Hindawi
Shock and Vibration
Volume 2020, Article ID 8888627, 16 pages
https://doi.org/10.1155/2020/8888627

Research Article

Sequential Network with Residual Neural Network for Rotatory Machine Remaining Useful Life Prediction Using Deep Transfer Learning

Hao Zhang,[1] Qiang Zhang,[2] Siyu Shao ,[2] Tianlin Niu,[2] Xinyu Yang,[2] and Haibin Ding[3]

[1]*The Graduate School of Air Force Engineering University, Xi'an 710051, China*
[2]*Air and Missile Defense College of Air Force Engineering University, Xi'an 710051, China*
[3]*Training Base of Army Engineering University, Xuzhou 221004, China*

Correspondence should be addressed to Siyu Shao; cathygx.sy@gmail.com

Received 26 June 2020; Revised 27 August 2020; Accepted 3 September 2020; Published 14 September 2020

Academic Editor: Changqing Shen

Deep learning has a strong feature learning ability, which has proved its effectiveness in fault prediction and remaining useful life prediction of rotatory machine. However, training a deep network from scratch requires a large amount of training data and is time-consuming. In the practical model training process, it is difficult for the deep model to converge when the parameter initialization is inappropriate, which results in poor prediction performance. In this paper, a novel deep learning framework is proposed to predict the remaining useful life of rotatory machine with high accuracy. Firstly, model parameters and feature learning ability of the pretrained model are transferred to the new network by means of transfer learning to achieve reasonable initialization. Then, the specific sensor signals are converted to RGB image as the specific task data to fine-tune the parameters of the high-level network structure. The features extracted from the pretrained network are the input into the Bidirectional Long Short-Term Memory to obtain the RUL prediction results. The ability of LSTM to model sequence signals and the dynamic learning ability of bidirectional propagation to time information contribute to accurate RUL prediction. Finally, the deep model proposed in this paper is tested on the sensor signal dataset of bearing and gearbox. The high accuracy prediction results show the superiority of the transfer learning-based sequential network in RUL prediction.

1. Introduction

Rotatory machine serves as a significant element in mechanical systems while its working conditions are directly related to the production efficiency and production safety [1–3]. However, due to complex operating environment and component wear, machine failure is inevitable during practical operation. Once failure occurs, it may have a negative effect on the whole mechanical system and even threaten the safety and stability of the industrial production. Therefore, it is necessary to conduct an effective and efficient predictive maintenance strategy which helps ensure reliability service for the mechanical system [4, 5]. Accurate prediction of the remaining useful life (RUL) for the rotatory machines contributes to preventing possible failures, as potential faults are identified in advance and removed in a timely manner. Hence, an appropriate maintenance strategy can be achieved based on the predicted results, realizing system efficiency, and quality improvement [6, 7].

Continued advancement in intelligent manufacturing has led to ever-increasing attention on system Prognostic and Health Management (PHM) in industry and academia [8]. Traditional fault prediction methods are mainly based on manual extraction of features, which requires prior knowledge as the basis. The performance of the traditional prediction model largely depends on the quality and applicability of the hand-crafted features. When the selected features are unsuitable for the certain task, the predictive accuracy may dramatically fall [9, 10]. Nowadays, data-driven RUL model is able to utilize historical data directly to

build prediction model without prior knowledge or feature extraction, which is able to model complex process of mechanical degradation [11–13]. Owing to the extraordinary feature learning ability, deep learning-based methods has gained great popularity in industrial applications as it overcomes the limitations of traditional prediction methods. By constructing the deep neural network with multiple hidden layers, the framework is able to learn hierarchical representations directly from the original data [14–16]. Deep neural networks automatically extract distinctive representations through model training and therefore obtain accurate prediction results [17–19]. By now, deep learning has various successful applications including computer vision (CV) [20], natural language processing (NLP) [21], speech recognition [22], machine translation [23], and automatic driving [24].

Similarly, deep learning has achieved remarkable achievements in the field of machine RUL prediction. Deutsch et al. proposed a sensor vibration signal RUL prediction method, which utilized learning ability and prediction ability of the deep belief network to realize automatic feature extraction and RUL prediction without manual intervention [25]. Qin et al. proposed an attention-based Long Short-Term Memory (LSTM) network which utilizes attention coefficients to evaluate the importance of intermediate information and its superiority in RUL prediction is verified by gear dataset [26]. Although deep learning-based methods have been successfully applied in the field of mechanical system degradation modeling, there are still several deficiencies. Firstly, due to the limited data availability, training samples are insufficient to conduct the adequate training of the deep model. Consequently, the depth of most deep model is limited, which directly affects the final prediction performance [27]. Secondly, with the increase of the number and scale of hidden layers, model parameters will also increase sharply, where training such a model from scratch is time-consuming. Besides, the selection of hyperparameters (learning rate, activation function, loss function, etc.) may also greatly influence the performance of the model. To overcome the difficulty of deep model training, transfer learning is applied.

To accelerate model training, model parameters that have been learned and trained in advance are transferred to the new model, which greatly improves the efficiency of feature learning procedure [28]. Transfer learning, which provides a reasonable initialization for the target model and simplifies the fine-tuning procedure, has been successfully applied in the field of fault prediction [29]. Sun utilized sparse autoencoder to build a deep model which is pretrained by run-to-failure data with RUL information from a cutting tool. The trained network is then transferred to a new model to achieve accurately RUL prediction [30]. Shao et al. proposed a transfer learning network using the structure and parameters of VGG-16 and conducted accurate fault classification among different mechanical datasets [31].

In order to achieve accurate prediction of sequential sensor signals, deep learning-based sequential models are necessary. LSTM network is commonly applied to deal with time series, which has a strong capability in discovering the underlying variation pattern. LSTM is a variant of Recurrent Neural Network (RNN) and is proposed to overcome the problem of information redundancy caused by long input, therefore achieving the desired performance in the applications of signal prediction [32]. Yuan et al. discussed several applications of RNN, LSTM, and Gated Recurrent Unit (GRU) models in aeroengine fault diagnosis, and experiments proved that the models based on LSTM and GRU had better prediction performance than RNN-based models [33]. Wang et al. designed a residual LSTM network framework to solve the degradation problem of the deep LSTM model and verified the superiority of the structure by experiments [34]. Zhang et al. proposed that a new approach based on the LSTM network models the system degradation process, and it has the capability to learn specific patterns from time series [35].

Inspired by these prior researches, a novel RUL prediction model based on Bidirectional LSTM and transfer learning strategy is proposed. The proposed model utilizes the first two convolution blocks of Residual Network (ResNet-50) as sensor signal feature extractor and outputs the predicted RUL values by Bidirectional LSTM. The main contributions of this paper are summarized as follows:

(1) A novel RUL prediction framework of rotatory machine based on sequential network is proposed and combined with transfer learning strategy to improve the training efficiency. A pretrained network trained by ImageNet dataset is designed as a feature extractor at the first stage and then the advanced parameters of the whole framework are fine-tuned by specific sensor signals of rotatory machine, which greatly reduces the difficulty of deep network training and realizes efficient RUL prediction.

(2) For accurate RUL prediction, the Bidirectional LSTM module is combined with pretrained feature extractor. By the powerful feature learning capability of CNN and the prediction capability of the Bidirectional LSTM, the RUL value of the rotatory machine can be accurately predicted.

(3) In the current research, transfer learning has not been applied to the field of rotatory machine remaining useful life prediction. Different from the traditional fault classification problem, the prediction of RUL requires the network output specific value rather than a category label, which greatly increases the difficulty of model training. The proposed method in this paper extends the application scope of transfer learning.

The rest of the paper is organized as follows. Section 2 introduces the theoretical background of the proposed method, including convolutional neural networks, residual networks, transfer learning, and Bidirectional LSTM. In Section 3, the overall RUL prediction framework is illustrated in detail. In Section 4, experimental studies are carried out to verify the effectiveness of the proposed model, together with performance comparisons with other methods. Conclusions and future work are presented in Section 5.

2. Methodology

2.1. Basic Theory of Convolutional Neural Network. Convolutional neural network (CNN) is a feedforward neural network using local connection and weights sharing strategy. The weights sharing strategy reduces the number of model parameters and improves the computing efficiency effectively. Deep convolutional neural networks automatically learn the potential abstract features from the input data and achieve accurate prediction by capturing the high-level feature representations. CNN contains three main types of building units, including convolutional layer, pooling layer, and fully connected layer. Figure 1 shows a typical CNN architecture.

