

Received 1 February 2024, accepted 25 April 2024, date of publication 2 May 2024, date of current version 17 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396156

## RESEARCH ARTICLE

# An Improved Method for Predicting the Remaining Useful Life Using a Spatial-Temporal Feature Extraction Network With Attention Mechanism

XIAOJIA YAN, WEIGE LIANG, AND SHIYAN SUN

College of Weapon Engineering, Naval University of Engineering, Wuhan 430033, China

Corresponding author: Weige Liang (1312021010@nue.edu.cn)


This work was supported in part by the National Natural Science Foundation of China under Grant 61640308, and in part by the Natural Science Foundation of Hubei Province under Grant 2023AFB900.

**ABSTRACT** The prediction of the remaining useful life (RUL) of mechanical equipment is of vital importance to its operation and maintenance. Deep learning methods can effectively extract degradation information closely related to equipment RUL from extensive monitoring data. However, when the data is nonlinear, multi-dimensional, long-term, and large-scale, the critical degradation information for RUL prediction may be obscured. Traditional deep learning networks do not perform well in predicting RUL. Therefore, a method for predicting the remaining useful life of mechanical equipment using a spatial-temporal feature extraction network is proposed. The main innovations can be classified into two aspects. Firstly, by training a unidirectional residual convolutional network (URCNN), the deep spatial features of the monitoring data are extracted. This network does not disrupt the temporal relevance of the monitoring data and can effectively avoid the phenomenon of vanishing gradient during the training process. Secondly, the weight parameters of the bidirectional long short-term memory network (BiLSTM) for extracting time-related features are optimized by introducing an attention mechanism. The attention mechanism can effectively enhance the expression of crucial degradation information for RUL prediction. Eventually, the benchmark dataset and the specialized transmission mechanism dataset validates the effectiveness and superiority of the proposed method. The analysis results indicate that for multi-dimensional monitoring data with complex operating conditions and variable fault modes, the proposed method can accurately locate degradation temporal points and effectively improve the RUL prediction accuracy of long-term operating equipment.

**INDEX TERMS** Remaining useful life, spatial-temporal feature, attention mechanism, unidirectional residual convolutional network.

## I. INTRODUCTION

Prognostics and health management (PHM) of mechanical equipment can enhance its safety and prevent catastrophic accidents [1], [2]. Accurate prediction of the remaining useful life (RUL) is a key task in the field of PHM research, providing decision-making support for establishing optimal maintenance strategies for equipment [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang .

RUL prediction can dynamically perceive future changes in the health status of mechanical equipment [4]. Currently, there are three popular methods: model-driven prediction methods, data-driven prediction methods, and hybrid-driven prediction methods [5]. Model-driven and hybrid-driven prediction methods require explicit mathematical models based on known mechanical principles to analyze the degradation process of equipment [6]. However, the structure of large mechanical equipment is complex, and the operating environment and failure modes are diverse, making the establishment

of accurate degradation models costly and significantly limiting their application scope [7]. In contrast, data-driven prediction methods only involve mining performance features from monitoring data during equipment operation, thus revealing the degradation of equipment performance caused by internal and external environments [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. This method avoids the mathematical modeling of complex mechanisms of mechanical equipment, gradually becoming a hot research topic in RUL prediction.

