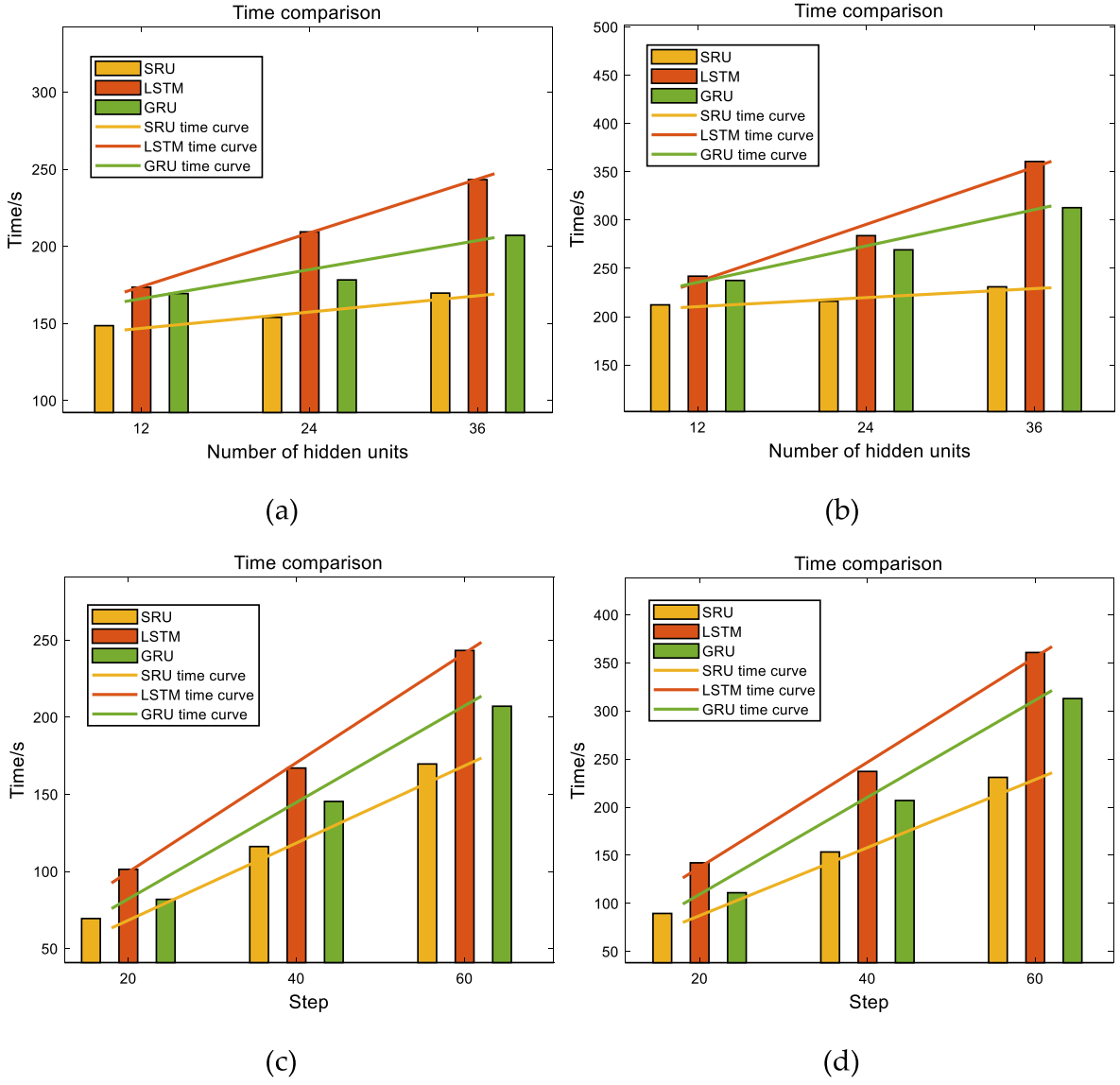

Fig. 23. Comparison of training times under different conditions: (a) Case 1 increases the number of hidden units, (b) Case 2 increases the number of hidden units, (c) Case 1 increases the number of steps, and (d) Case 2 increases the number of steps.

4.4. Model training speed evaluation

To verification the performance and training speed of the SRU model, several parameters including the data volume of the training set, the number of hidden layer units, and the step size were adjusted, and the obtained results are presented in Table 6. A part of the data was also selected for training time comparison (Fig. 23).