Convolutional layer realizes convolution operation, and each convolutional layer generates multiple feature maps. The input data are mapped into different feature space through various convolutional kernels, where each convolutional kernel represents one certain feature extraction. Within one convolutional layer, only a certain part of the input is connected to the corresponding neuron, called local connection. The weights associated with different neurons within the same layer are the same, called weights sharing, and weights and bias shared within each layer form a convolution kernel. This sharing strategy greatly reduces the number of model parameters and helps deep network training. Formula (1) is the calculation method of feature mapping:

$$\alpha_n = W_n \otimes x + b_n, \tag{1}$$

$$Z_k = f(\alpha_k), \tag{2}$$

where $\alpha_n$is the characteristic value extracted by the n-th convolutional layer, Z_k is the feature mapping value extracted after activation function, $f(\,)$ represents the nonlinear activation function where Rectified Linear Unit (ReLU) is widely used in deep architectures, x is the input image, W_n and b_n represent weights and bias values, and $\otimes$ is a two-dimensional convolution operation, which performs dot product of convolution kernels and the input. Pooling layer usually follows the convolutional layer to reduce feature dimension and remove redundant information. By compressing the features appropriately, the network computation complexity is simplified and the robustness of feature extraction is improved.

2.2. ResNet-50. As the network deepens, constructing more hidden layers may not achieve performance improvement or even cause model degradation. Redundant hidden layers may lead to a decrease in model convergence rate in the training process and affect the predictive accuracy consequently. To address this issue, Residual Neural Network (ResNet) is proposed, which adopts a skip-connection strategy to reinforce feature learning ability and therefore effectively expands the depth of the network, improving model feature learning ability. Skip connection between stacked cells sends useful information directly to the next layer which is an effective way to avoid gradient vanishing.

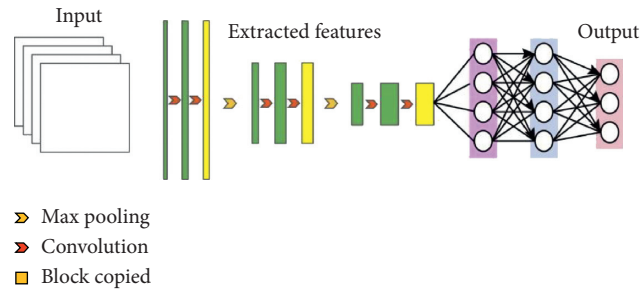

Figure 1: The typical architecture of CNN.

Figure 2 shows the skip-connection idea of ResNet-50. The output y is expressed as combining the linear superposition of the input x with a nonlinear transform $F(x)$ of the input. Instead of learning direct features from input x, ResNet tries to learn the difference between the expected features and input x, which is called residual. When the so-called residual approaches 0, it means that the stacked layer conducts the identity mapping which at least ensures that the stacked model will not degrade in performance as the network deepens. Actually, the residual part is hardly null and based on the learned representations from previous layers, it can help the model learn new features.

Table 1 and Figure 3 show the detailed information of ResNet-50 [36]. ResNet-50 is mainly composed of four residual blocks and one fully connected layer. Each residual block consists of several convolutional layers, which have different convolution kernel sizes. Convolution operation is performed on the input, and then features are extracted through different residual blocks. Finally, fully connected layer is leveraged to output the corresponding targets. Different from conventional neural network where the output of the $(n-1)$-th layer is only connected to the n-th layer as the input, the skip-connection structure of Residual Network enables the output directly cross several layers, which solves the gradient dissipation problem in the backpropagation process and makes it easy to train.

2.3. Transfer Learning. For a large deep neural network, training from scratch requires sufficient labeled data and the training procedure is time-consuming. To overcome this problem, transfer learning is used to boost the training performance of large model, which aims at solving the training difficulty when training data is limited. The problem of overfitting may occur when the training samples are insufficient, thus limiting the generalization ability of the deep model. Transfer learning strategy pretrains model parameters with a large set of sufficient data and transfers the well-trained model to a new specific task. Through fine-tuning the weight parameters of the higher hidden layer with a small number of new data, the higher-level representations can be achieved. Whether the transfer learning strategy is effective depends on the difference between the data used in the pretrained process and the data for fine-tuning. For datasets with similar tasks, only the fully connected layer parameters at the end of the network need to be fine-tuned, without changing the parameters of the overall model. For

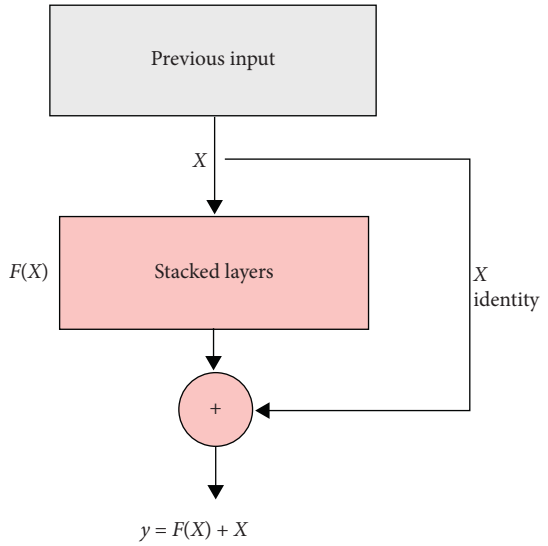

Figure 2: The diagram of residual connection.

Table 1: The detailed information about ResNet-50.

Layer name	Output size	Type
Input	$224 \times 224 \times 3$	None
Conv1	112×112	7×7, 64, stride 2 3×3 max pool, stride 2
Conv2	56×56	$\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$
Conv3	28×28	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$
Conv4	14×14	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 6$
Conv5	7×7	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$
–	1×1	Avg pool, 1000-d FC

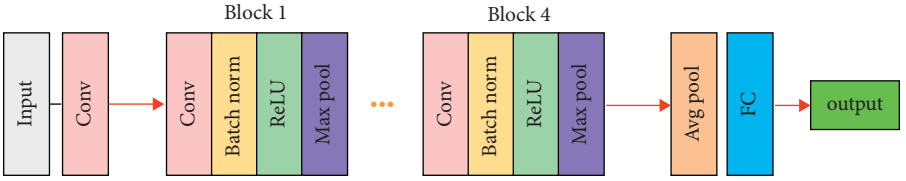

Figure 3: The detailed structure of ResNet-50.

datasets with great differences, most of the convolution block parameters need to be appropriately updated. In general, compared with the methods of trained from scratch, transfer learning reduces the number of parameters that need to be trained and enables the model to converge faster.

Figure 4 shows the process of transfer learning. Model parameters are transferred from source domain (image data in the picture) to target domain (vibration signal in the picture), where parameter transfer helps new model realize quick convergence. In general, training data in the source domain is sufficient, while training data for target domain is limited. Based on the similarity between target domain and source domain, fine-tuning corresponding weights and bias in the new model can greatly improve the predictive accuracy in the case of insufficient samples. The pretrained model leverages public dataset to learn the general features at the lower layers, and then the transfer model extracts abstract features at the higher layers through the limited

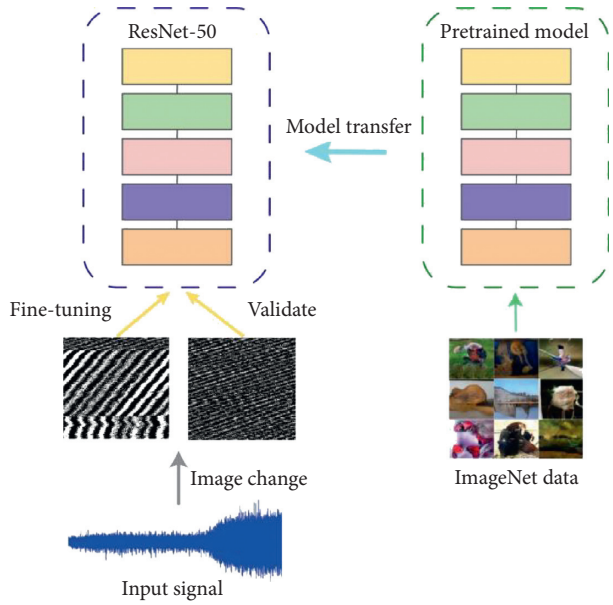

Figure 4: The process of transfer learning.

specific data. By applying the model trained by sufficient training data to the target domain for prediction, the similarity between the models can be fully utilized and model efficiency can be achieved.

2.4. Bidirectional LSTM. RNN is an effective tool to analyze time-series signal. However, it is difficult to reasonably deal with the long and complex time series due to the limitation of its inherent characteristic. LSTM network effectively overcomes the problems existing in the traditional recurrent network through the standard recurrent layer and the internal unique gate structure. However, sensor signal data in machine health monitoring system has strong time dependence, while the basic LSTM can only access the information in specific time step but are unable to build an overall comprehension. As a variant of the traditional LSTM, Bidirectional LSTM (Bi-LSTM) improves model capability in dealing with long sequence, which has stable dynamic learning ability and strong robustness in extracting useful features from complex sequential data. A typical Bi-LSTM model is shown in Figure 5, where Bi-LSTM model adopts bidirectional connection. Each input sequence propagates forward and backward in an independent LSTM, and the output is presented in series. Bidirectional propagation allows each time-series sample to access complete information as it travels in each direction, and the backpropagated LSTM can further smooth the data and reduce the impact of noise.