Data-driven prediction methods mainly fall into two categories: statistical analysis and machine learning [8]. Statistical analysis methods predict RUL by constructing a distribution function of the remaining life through parameter estimation based on equipment failure time data, and then realize RUL prediction by fitting the relationship of random variables [9]. Li et al. [10] obtained the remaining life of aviation engines and cutting tools in the form of probability density functions based on the characteristics of Wiener processes. Zhao et al. [11] updated the drift coefficient of the Wiener process in real-time based on similarity measurement, effectively solving nonlinearity and multi-stage problems. Wu et al. [12] proposed an adaptive nonlinear Wiener process model with degradation drift to simulate non-normal and asymmetric degradation characteristics. Feng et al. [13] proposed a semi-random filtering-expectation maximization algorithm, successfully solving the problem of online prediction of remaining life in hidden degradation processes. Kundu et al. [14] considered the working parameters and monitoring signals of the equipment during the model parameter estimation process, thereby establishing a Weibull accelerated damage regression model to describe the failure characteristics of the equipment under different working conditions. Zhang et al. [15] established an RUL prediction model based on the Gamma degradation process using integrated monitoring information of equipment. Wang et al. [16] introduced the ratio of the diffusion coefficient to the drift coefficient in the traditional Wiener degradation model, thereby achieving synchronous updating of the drift coefficient and the diffusion coefficient. Tang et al. [17] used Dempster-Shafer theory of evidence to deal with prediction uncertainty with good convergence. However, with the development of intelligent sensors, multiple networked sensors are usually used to collect multi-source data to fully reflect the operating status of the equipment. Statistical analysis methods struggle to extract degradation feature information from extensive data with large signal differences, multiple sampling strategies, and low data value, while the problem of solving the remaining life distribution based on coupled multi-dimensional variables remains unresolved. Machine learning, represented by deep learning, can extract features and train from raw monitoring data, simulating the performance degradation process without the need to construct specific degradation models. Meanwhile, deep learning methods overcome the difficulty of shallow machine

learning methods in extracting deep features [18]. By extracting deep features from multi-dimensional data composed of performance parameters of multiple sensors, they can fully learn the information contained in the time series for more accurate RUL prediction.

Deep Recurrent Neural Networks (RNN) are suitable for handling time series, mining latent features that reflect equipment performance degradation. Its derivative model, the Long Short Term Memory (LSTM) network, can solve problems such as RNN models being unable to learn long-term feature relationships and vanishing gradient problems, demonstrating significant advantages in solving the RUL prediction problem for equipment [19]. Shuai et al. [20] improved RUL prediction accuracy by stacking multiple layers of LSTM to construct a deep LSTM structure. Zhao et al. [21] combined a generalized learning system algorithm and LSTM to develop a fusion neural network model that excellently predicts the capacity and RUL of lithium-ion batteries. Xiang et al. [22] constructed a multi-scale LSTM capable of mining information on different degradation states, demonstrating better prediction performance. Song et al. [23] proposed a fusion model of Autoencoder and Bidirectional Long Short Term Memory Network (BiLSTM), achieving good results in handling high-dimensional massive state monitoring data. Chen et al. [24] used Convolutional Neural Networks (CNN) to extract spatial features of monitoring data, thereby accurately predicting the remaining life of the equipment.

The above research indicates that LSTM performs well when processing time-series data. However, when dealing with long time series generated by long-term operation of equipment, it struggles to capture key degradation information, leading to a low recognition accuracy rate of performance degradation time points and a tendency to lag in RUL prediction. To solve this problem, this paper proposes a method for predicting the RUL of mechanical equipment based on a fusion of Unidirectional Residual Convolutional Neural Network and Attentional Bidirectional Long Short-Term Memory Network (URCNN-ABiLSTM). The model used in this method is called the Spatial-temporal Feature Extraction Network. The main advantages of this model are as follows:

- 1) The 1D-CNN networks commonly used to process sequence data use a 1-dimensional convolutional kernel to extract features from the data along the time dimension. The same network cannot adapt to data of different dimensions. Therefore, we propose unidirectional residual convolutional network capable of spatial feature extraction in both temporal and spatial directions without destroying the feature correlation of the time series. In addition, this network can effectively solve the gradient vanishing problem caused by deep CNN.

- 2) For the temporal correlation of massive data, BiLSTM is used to learn the degradation information of spatial features provided by CNN to capture the distantly correlated features in the time series, instead of the traditional method that

utilizes BiLSTM to predict the RUL for each time step. Secondly, traditional methods use attention mechanism (AM) to assign weights to values at different time points. In contrast, the purpose of introducing AM in this paper is to strengthen the linkage enhancement between spatial-temporal features and key degradation information, so as to effectively improve the RUL prediction accuracy of the model.