With the increase of the training data, the training time of the three models increased; however, the time consumption of the SRU model was the least. Moreover, the increase of the number of hidden layer units and the training step size had great impacts on the training time of the models. As the step size increased, the training time of the three models increased. LSTM had the longest training time followed by GRU, SRU model training time is also increased, but it consumes the least time. Furthermore, this change was more obvious with the increase of the number of hidden layer units. As both LSTM and GRU are serial input structures, their training time increased with the increase of hidden layer units and the curve amplitude and slope were also large. However, the curve amplitude and slope of the SRU model were very small, this is because its unique parallel structure is not affected by this.

When the same input dataset was used, no significant difference in the prediction effects of SRU and LSTM was observed; however, the prediction effect of GRU was slightly poor because it does not rely on the cell state in information transmission. On the premise of ensuring certain prediction accuracy, the SRU model had the least time consumption and the fastest training speed.

However, it should be noted that there is no direct relationship between the change of evaluation training speed and the improvement of prediction accuracy. The data in Table 5 only prove the superior performance in training speed of the SRU model. In fact, in any prediction, the first premise is to ensure prediction accuracy.

5. Conclusions

In the present paper, from the perspectives of improving the intelligence of roller bearing remaining useful life prediction, reducing manual intervention, and saving time and labor cost, a novel prediction method based on a 1D-CNN and a simple recurrent unit was proposed. After experimental verification, the following inferences were drawn.

(1) For feature extraction, the 1D-CNN model was used to replace traditional manual feature extraction methods. On the premise of ensuring feature effect, it reduced manual intervention, saved labor cost and time, and improved prediction intelligence.
(2) In the prediction process, the SRU model was used as the prediction algorithm. On the premise of ensuring the accuracy of prediction, the SRU model was faster in training than other models. It can significantly improve the operation efficiency and reduce the time consumption.

Experimental results revealed that in comparison to traditional methods, the 1D-CNN-SRU method has certain improvements in data front-end processing and back-end prediction, providing an economic and intelligent idea roller bearing remaining useful life prediction.

In future works, our research will attempt to improve the ability of SRU model to integrate information, build a bidirectional recurrent layer, hoping that it can focus on the information of the past and integrate the information of the future. Meanwhile, some innovation mechanisms will also be introduced into the model to strengthen the impact of important moment data on the results.

CRediT authorship contribution statement

Dechen Yao: . **Boyang Li:** . **Hengchang Liu:** Software. **Jianwei Yang:** . **Limin Jia:** Supervision, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The present work was funded by the National Natural Science Foundation of China (Grant No. 51975038), the Nature Science Foundation of Beijing, China (Grant No. L191005), the Support plan for the development of high-level teachers in Beijing municipal universities (Grant Nos. CIT&TCD201904062 and CIT&TCD201704052), the General Project of Scientific Research Program of Beijing Education Commission Grant No. (KM201810016015), the Scientific Research Fund of Beijing University of Civil Engineering Architecture (Grant No. ZF15068), the BUCEA Post Graduate Innovation Project(PG2020088), the Fundamental Research Funds for Beijing Universities(X20071) and the Fundamental Research Funds for Beijing University of Civil Engineering and Architecture (Grant No. X18133).