3. Proposed Method

In this paper, a high-precision RUL prediction framework for rotatory machine based on deep neural network is proposed, which is able to automatically learn fault signatures and identify the degradation process directly from the original vibration signal. The proposed framework is able to achieve quick and accurate RUL prediction.

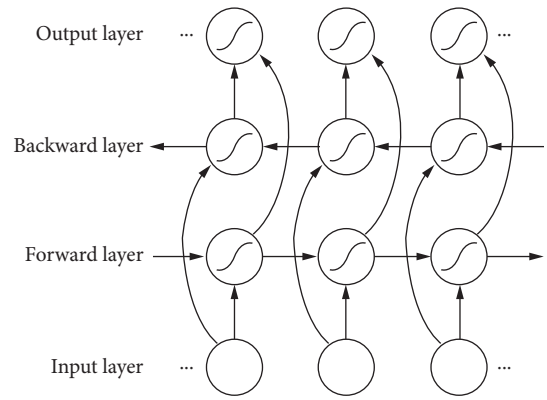

Figure 5: The structure of Bi-LSTM.

The overall pipeline of the prediction framework is shown in Figure 6, which mainly contains 5 stages, including data acquisition, data processing, data partitioning, pretraining feature extractor, and RUL prediction. Algorithm by proposed model for RUL prediction is summarized in Algorithm 1.

$$t = \frac{\left[t - \mathrm{Min}\left(t_{\mathrm{training}}\right)\right]}{\left[\mathrm{Max}\left(t_{\mathrm{training}}\right) - \mathrm{Min}\left(t_{\mathrm{training}}\right)\right]}. \tag{3}$$

Step 1. Data acquisition. The dataset used for RUL prediction is the vibration signals collected by acceleration sensor mounted on rotatory machine. The vibration signal records the whole run-to-fail process of the mechanical system. By analyzing the vibration signals in each time period, the RUL of the rotatory machine is predicted reasonably.

Step 2. Data processing. As the input data format of the pretrained ResNet-50 network is required to be a two-dimensional image, the original one-dimensional signal data need to be processed and converted into two-dimensional images. Different from the time-frequency imaging method adopted by Shao et al. for fault classification tasks, the RUL prediction task is to explore the signal variation characteristic. It is difficult to predict RUL value from the same signal with a small difference in time-frequency change, so the proposed method leverages the amplitude changes of the signals to conduct data processing for the one-dimensional sensor signal. The specific method is designed as follows. Assuming that the original sensor signal is $l = [x_1, x_2, \ldots, x_n]$ and the step length is k, the first sample extracted is $l_1 = [x_1, x_2, \ldots, x_a]$ and the next sample is $l_2 = [x_{1+k}, x_2, \ldots, x_{a+k}]$. Then, the sample is converted into $\sqrt{a} * \sqrt{a}$ and the amplitude of the original vibration signal can be better retained, which is conducive to realize the data analysis and the RUL prediction. Since the converted image is distributed as a grey image of one channel, it is necessary to duplicate the grey image into three channels by means of channel enhancement, so as to realize the two-dimensional image with three channels.

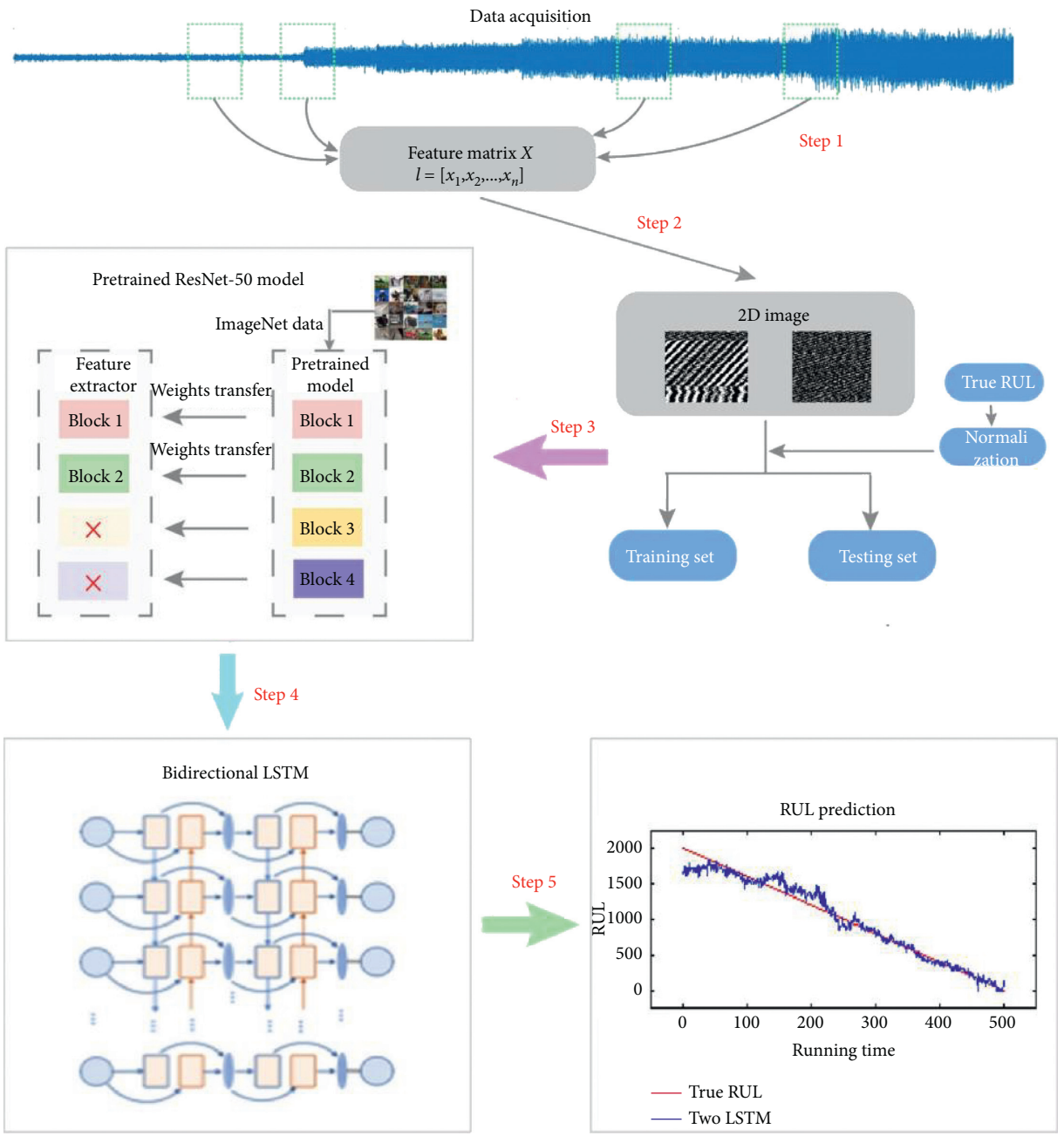

Figure 6: The flowchart of the proposed RUL method.

Step 3. Data partitioning. Each image needs specific RUL labeling. Compared with the amplitude of the vibration signal sequence, the real RUL are relatively large values, where it is difficult for the deep neural network to model the mapping relationship between the input and output. In order to achieve quick convergence of the model and accurate RUL prediction, normalization is performed to the real RUL values, defined as

Through normalization, the output label corresponding to the real RUL values can be obtained. After model training, the proposed framework is able to output predicted labels. To achieve RUL from the predicted values, inverse normalization is utilized as

$$t = t \times \left[\mathrm{Max}\left(t_{\mathrm{training}}\right) - \mathrm{Min}\left(t_{\mathrm{training}}\right)\right] + \mathrm{Min}\left(t_{\mathrm{training}}\right). \tag{4}$$

Through these transformations, effective model training and RUL prediction can be obtained. In addition, the processed images are divided into two parts: the training dataset and the test dataset. The training set is used to fine-tune the pretrained model and update the model parameters, while the testing set is only used to verify the performance of the effectiveness of the proposed architecture.