The rest of this paper is organized as follows. Section II reviews the relevant content of URCNN and BiLSTM. Section III details the Spatial-temporal Feature Extraction Network. Section IV applies the proposed model to the benchmark dataset and the specialized transmission mechanism dataset and provides analysis. Section V presents some conclusions and future work.

## II. THEORETICAL BACKGROUND

### A. UNIDIRECTIONAL RESIDUAL CONVOLUTIONAL NETWORK

The structure of a traditional Convolutional Neural Network (CNN) is shown in Figure 1, mainly composed of convolutional layer (CL), pooling layer (PL), and fully connected layer (FCL) [25]. The CL extracts feature from the input layer (IL) data, the pooling layer selects and filters these features according to specific rules, and the fully connected layer generally connects with the output layer (OL), equivalent to the hidden layer of traditional feed forward neural networks, implementing the final output dimension transformation. The model can adaptively extract spatial features from multi-dimensional samples, making it easy to implement sample reconstruction and feature extraction, and effectively obtain performance degradation information [26].

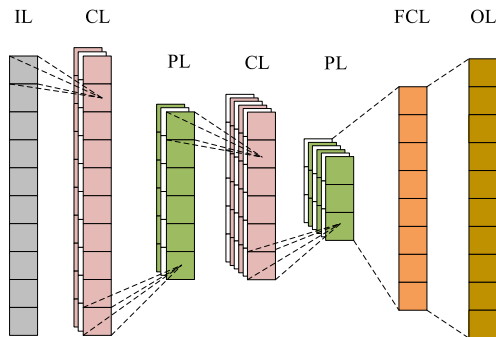


FIGURE 1. Structure of CNN.

The more convolutional layers in a CNN, the richer the deep features that can be extracted. However, deepening the network can lead to a vanishing gradient problem, which in turn reduces the accuracy of remaining life prediction [27], [28]. The deep residual network (ResNet) can effectively overcome the impact of vanishing gradient and can improve feature extraction and learning capabilities [29]. Constructing a deeper network architecture based on residual thinking can obtain more robust feature expressions, thereby improving the accuracy of remaining life prediction. Figure 2 (a) is a schematic diagram of the structure of a single module

of a traditional residual network, where regular convolution operations disrupt the time sequence of input data. To solve this problem, this paper proposes a unidirectional residual convolutional network structure where stacking several unit modules can construct networks of different depths that can extract spatial features in both temporal and spatial directions without destroying the feature correlation of the time series. Simultaneously, batch normalization operations are introduced to accelerate module training speed and improve model training accuracy.

The structure of a single module is shown in Figure 2(b). If  $\tilde{X}$  is the input of the unidirectional residual module, the definition of the nonlinear mapping function  $F(\cdot)$  in the convolution module (Conv) is given by Equation (1):

$$F(\tilde{X}, \{W'_i\}) = W_3 \otimes f(W_2 \otimes f(W_1 \otimes \tilde{X})) \quad (1)$$

where  $W_1$ ,  $W_2$  and  $W_3$  are the weight matrices of the three convolutional layers of the unidirectional residual module, and  $f(\cdot)$  represents the functional mapping of BN operation and ReLU activation function. Then the nonlinear mapping function of a single module is given by Equation (2):

$$\tilde{Y} = F(\tilde{X}, \{W'_i\}) + \tilde{X} \quad (2)$$

Taking the partial derivative of Equation (2), we get Equation (3):

$$\begin{aligned} \frac{\partial \tilde{Y}_L}{\partial \tilde{X}_l} &= \frac{\partial F(\tilde{X}_l, W_l) + \partial \tilde{X}_l}{\partial \tilde{X}_l} \\ &= 1 + \frac{\partial F(\tilde{X}_l, W_l)}{\partial \tilde{X}_l} \end{aligned} \quad (3)$$