References

[1] A. Rai, S. Upadhyay, A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings, Tribol. Int. (2016) 289–306, https://doi.org/10.1016/j.triboint.2015.12.037.
[2] D. Yao, H. Liu, J. Yang, X. Li, A lightweight neural network with strong robustness for bearing fault diagnosis, Measurement 159 (2020), https://doi.org/10.1016/j.measurement.2020.107756.
[3] H. Liu, D. Yao, J. Yang, X. Li, Lightweight convolutional neural network and its application in rolling bearing fault diagnosis under variable working conditions, Sensors (Basel) 22 (2019), https://doi.org/10.3390/s19224827.
[4] L. Dhamande, M. Chaudhari, Compound gear-bearing fault feature extraction using statistical features based on time-frequency method, Measurement 125 (2018) 63–77, https://doi.org/10.1016/j.measurement.2018.04.059.
[5] L. Saidi, J. Ali, E. Bechhoefer, M. Benbouzid, Wind turbine high-speed shaft bearings health prognosis through a spectral Kurtosis-derived indexes and SVR, Appl. Acoust. 120 (2017) 1–8, https://doi.org/10.1016/j.apacoust.2017.01.005.
[6] G. Qiu, Y. Gu, J. Chen, Selective health indicator for bearings ensemble remaining useful life prediction with genetic algorithm and Weibull proportional hazards model, Measurement 150 (2019) 107097, https://doi.org/10.1016/j.measurement.2019.107097.
[7] S. Gao, S. Zhang, Y. Zhang, Y. Gao, Operational reliability evaluation and prediction of rolling bearing based on isometric mapping and NoCuSa-LSSVM, Reliab. Eng. Syst. Saf. 201 (2020) 106968, https://doi.org/10.1016/j.ress.2020.106968.
[8] H. Wang, G. Ni, J. Chen, J. Qu, Research on rolling bearing state health monitoring and life prediction based on PCA and Internet of things with multi-sensor, Measurement 157 (2020), https://doi.org/10.1016/j.measurement.2020.107657.
[9] A. Rai, S. Upadhyay, An integrated approach to bearing prognostics based on EEMD-multi feature extraction, Gaussian mixture models and Jensen-Rényi divergence, Appl. Soft Comput. 71 (2018) 36–50, https://doi.org/10.1016/j.asoc.2018.06.038.
[10] K. Noman, D. Wang, Z. Peng, Q. He, Oscillation based permutation entropy calculation as a dynamic nonlinear feature for health monitoring of rolling element bearing, Measurement 172 (2021), https://doi.org/10.1016/j.measurement.2020.108891.
[11] L. Xu, P. Pennacchi, S. Chatterton, A new method for the estimation of bearing health state and remaining useful life based on the moving average cross-correlation of power spectral density, Mechanical systems and signal processing 139 (May) (2020) 106617.1–106617.17, https://doi.org/10.1016/j.ymssp.2020.106617.
[12] J. Castilla-Gutiérrez, J. Fortes, J. Davila, Control and prediction protocol for bearing failure through spectral power density, Eksploatacja i Niezawodnosc – Maintenance Reliab. 22 (4) (2020) 651–657. https://doi.org/10.17531/ein.2020.4.8.
[13] J. Blaut, Ł. Breńkacz, Application of the Teager-Kaiser energy operator in diagnostics of a hydrodynamic bearing, Eksploatacja i Niezawodnosc – Maintenance Reliab. 22 (4) (2020) 757–765. https://doi.org/10.17531/ein.2020.4.20.
[14] B. Rezaeianjouybari, Y. Shang, Deep learning for prognostics and health management: state of the art, challenges, and opportunities, Measurement (2020) 163, https://doi.org/10.1016/j.measurement.2020.107929.
[15] A. Hinchi, M. Tkiouat, Rolling element bearing remaining useful life estimation based on a convolutional long-short-term memory network, Procedia Comput. Sci. 127 (2018) 123–132, https://doi.org/10.1016/j.procs.2018.01.106.
[16] N. Ding, H. Li, Z. Yin, N. Zhong, L. Zhang, Journal bearing seizure degradation assessment and remaining useful life prediction based on long short-term memory