RUL algorithm for training is defined as follows:
Input: $l = [x_1, x_2, \ldots, x_n]$ # Historical run-to-fail sensors data of rotatory machine
For $i = 1, \ldots, 2500$ do:
$l_i = [x_{1+ij}, x_{2+ij}, \ldots, x_{1024+ij}]$ # j is size of the time window
$l_i \longrightarrow x_i = \begin{bmatrix} x_{1+ij} & \cdots & x_{32+ij} \\ \vdots & \ddots & \vdots \\ x_{992+ij} & \cdots & x_{1024+ij} \end{bmatrix}$ # Convert time-series sample into grey image
$x_i = \text{copy}(x_i).\text{reshape}(32, 32, 3)$ # Duplicate and extend channel into 3
$X = [\phi]$
$X = X.\text{append}(x_i)$ # The whole datasets
end
$t = [t_1, t_2, \ldots, t_i]$ #Define corresponding RUL value
Function TRAIN (X, N, d):
$t = [t - \text{Min}(t_{\text{training}})]/[\text{Max}(t_{\text{training}}) - \text{Min}(t_{\text{training}})]$ # Normalization t
N # N means Maximum number of iterative training d # represents minimum allowable training error
Res-Conv2 # rain the ResNet-50 model by ImageNet datasets and use the first two blocks of ResNet-50 as pretrained feature extractor.
w, b # Initialize parameters of Bi-LSTM module and fully connected layers
$Z = \text{Random. Sample(X,2000)}$ # randomly select 2000 samples as training sets Z
While $N > n$ or $L < d$:
For $j = 1, \ldots, 2000$ do: $c = \text{Res-Conv2}(Z)$
$h = \text{Bi-LSTM}\ (c) \longrightarrow h$ is the vector extracted by Bi-LSTM
$\text{RUL} = \text{MLP}(h) \longrightarrow$ The RUL value is output from fully connected layer.
$L = \sqrt{(1/n)\sum_{i=1}^{n} (t^i - \text{RUL}^i)^2}$
$w \longleftarrow \text{Adam}\,(\nabla_w \sum_{r=1}^{n} L, w)$
$b \longleftarrow \text{Adam}\,(\nabla_b \sum_{r=1}^{n} L, b)$
Output: the trained proposed model.

Algorithm 1: The RUL algorithm for training.

Step 4. Pretraining feature extractor. The pretrained model used in the proposed method is the first two convolution blocks of ResNet-50 which has obtained accurate classification results on the ImageNet data, which indicates that ResNet-50 has effective and efficient feature learning ability. Detailed information about the selected convolution blocks is shown in Algorithm 1. Different from task-specific features, general features are extracted by the first two blocks of the pretrained model and suitable for various tasks.

Step 5. RUL prediction. As the ImageNet dataset differs greatly from the sensor vibration signal dataset, the issue of RUL prediction is much more difficult than the simple image classification prediction. Thus, the features extracted by the pretrained network will be used as the input of the Bidirectional LSTM and the sequential network is designed to realize the final RUL prediction. The whole network needs to be fine-tuned for the specific rotatory machine dataset during the training process. The fully connected layer is located at the end of the network to output RUL results. In this paper, as shown in formulas (5) and (6), ReLU function is used as the activation function, and the last layer outputs the specific value of RUL:

$$y = f(Wa + b), \tag{5}$$

$$f(n) = \max\{0, n\}. \tag{6}$$

4. Experiment and Analysis

To explore the performances of model and verify the effectiveness of its prediction, several experiments were carried out among the bearing dataset and gearbox dataset, respectively. The bearing dataset contains vibration data of various bearings collected at Xi'an Jiaotong University (XJTU) and the Changxing Sumyoung Technology (SY). The gear dataset was carried out by Chongqing University on the contact fatigue testing machine. These two different datasets contain the run-to-fail process of bearing and gear operating under different conditions. By comparing with the existing mainstream forecasting methods, the validity of the proposed method is verified.

4.1. Case 1: XJTU Bearing Dataset

4.1.1. Data Description. As shown in Figure 7, the experimental data were collected by the bearing testbed. The testbed is designed for the degradation test of rolling bearings under different working conditions. The dataset contains run-to-fail sensor data of multiple rolling bearings, and the whole dataset was obtained through a number of accelerated degradation experiments. Table 2 shows the detailed description of the experimental data. Four different fault types are set, including inner race wear, cage fracture, outer race fracture, and outer race wear.

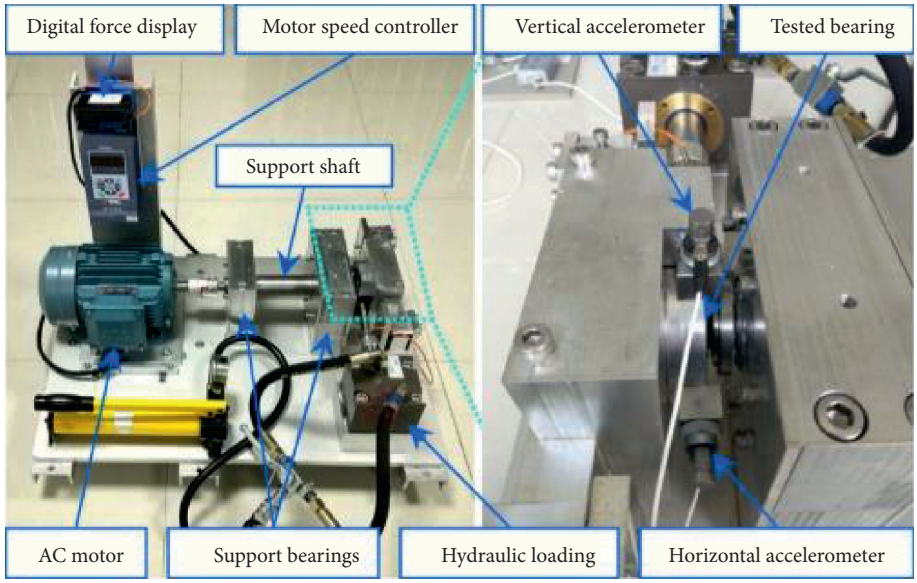

Figure 7: Bearing contact fatigue testing machine.

Table 2: Details of bearing datasets [37].

Operating condition	Radial force (kN)	Rotating speed (rpm)	
Condition 1	12	2100	Bearings 1–1 and 1–2 Bearings 1–3 and 1–4 Bearing 1–5
Condition 2	11	2250	Bearings 2–1 and 2–2 Bearings 2–3 and 2–4 Bearing 2–5
Condition 3	10	2400	Bearings 3–1 and 3–2 Bearings 3–3 and 3–4 Bearing 3–5

4.1.2. Building and Training. For the data preparation, the time window is set to be 1024, which means that one single sample has 1024 data points. The selected sensor signal within the time window is processed to be an amplitude image with size of 32×32 which is suitable for the pretrained network. The format of the processed images is $32 \times 32 \times 3$ by duplication of the origin grey-scale images. After that, preprocessed image samples are separated into two parts for training and testing, respectively. The size of the training samples is 2000 for each condition while the size of the testing samples is 500. Adam optimizer is adopted for parameters updating with learning rate 0.001 and the batch size was set to 16.

For model comparison, the following methods are also conducted using the same dataset:

RNN: basic RNN

LSTM: a one-layer LSTM network

TLSTM: a two-layer LSTM network

GRU: a Gated Recurrent Unit-based network

Bi-LSTM: a Bidirectional LSTM network

ResSN: a Sequential Network with ResNet trained from scratch

ResSN-TL: a Sequential Network with ResNet using Transfer Learning

4.1.3. RUL Prediction for Bearing Dataset. To verify the validity of the proposed architecture, we took two different fault datasets under three conditions as specific experimental datasets, and RMSE loss is adopted to evaluate the RUL prediction performance. Figure 8 shows the comparison of training efficiency between the proposed model and the model training from scratch, and the time needed for the two models to reach the specific loss value is compared. In these experiments, we set RMSE threshold as 65 to record the time required for model training.

From the results shown in Figure 8, the training time of the proposed ResSN-TL model is much lower than that of the network trained from scratch, which verifies the advantages of transfer learning strategy in training efficiency.

Table 3 shows the final RMSE loss of each model. Compared with the current mainstream prediction model RNN and its variants, the proposed model converts one-dimensional sensor signals into two-dimensional images, effectively leveraging the advantages of convolutional neural network in feature learning, and is superior to most sequential networks in terms of RUL prediction. By means of transfer learning, the feature extraction module of the network is pretrained, which greatly improves the training efficiency of the model. Compared with the model training from scratch, the effectiveness of the transfer learning method is verified in terms of both training time and prediction accuracy.

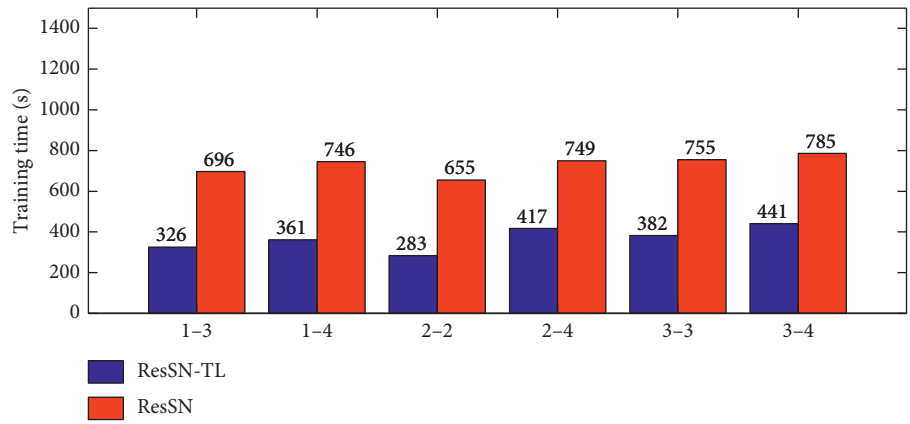

Figure 8: Training time of ResSN and ResSN-TL.