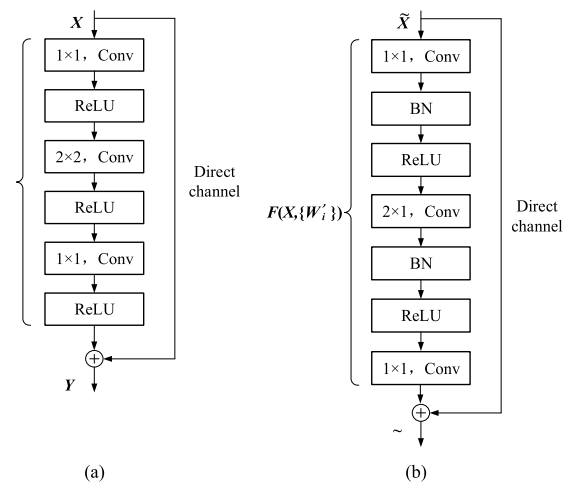


FIGURE 2. Residual network module. (a) Classic residual module; (b) Unidirectional residual module.

From Equation (3), it can be seen that the problem of vanishing gradients can be effectively solved by constructing a direct channel from input to output. Comparing Figures 2 (a) and (b), according to the characteristics of the full life cycle dataset, the unidirectional residual module

replaces the convolution kernel size of the second convolutional layer of the classic residual module from  $2 \times 2$  to  $2 \times 1$ . At the same time, the actual convolution process realizes the unidirectional movement of the convolution kernel, which does not disrupt the temporal correlation of the time series data, i.e., the information fusion of a single time series of mechanical equipment sensor data is carried out without changing the temporal relationship of the data, thereby extracting features related to the remaining life.

**B. BIDIRECTIONAL LONG SHORT-TERM MEMORY NETWORK**

The main approach to remaining life prediction is to learn spatial-temporal information from mechanical equipment sensor data related to equipment performance degradation, and the RNN can fully utilize the time-related characteristics of the data to process time-domain sequence data. However, when the length of a single time series is long or the time interval is short, the RNN may have problems such as vanishing gradient and exploding gradient [30].

LSTM introduces the concept of memory cells, which is composed of four inter-connected structures: the input gate, forget gate, selection gate, and output gate. By comparing memory information and current information, LSTM chooses important information and forgets secondary information, giving the network a stronger memory ability. This can alleviate the problems of vanishing gradient and exploding gradient during the training process of long sequence RNN networks. Its network topology structure is shown in Figure 3.

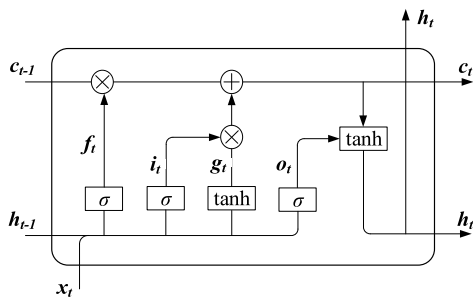


FIGURE 3. Structure of LSTM network.

In Figure 3,  $h_{t-1}$  and  $h_t$  are the hidden states at times  $t-1$  and  $t$ , respectively;  $c_{t-1}$  and  $c_t$  are the gate unit states at times  $t-1$  and  $t$ , respectively;  $x_t$  is the input at time  $t$ . The forget gate  $f_t$  is responsible for selectively discarding the information input from the previous node; the input gate  $i_t$  merges memory information and current information and inputs it into the selection gate  $g_t$ ; the selection gate  $g_t$  selectively remembers the merged information; the output gate  $o_t$  determines the impact of the current gate unit state on the hidden layer output. The current gate unit state  $c_t$  is affected by the gate unit state at the previous moment, the forget gate, the input gate, and the selection gate, while the output of the memory unit is determined by the output gate  $o_t$  and the gate unit state  $c_t$ . The specific calculation process

is shown in Equation (4).

$$\begin{cases} f_t = \sigma(W_f[x_t, h_{t-1}] + b_f) \\ i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \\ g_t = \tanh(W_g[x_t, h_{t-1}] + b_g) \\ o_t = \sigma(W_o[x_t, h_{t-1}] + b_o) \\ c_t = f_t \times c_{t-1} + i_t \times g_t \\ h_t = o_t \times \tanh(c_t) \end{cases} \quad (4)$$