neural network, Measurement 166 (2020) 108215, https://doi.org/10.1016/j.measurement.2020.108215.
[17] F. König, J. Marheineke, G. Jacobs, C. Sous, M. Zuo, Z. Tian, Data-driven wear monitoring for sliding bearings using acoustic emission signals and long short-term memory neural networks, Wear (2021), https://doi.org/10.1016/j.wear.2021.203616.
[18] J. Chen, H. Jing, Y. Chang, Q. Liu, Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process, Reliab. Eng. Syst. Saf. 185 (may) (2019) 372–382, https://doi.org/10.1016/j.ress.2019.01.006.
[19] L. Xiao, Z. Liu, Y. Zhang, Y. Zheng, C. Cheng, Degradation assessment of bearings with trend-reconstruct-based features selection and gated recurrent unit network, Measurement 108064 (2020), https://doi.org/10.1016/j.measurement.2020.108064.
[20] M. Dangut, Z. Skaf, I. Jennions, Rare failure prediction using an integrated auto-encoder and bidirectional gated recurrent unit network, IFAC-PapersOnline 53 (3) (2020) 276–282, https://doi.org/10.1016/j.ifacol.2020.11.045.
[21] D. She, M. Jia, A BiGRU method for remaining useful life prediction of machinery, Measurement 167 (2020) 108277, https://doi.org/10.1016/j.measurement.2020.108277.
[22] S. Behera, R. Misra, A. Sillitti, Multiscale deep bidirectional gated recurrent neural networks based prognostic method for complex non-linear degradation systems, Inf. Sci. 554 (2021) 120–144, https://doi.org/10.1016/j.ins.2020.12.032.
[23] M. Aliabadi, H. Emami, M. Dong, Y. Huang, Attention-based recurrent neural network for multistep-ahead prediction of process performance, Comput. Chem. Eng. 140 (2020), https://doi.org/10.1016/j.compchemeng.2020.106931.
[24] J. Jiang, J. Lee, Y. Zeng, Time series multiple channel convolutional neural network with attention-based long short-term memory for predicting bearing remaining useful life, Sensors 20 (1) (2019), https://doi.org/10.3390/s20010166.
[25] Y. Chen, G. Peng, Z. Zhu, S. Li, A novel deep learning method based on attention mechanism for bearing remaining useful life prediction, Appl. Soft Comput. 86 (2020), https://doi.org/10.1016/j.asoc.2019.105919.
[26] A. ElSaid, F. Jamiy, J. Higgins, B. Wild, T. Desell, Optimizing long short-term memory recurrent neural networks using ant colony optimization to predict turbine engine vibration, Appl. Soft Comput. 73 (2018) 969–991, https://doi.org/10.1016/j.asoc.2018.09.013.
[27] H. Nguyen, J. Liu, E. Zio, A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators, Appl. Soft Comput. 89 (2020), https://doi.org/10.1016/j.asoc.2020.106116.
[28] A. Rifai, H. Aoyama, N. Tho, S. Md Dawal, N. Masruroh, Evaluation of turned and milled surfaces roughness using convolutional neural network, Measurement 161 (2020), https://doi.org/10.1016/j.measurement.2020.107860.
[29] B. Wang, Y. Lei, N. Li, T. Yan, Deep separable convolutional network for remaining useful life prediction of machinery, Mech. Syst. Sig. Process. 134 (2019), https://doi.org/10.1016/j.ymssp.2019.106330.
[30] H. Cheng, X. Kong, G. Chen, Q. Wang, R. Wang, Transferable convolutional neural network based remaining useful life prediction of bearing under multiple failure behaviors, Measurement 168 (2021), https://doi.org/10.1016/j.measurement.2020.108286.
[31] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D. Inman, 1D convolutional neural networks and applications: a survey, Mech. Syst. Sig. Process. 151 (2021), https://doi.org/10.1016/j.ymssp.2020.107398.
[32] P. Kumar, A. Hati, Deep convolutional neural network based on adaptive gradient optimizer for fault detection in SCIM, ISA Trans. (2020), https://doi.org/10.1016/j.isatra.2020.10.052.
[33] P. Kundu, A. Darpe, M. Kulkarni, Weibull accelerated failure time regression model for remaining useful life prediction of bearing working under multiple operating conditions, Mech. Syst. Sig. Process. 134 (2019), https://doi.org/10.1016/j.ymssp.2019.106302.
[34] V. Kulkarni, A. Khadersab, Spectrum analysis of rolling element bearing faults, Mater. Today: Proc. (2020), https://doi.org/10.1016/j.matpr.2020.10.296.
[35] O. Fink, Q. Wang, M. Svensén, P. Dersin, W. Lee, M. Ducoffe, Potential, challenges and future directions for deep learning in prognostics and health management applications, Eng. Appl. Artif. Intell. 92 (2020), https://doi.org/10.1016/j.engappai.2020.103678.
[36] X. Cui, Z. Chen, F. Yin, Speech enhancement based on simple recurrent unit network, Appl. Acoust. 157 (2020), https://doi.org/10.1016/j.apacoust.2019.107019.
[37] K. Cho, B. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP 2014–2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, pp. 1724–1734, https://doi.org/10.3115/v1/d14-1179.
[38] T. Lei, Y. Zhang, S. Wang, H. Dai, Y. Artzi, Simple recurrent units for highly parallelizable recurrence. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.
[39] D. Yao, Q. Sun, J. Yang, H. Liu, J. Zhang, C. Shen, Railway fastener fault diagnosis based on generative adversarial network and residual network model, Shock Vib. (2020) 1–15, https://doi.org/10.1155/2020/8823050.
[40] D. Yao, H. Liu, J. Yang, J. Zhang, Implementation of a novel algorithm of wheelset and axle box concurrent fault identification based on an efficient neural network with the attention mechanism, J. Intell. Manuf. (2020), https://doi.org/10.1007/s10845-020-01701-y.
[41] P. Kumar, L. Kumaraswamidhas, S. Laha, Selection of efficient degradation features for rolling element bearing prognosis using Gaussian Process Regression method, ISA Trans. (2020), https://doi.org/10.1016/j.isatra.2020.12.020.
[42] B. Wang, Y. Lei, N. Li, N. Li, A hybrid prognostics approach for estimating remaining useful life of rolling element bearings, IEEE Trans. Reliab. 69 (1) (2020) 401–412, https://doi.org/10.1109/tr.2018.2882682.
[43] W. Wang, Y. Lu, Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in assessing rounding model, IOP Conf. Ser.: Mater. Sci. Eng. 324 (2018), https://doi.org/10.1088/1757-899X/324/1/012049.
[44] P. Nectoux, R, Gouriveau, K. Medjaher, E. Ramasso, B. Chebel-Morello, N. Zerhouni, C. Varnier, PRONOSTIA: An experimental platform for bearings accelerated degradation tests, in: IEEE International Conference on Prognostics and Health Management, IEEE, 2012.