Table 3: RMSE for different methods on RUL prediction.

Method/dataset	1–3	1–4	2–2	2–4	3–3	3–4
RNN	108.56	113.76	105.32	135.96	139.37	145.73
LSTM	65.13	70.34	51.08	64.12	69.14	85.65
Two LSTM	59.97	65.76	50.92	59.38	65.17	83.12
GRU	67.32	73.33	62.17	69.31	75.38	90.13
Bi-LSTM	51.03	61.25	43.41	58.14	60.79	68.88
ResSN	40.11	49.73	34.60	50.33	53.74	61.39
ResSN-TL	**35.97**	**38.86**	**26.06**	**42.41**	**46.45**	**48.98**

To provide convincing experimental verification and discuss the model learning ability, we conduct 100 times hand-out cross validation using the proposed framework and several comparison methods and generate RMSE confidence interval with 95% confidence level to evaluate the model learning ability for RUL probability distribution, shown in Table 4 and Figure 9.

In Figure 9, the color bars represent the average RMSE across the 100-time cross-validation experiments and the black lines denote their confidence interval. Based on the estimation results of confidence interval for RUL prediction, the proposed framework ResSN-TL has relatively superior performance. For the proposed model, there is a 95% likelihood that the RMSE of the predicted RUL ranges between 39.64 and 55.12. Compared with other models, the proposed framework has more accurate prediction results and the results have verified the superiority in learning ability of the proposed method.

Figure 10 shows the RUL prediction results of bearing dataset 2–4 by the proposed model. It can be seen from the results that the RNN model has a slightly weaker performance in analysis and prediction of complex signals with long sequence than its variants. Although LSTM and GRU model are able to solve the problems existing in RNN such as gradient explosion and gradient disappearance to some degree, the prediction performance of them is still unsatisfactory. It can be explained that the signal beyond a certain length may result in information loss during the long-distance transmission process. Compared with LSTM and GRU networks, the prediction performance of the Bi-LSTM model is significantly improved, which verifies that it is effective to analyze the long period of time-series signals through bidirectional propagation. Convolutional neural network has its advantages in the field of feature extraction compared with time-series model. Combined with the powerful feature extraction capability of convolutional neural network and the analytical capability of Bidirectional LSTM to sequence signals, the model proposed in this paper is superior to most time-series models in prediction effect, which proves the effectiveness and efficiency of the proposed method.

4.1.4. Model Generalization Ability. In order to further investigate the model performance for a more general task, several additional experiments have been conducted. Datasets under different working conditions for training and testing are utilized to investigate the model generalizability. Table 5 shows various datasets defined for training procedure and testing prediction verification.

Experiments have been carried out where model parameters and hyperparameters are initialized the same as the original model initialization. Training datasets are utilized to train the proposed model and the trained model is saved. After sufficient training, the proposed model has obtained the key feature learning ability and RUL prediction ability. Testing datasets under different working conditions are applied to test the model performances in RUL prediction. Specifically, the mixture of sensor data from bearings 1–1, 1–2, 1–3, 1–4, and 1–5 is used as training datasets to well train the proposed model, and the sensor data from bearings 2–1, 2–2, 2–3, 2–4, and 2–5 are used as the testing data separately to test the trained model.

The RUL prediction results under different bearing datasets are shown in Figure 11 and Table 6. Several comparisons are also carried out and the experimental results are shown in Figure 12.

From the results, our model is able to achieve approximate RUL prediction among various working conditions and outperforms other frameworks. However, compared with the experimental results based on the same unit, model prediction accuracy using testing data from different working conditions is lower, which owes to the various key

Table 4: RMSE results of hand-out cross validation.

Method	Dataset					
	1–3	1–4	2–2	2–4	3–3	3–4
RNN	112.14 (± 19.64)	109.32 (± 15.22)	112.58 (± 16.26)	126.63 (± 19.48)	135.47 (± 24.71)	132.19 (± 20.25)
GRU	65.28 (± 9.99)	78.19 (± 8.47)	68.51 (± 11.21)	65.11 (± 9.77)	71.69 (± 11.76)	88.45 (± 9.49)
Bi-LSTM	55.79 (± 8.87)	59.63 (± 8.62)	48.51 (± 9.94)	54.36 (± 9.92)	57.78 (± 10.23)	67.98 (± 17.53)
ResSN-TL	37.26 (± 7.63)	38.12 (± 8.48)	30.25 (± 7.12)	38.97 (± 6.25)	42.36 (± 7.97)	47.38 (± 7.74)

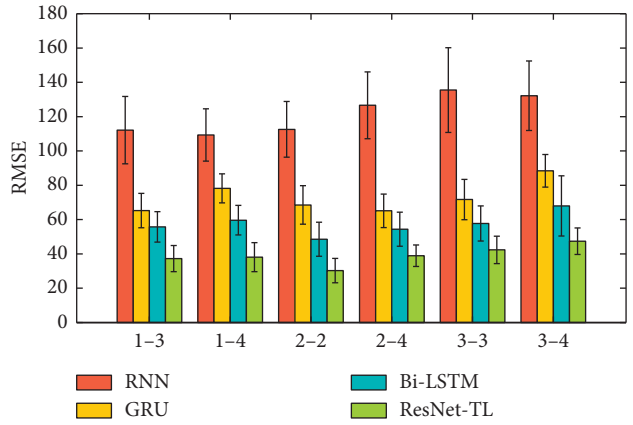

Figure 9: Confidence interval of RMSE for different methods.

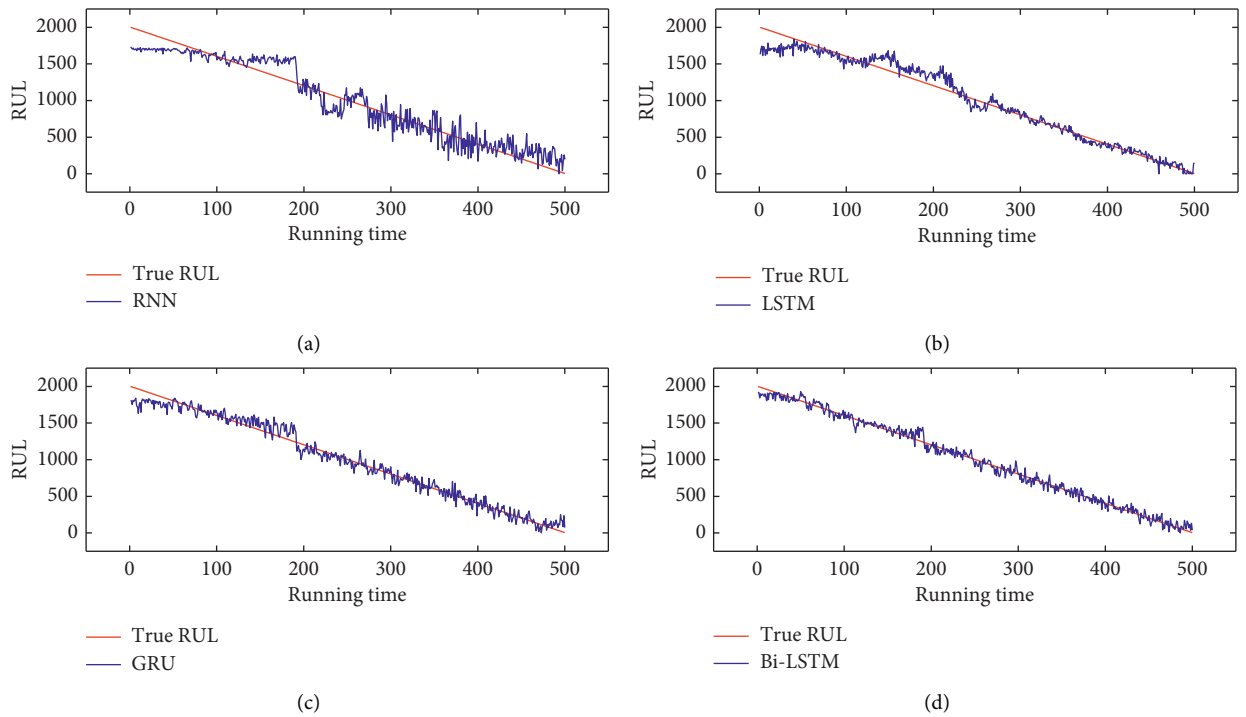

Figure 10: Continued.