In the equation,  $W_f$  and  $b_f$ ,  $W_i$  and  $b_i$ ,  $W_g$  and  $b_g$ ,  $W_o$  and  $b_o$  are the weights and bias matrices of the forget gate, input gate, selection gate, and output gate, respectively. The Sigmoid activation function  $\sigma(\cdot)$  transforms the output to the  $[0,1]$  interval;  $\tanh(\cdot)$  is the hyperbolic tangent activation function, which transforms the output to the  $[-1,1]$  interval.

Memory units or gate units (cell) are the core of the LSTM network. By forming a forward propagation chain structure through several memory units, it can achieve the overall coordination and transmission of information [31]. As can be seen from Equation (4), the memory unit can precisely control the information flow by integrating all internal state information and input information at all times, ensuring the stability of gradient descent during the model training process. In order to deepen the extraction of original time series features and further improve the accuracy of the model output, two independent LSTMs in different directions are stacked together to form a BiLSTM network, as shown in Figure 4.

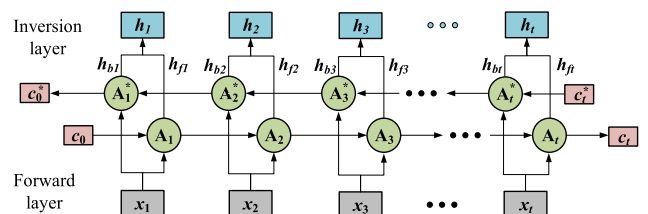


FIGURE 4. BiLSTM network structure diagram.

The input  $x_t$  is input into the forward layer, and the output  $h_f$  of the forward hidden layer is calculated forward from time 0 to  $t$ ; it is input into the backward layer, and the output  $h_b$  of the backward hidden layer is calculated backward from time  $t$  to 0. Finally, the output results of the forward layer and the backward layer are input into the fully connected layer to obtain the final output  $h$ .

$$h = f(h_f, h_b) \quad (5)$$

In the equation,  $f(\cdot)$  is the mapping function of the fully connected layer.

**III. REMAINING USEFUL LIFE PREDICTION MODEL FOR MECHANICAL EQUIPMENT**

**A. OVERVIEW OF THE PREDICTION MODEL**

In the prediction of the remaining useful life of mechanical equipment, important information about the degradation pattern is mainly stored in the historical time series. Traditional



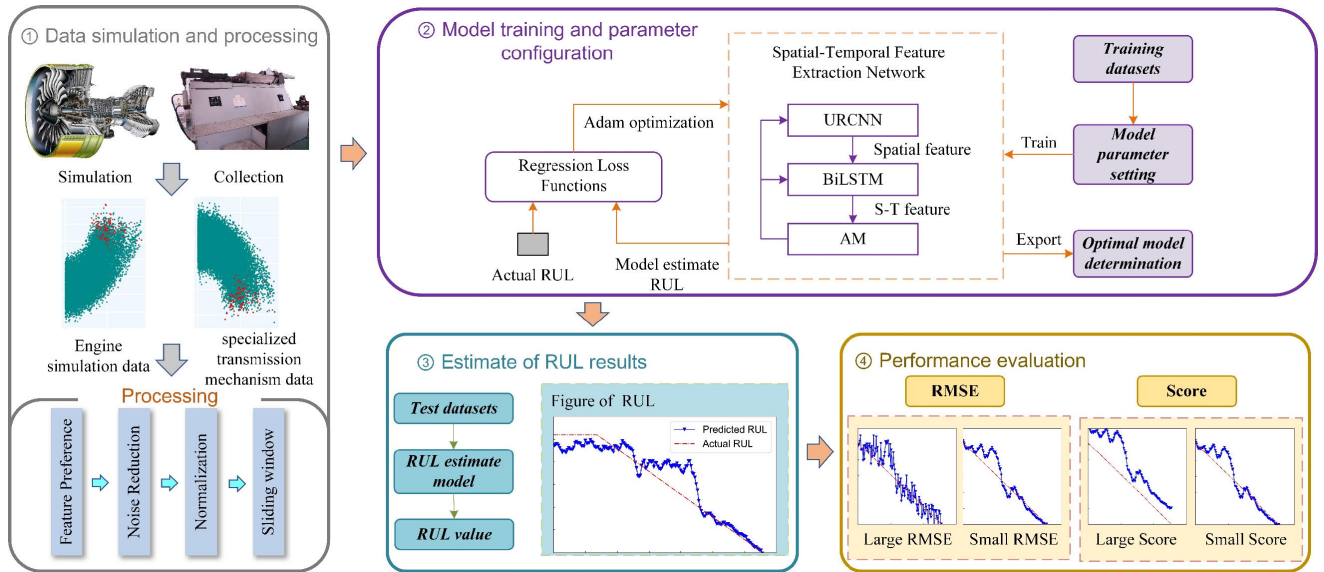


FIGURE 5. Flowchart of the proposed RUL prediction method.

machine learning methods generally select features related to performance degradation from the time series based on prior knowledge, which can disrupt the temporality of the historical data, leading to lower prediction accuracy. Therefore, the method proposed in this paper for predicting the RUL of mechanical equipment based on a spatial-temporal feature extraction network can deeply mine the rich pattern information and degradation trends in sensor data using a fusion model, improving the accuracy of remaining life prediction. Using the spatial-temporal feature extraction network, the RUL can be obtained in four steps as shown in Figure 5.

### 1) STEP 1: DATA SIMULATION AND PROCESSING

In this study, we divided the dataset into a training set and a test set for training and testing the proposed model. Secondly, the preprocessing of the engine simulation data and the specialized transmission mechanism dataset includes four aspects as follows:

#### 1. Feature preference

Taking FD001 of engine simulation data as an example to demonstrate the validation process. It can be observed that the dataset contains setting\_3 and seven sensor values that do not change. To speed up computation, meaningless data are discarded, and the operating cycle is also considered as one of the input features, thus obtaining an optimized dataset with 17 input features.

#### 2. Noise reduction

The original data contains a large amount of Gaussian random noise. A filter function with a window width of 10 is used to denoise and smooth the data to reduce data fluctuations.

#### 3. Normalization

As multi-dimensional monitoring data have different dimensions, normalization preprocessing must be performed

before model construction. In this paper, the Min-Max normalization method is used to unify the data to the [1, 0] range. The specific equation is as follows:

$$\tilde{x}_i^j = \frac{x_i^j - x_{\min}^j}{x_{\max}^j - x_{\min}^j} \quad (6)$$

In the equation:  $\tilde{x}_i^j$  is the normalized data;  $x_{\max}^j$  and  $x_{\min}^j$  are the maximum and minimum values of the  $j$  sensor data.

#### 4. Sliding window

In order to deeply mine the degradation pattern from the limited series and transform the time series into a three-dimensional input format that the BiLSTM network is good at processing, a sliding time window segmentation method is used to process the normalized data. This can fully retain the time correlation between adjacent sequences and increase the number of training set samples, making the model more robust and generalizable. Let the original time series length be  $T$ , the feature dimension be  $N$ , and use a sliding window with a width of  $S$  to slide along the time series, stacking the time series cut for each sliding step to the third dimension to form a three-dimensional tensor of  $(T - S, S, N)$ . The calculation process is as follows:

$$X_{1:T-S} = x_{1:1+S} \oplus x_{2:2+S} \oplus \dots \oplus x_{i:i+S} \dots \oplus x_{T-S:T} \quad (7)$$

In the equation:  $X_{1:T-S}$  is the transformed three-dimensional tensor;  $x_{i:i+S}$  is the sequence of length  $S$  starting from the  $i$ -th time period;  $\oplus$  represents the connection of data within the window in the third dimension to form a three-dimensional tensor.