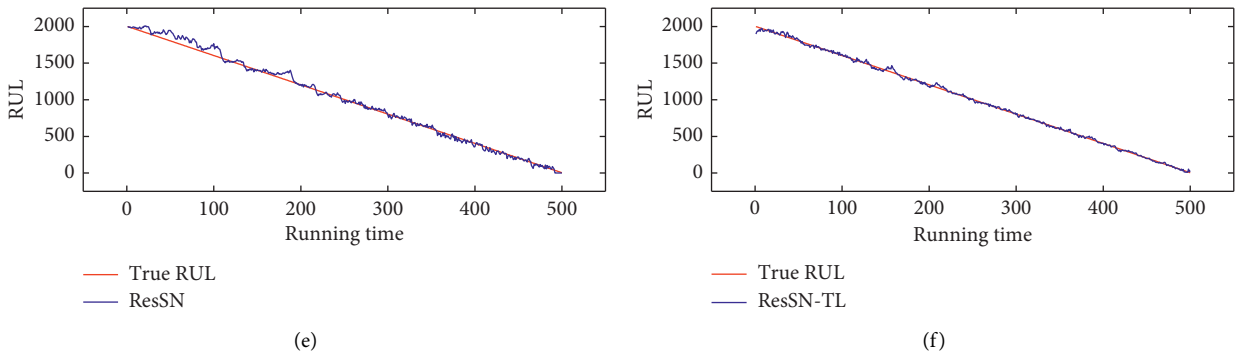

Figure 10: RUL prediction results of bearing dataset 2–4. (a) RNN, (b) LSTM, (c) GRU, (d) Bi-LSTM, (e) ResSN, and (f) ResSN-TL.

Table 5: Experimental data.

Dataset	Radial force (kN)	Rotating speed (rpm)	
Training datasets	12	2100	Bearings 1–1 and 1–2 Bearings 1–3 and 1–4 Bearing 1–5
Testing datasets	11	2250	Bearings 2–1 and 2–2 Bearings 2–3 and 2–4 Bearing 2–5

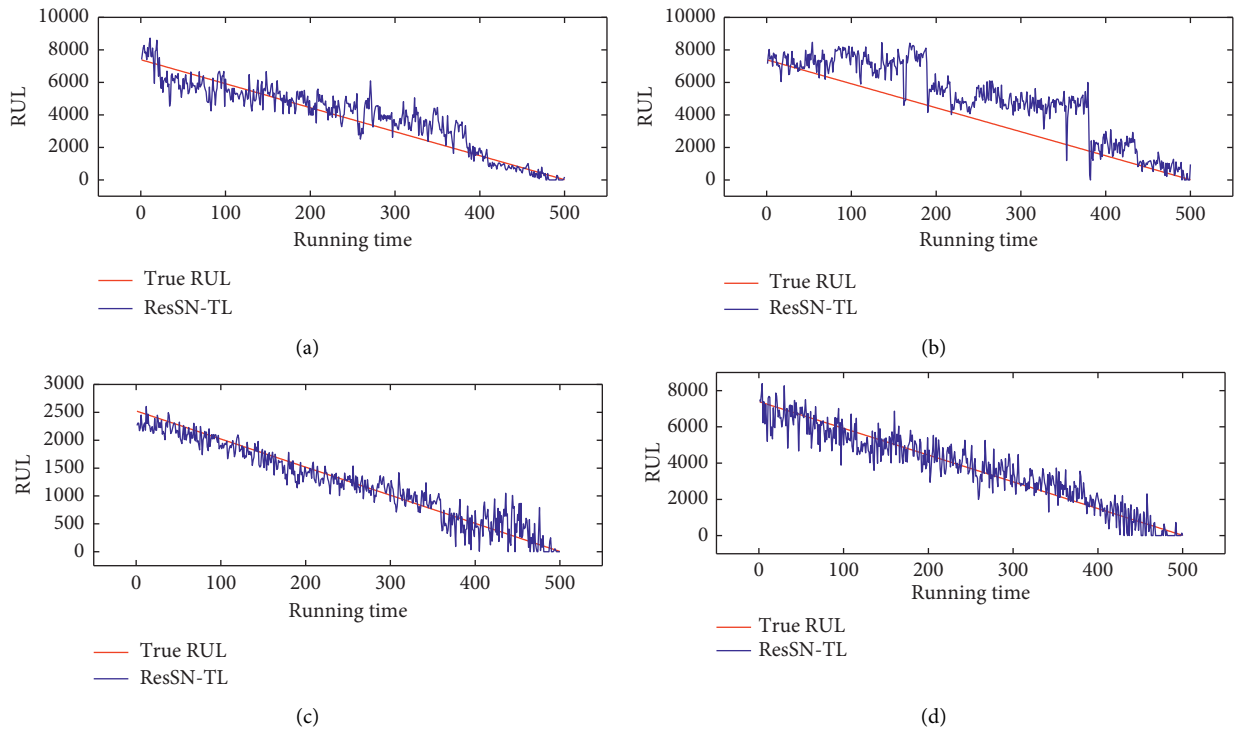

Figure 11: RUL prediction results of experimental data. (a) Bearing 2-1, (b) bearing 2–3, (c) bearing 2–4, and (d) bearing 2–5 experimental data.

Table 6: RMSE results of Ea.

Method/dataset	2–1	2–2	2–3	2–4	2–5
RNN	458.96	378.21	497.15	421.39	474.45
GRU	384.29	346.94	428.68	370.85	421.98
Bi-LSTM	361.97	285.99	397.62	366.12	398.64
ResSN-TL	**252.61**	**147.39**	**325.36**	**201.74**	**223.58**

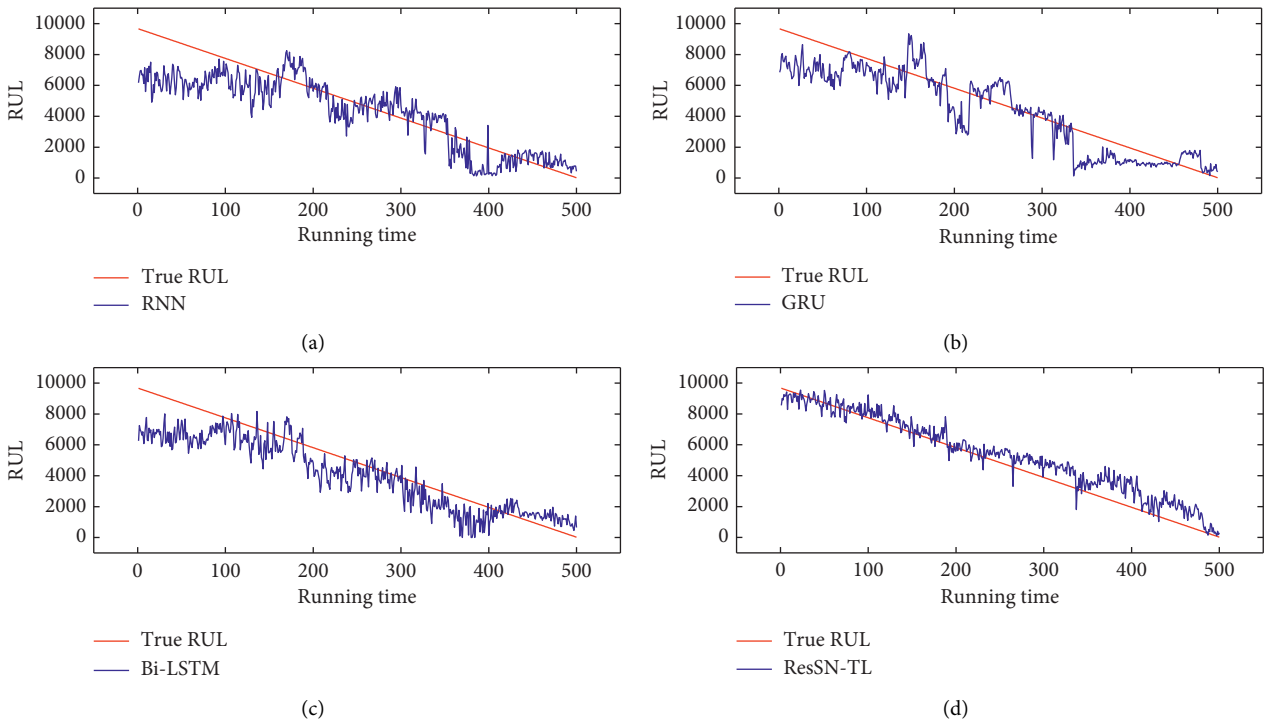

Figure 12: RUL prediction results of bearing dataset 2–2, (a) RNN, (b) GRU, (c) Bi-LSTM, and (d) ResSN-TL.

features among various working conditions and different failure types. The complexity of degradation features increases the difficulty of model generalization.

4.2. Case 2: Gearbox Dataset

4.2.1. Data Description. In this case study, the gearbox sensor signal dataset is used to verify the performance of the proposed model. As shown in Figure 13, vibration signals are collected by sensors placed on a gearbox and the sampling frequency was set to 50 kHz. The experiment is set to stop when the amplitude of the collected vibration signals exceeds the given threshold. The specific description of the experimental data is shown in Table 7, including a total of four run-to-failure datasets under two different working conditions.

4.2.2. Building and Training. Compared with bearing signal dataset from XJTU, the vibration signals of gearbox have no obvious trend of amplitude change, and the sampling frequency is also different from the bearing dataset. To effectively analyze the dataset and evaluate the prediction performance of each model, a time window containing 10000 data points is chosen to be one sample and every sample is converted to an image with the size of 100×100. The other settings are the same as those mentioned in the previous section.

4.2.3. RUL Prediction for Bearing Dataset. Figure 14 shows the comparisons in training time of the proposed pipeline with the ResSN model trained from scratch. Similar to the experimental conclusion in case study 1, the transfer learning-based method is able to accelerate the training procedure and therefore improve model training efficiency.

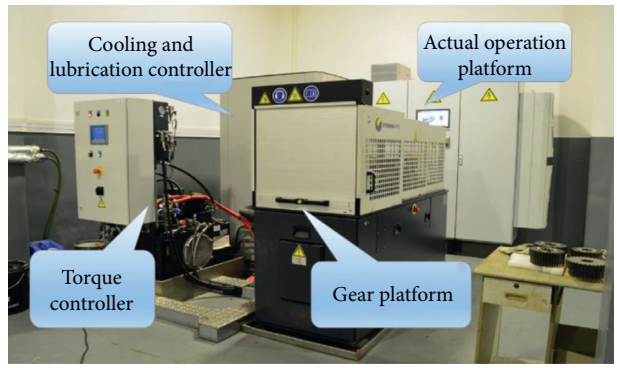

Figure 13: Gearbox testing machine.

Table 7: Details of gearbox datasets [26].

	Data A	Data B	Data C	Data D
Load (N)	1400	1400	1300	1300
Speed (rpm)	500	520	1000	1000
Experimental time (min)	814	820	789	796
Number of sample points	1221	1230	1183	1193

Table 8 shows the RMSE loss of four different datasets. It can be seen that RNN has relatively poor predictive performance due to its insufficient processing capacity for long and complex sequence signals. The prediction performance of LSTM and GRU is better than that of RNN. By stacking LSTM or using bidirectional units, the predictive ability of the network significantly improved.

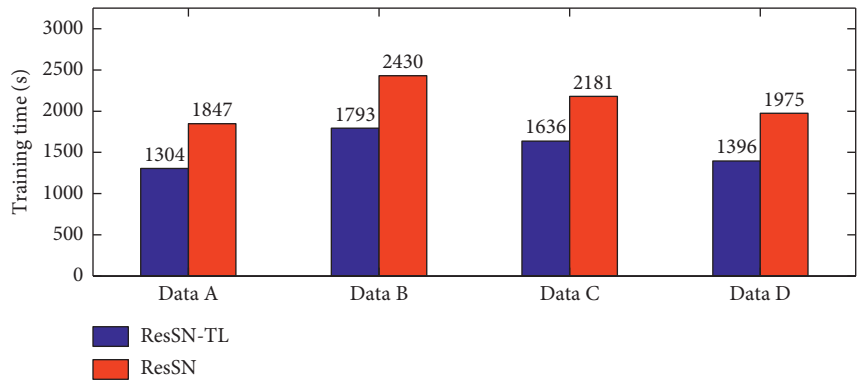

Figure 14: Training time of ResSN and ResSN-TL.

Table 8: Gearbox dataset RMSE for different methods on RUL prediction.

Method/dataset	Data A	Data B	Data C	Data D
RNN	213.01	268.50	235.65	220.14
LSTM	102.73	140.44	121.38	114.36
Two LSTM	85.99	115.36	103.39	96.11
GRU	90.23	131.85	115.93	109.98
Bi-LSTM	73.18	87.39	86.17	83.56
ResSN	58.13	79.33	75.36	68.36
ResSN-TL	**53.80**	**75.12**	**69.32**	**57.98**

Table 9: RMSE results of hand-out cross validation.

Method	Dataset			
	Data A	Data B	Data C	Data D
RNN	210.79 (± 25.34)	255.91 (± 21.29)	241.85 (± 30.15)	225.16 (± 26.13)
GRU	92.47 (± 12.36)	125.32 (± 18.24)	112.74 (± 14.96)	108.53 (± 13.85)
Bi-LSTM	76.15 (± 10.98)	88.62 (± 11.34)	82.49 (± 9.54)	85.17 (± 12.79)
ResSN-TL	54.69 (± 7.85)	71.48 (± 7.16)	67.13 (± 6.35)	60.88 (± 11.47)

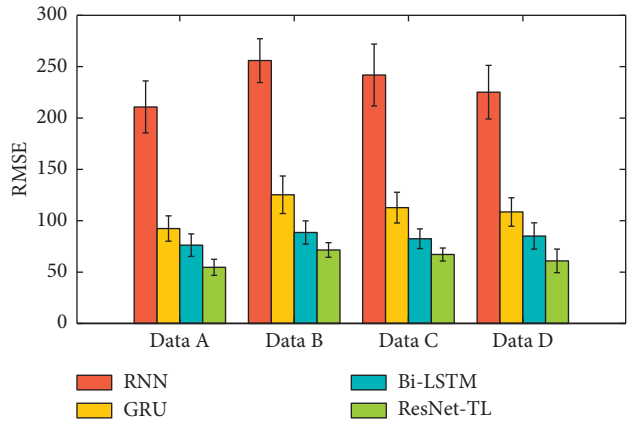

Figure 15: Confidence interval of RMSE for different methods.

Table 9 and Figure 15 show the color bars about the confidence interval of the RMSE to evaluate the model prediction ability of RUL. Results have shown that there is a 95% likelihood that the RMSE of the predicted RUL ranges between 64.32 and 78.64 using the proposed framework which is smaller than other methods, ranging between 65.17 and 87.13, which has verified the model learning ability from historical sensor data.

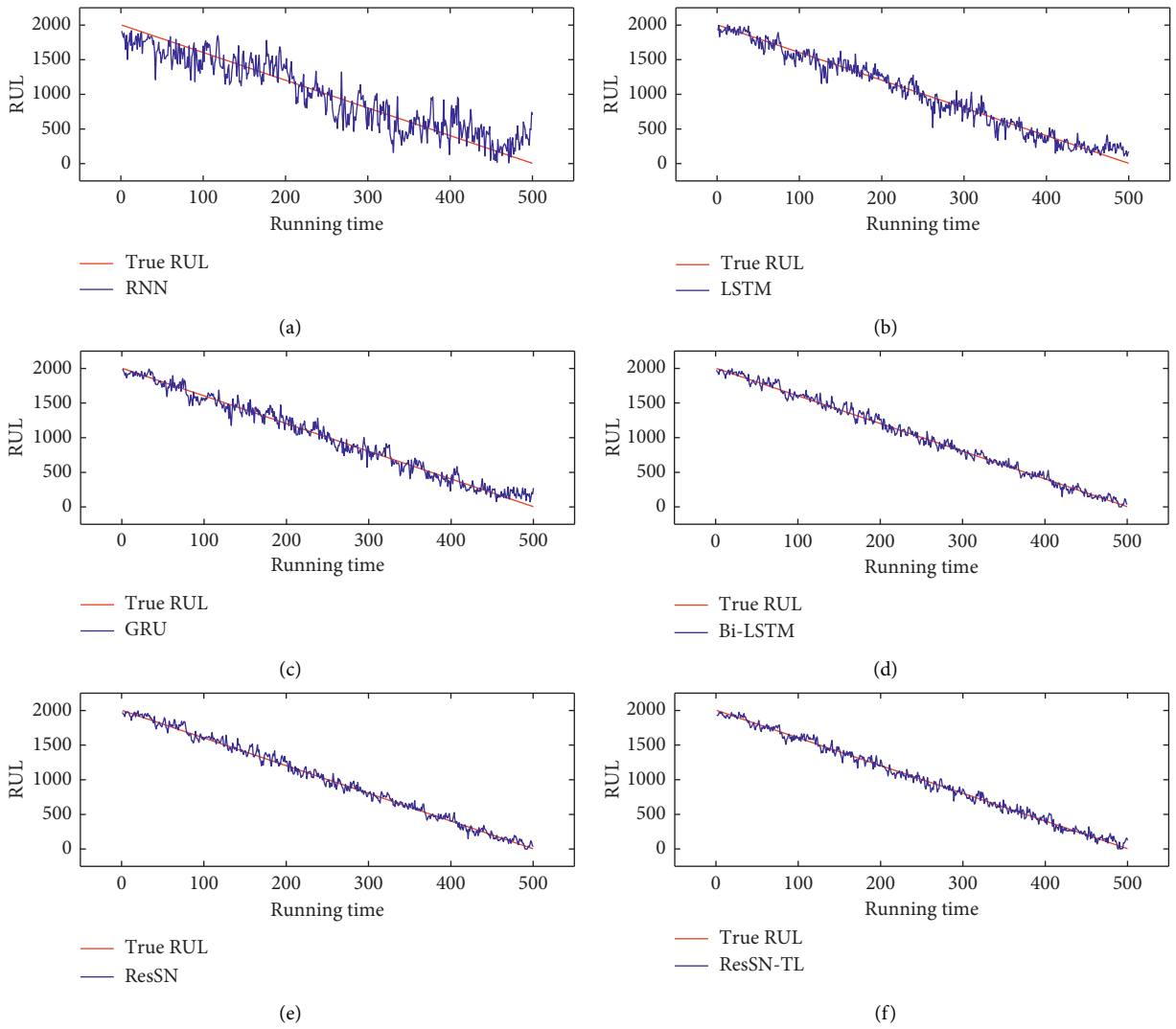

Figure 16: RUL prediction results of gearbox Data A, (a) RNN, (b) LSTM, (c) GRU, (d) Bi-LSTM, (e) ResSN, and (f) ResSN-TL.

Figure 16 shows the RUL prediction results of Data A. The conclusion is consistent with the one driven from bearing dataset, where the prediction performance of the proposed model combined pretrained feature extractor with the Bidirectional LSTM is the best, which further verifies the effectiveness of the method proposed in this paper.

5. Conclusion

In conclusion, an RUL prediction model framework based on transfer learning is proposed in this paper. By means of transfer learning, certain model parameters are initialized reasonably, which solves the problem of training instability existing in random initialization and greatly reduces the training burden of the deep architecture. Combining the pretrained feature extractor using residual blocks with the sequential model using Bi-LSTM architecture, an efficient and accurate RUL prediction model for rotatory machine is established, which is advanced in training efficiency and prediction accuracy.

The advantages of the proposed model are proved among the datasets of bearing and gear run-to-fail sensor signals. In the future, transfer learning can be expected to play a useful role in fault detection and RUL prediction across various mechanical systems. Besides, transfer guidelines will be further investigated.

Data Availability

The data used in this paper are downloaded from XJTU-SY Bearing Datasets at https://biaowang.tech/xjtu-sy-bearing-datasets/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61372033.

References

[1] T. Wang, Q. Han, F. Chu, and Z. Feng, "Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: a review," *Mechanical Systems and Signal Processing*, vol. 126, pp. 662–685, 2019.

[2] L. Xie, J. Si, Y. Hu, and H. Feng, "Helical motion analysis of the 2-Degree-of-Freedom split-stator induction motor," *IEEE Transactions on Magnetics*, vol. 55, no. 6, pp. 1–5, 2019.

[3] X. Jiang, C. Shen, J. Shi, and Z. Zhu, "Initial center frequency-guided VMD for fault diagnosis of rotating machines," *Journal of Sound and Vibration*, vol. 435, pp. 36–55, 2018.

[4] Y. Diao, X. Men, Z. Sun et al., "Structural damage identification based on the transmissibility function and support vector machine," *Shock and Vibration*, vol. 2018, Article ID 4892428, 13 pages, 2018.

[5] C. Li, R.-V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, "Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals," *Mechanical Systems and Signal Processing*, vol. 76-77, pp. 283–293, 2016.

[6] J. Carroll, S. Koukoura, A. Mcdonald, A. Charalambous, S. Weiss, and S. McArthur, "Wind turbine gearbox failure and remaining useful life prediction using machine learning techniques," *Wind Energy*, vol. 22, no. 3, pp. 360–375, 2019.

[7] P. Kundu, S. Chopra, and B. K. Lad, "Multiple failure behaviors identification and remaining useful life prediction of ball bearings," *Journal of Intelligent Manufacturing*, vol. 30, no. 4, pp. 1795–1807, 2019.

[8] K. Javed, R. Gouriveau, and N. Zerhouni, "State of the art and taxonomy of prognostics approaches, trends of prognostics applications and open issues towards maturity at different technology readiness levels," *Mechanical Systems and Signal Processing*, vol. 94, pp. 214–236, 2017.

[9] Y. Lei, N. Li, S. Gontarz, J. Lin, S. Radkowski, and J. Dybala, "A model-based method for remaining useful life prediction of machinery," *IEEE Transactions on Reliability*, vol. 65, no. 3, pp. 1314–1326, 2016.

[10] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques-Part I: fault Diagnosis with model-based and signal-based approaches," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3757–3767, 2015.

[11] S.-Y. Shao, W.-J. Sun, R.-Q. Yan, P. Wang, and R. X. Gao, "A deep learning approach for fault diagnosis of induction motors in manufacturing," *Chinese Journal of Mechanical Engineering*, vol. 30, no. 6, pp. 1347–1356, 2017.

[12] J. M. Kim and M. Sohaib, "Reliable fault diagnosis of rotary machine bearings using a stacked sparse autoencoder-based deep neural network," *Shock and Vibration*, vol. 2018, Article ID 2919637, 11 pages, 2018.

[13] R. Zhao, Z. Chen, R. Yan et al., "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019.

[14] F. Sun, N. Wang, X. Li, and W. Zhang, "Remaining useful life prediction for a machine with multiple dependent features based on bayesian dynamic linear model and copulas," *IEEE Access*, vol. 5, pp. 16277–16287, 2017.

[15] 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 classificationfication," *Measurement*, vol. 89, pp. 171–178, 2016.

[16] A. R. Bastami, A. Aasi, and H. A. Arghand, "Estimation of remaining useful life of rolling element bearings using wavelet packet decomposition and artificial neural network," *Iranian Journal of Science & Technology Transactions of Electrical Engineering*, vol. 43, pp. 233–245, 2019.

[17] C. Shen, Y. Qi, J. Wang et al., "An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder," *Engineering Applications of Artificial Intelligence*, vol. 76, pp. 170–184, 2018.

[18] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Reliability Engineering & System Safety*, vol. 172, pp. 1–11, 2018.

[19] H. M. Ertunc, H. Ocak, and C. Aliustaoglu, "ANN- and ANFIS based multi-staged decision algorithm for the detection and diagnosis of bearing faults," *Neural Comput Appl*, vol. 22, no. 1, pp. 435–446, 2013.

[20] S. Nie, M. Zheng, and Q. Ji, "The deep regression bayesian network and its applications: probabilistic deep learning for computer vision," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 101–111, 2018.

[21] Y. Tom, H. Devamanyu, P. Soujanya et al., "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[22] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[23] D. Monroe, "Deep learning takes on translation," *Communications of the Acm*, vol. 60, no. 6, pp. 12–14, 2017.

[24] D. Zhang, W. Ding, B. Zhang et al., "Automatic modulation classification based on deep learning for unmanned aerial vehicles," *Sensors*, vol. 18, no. 3, pp. 924–939, 2018.

[25] J. Deutsch and D. He, "Using deep learning-based approach to predict remaining useful life of rotating components," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 1, pp. 11–20, 2017.

[26] Yi Qin, S. Xiang, Yi Chai et al., "Macroscopic-microscopic attention in LSTM networks based on fusion features for gear remaining life prediction," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 12, pp. 10865–10875, 2020.

[27] X. Ding and Q. He, "Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosisfluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 8, pp. 1926–1935, 2017.

[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3320–3328, Montreal, Canada, December 2014.

[29] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse autoencoder for fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 136–144, 2017.

[30] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, "Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2416–2425, 2019.

[31] S. Shao, S. McAleer, R. Yan et al., "Highly-accurate machine fault diagnosis using deep transfer learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2446–2455, 2018.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[33] M. Yuan, Y. Wu, and L. Lin, "Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network," in *Proceedings of the IEEE/CSAA International Conference on Aircraft Utility Systems*, pp. 135–140, Beijing, China, October 2016.
[34] J. Wang, B. Peng, and X. Zhang, "Using a stacked residual LSTM model for sentiment intensity prediction," *Neurocomputing*, vol. 322, pp. 93–101, 2018.
[35] J. Zhang, P. Wang, R. Yan, and R. X. Gao, "Long short-term memory for machine remaining life prediction," *Journal of Manufacturing Systems*, vol. 48, pp. 78–86, 2018.
[36] K. He, X. Zhang, S. Ren et al., "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, pp. 630–645, Amsterdam, Netherlands, October 2016.
[37] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Transactions on Reliability*, vol. 69, no. 1, pp. 401–412, 2020.