### 2) STEP 2: MODEL TRAINING AND PARAMETER CONFIGURATION

This study uses a spatial-temporal feature extraction network to learn the complex relationship between the feature

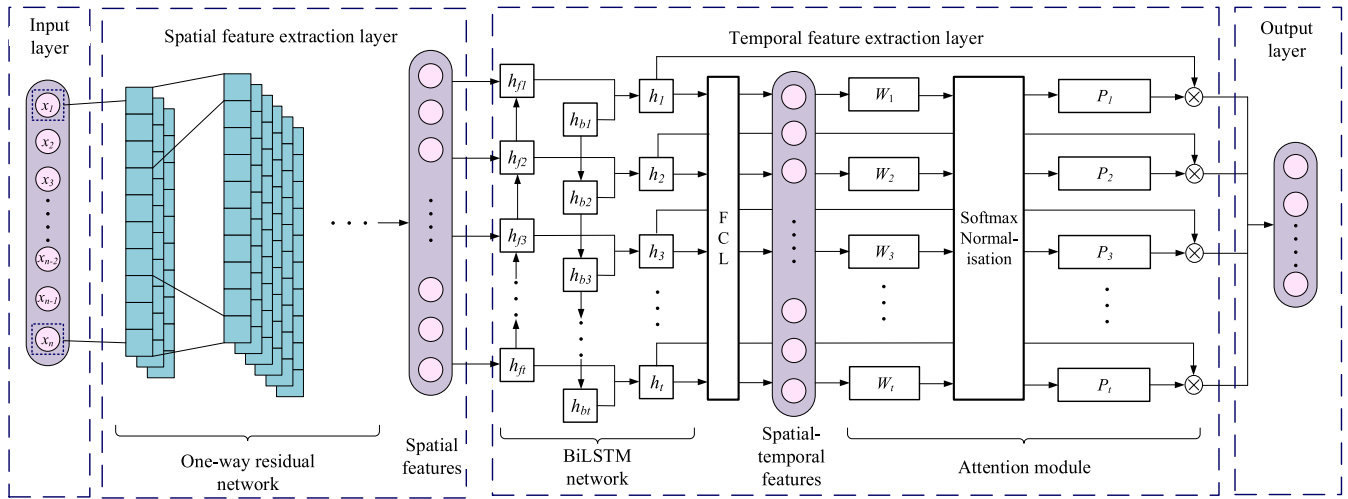


FIGURE 6. Structure diagram of the spatial-temporal feature extraction network.

variables and RUL. First, set the number of residual convolutional layers, BiLSTM layers, and attention layers, and then determine the model’s hyperparameters, including network structure, learning rate, batch size, number of iterations, Dropout rate, etc. Finally, Adam is chosen as the optimizer to train the model, and the Mean Square Error, a commonly used regression loss function, is used to optimize the network and determine the optimal hyperparameters of the model.

3) STEP 3: ESTIMATE OF RUL RESULTS

After training the spatial-temporal feature extraction network with the selected optimal model parameters, the test input sequence is input into the trained model to obtain the final results of the test set.

4) STEP 4: PERFORMANCE EVALUATION

Two standards are used to comprehensively evaluate the performance of the proposed model. The root mean square error (RMSE) is used to measure the deviation between the predicted values and the actual values. The smaller the RMSE deviation value, the more accurate the method’s prediction result. Secondly, Score is introduced to evaluate the lag of the prediction result. The lower the Score, the better the method can overcome the impact of prediction lag on equipment safety.

B. SPATIAL-TEMPORAL FEATURE EXTRACTION NETWORK

The structure of the spatial-temporal feature extraction network designed and built in this paper is shown in Figure 6. It is mainly formed by the fusion of two typical models, URCNN and BiLSTM. URCNN can not only fully mine the potential rules of data, automatically extract important features, but also can improve vanishing gradient problem in the training process of deep CNN networks. The BiLSTM network based on the attention mechanism can better learn the degradation pattern from the features extracted by URCNN, and can easily capture long-distance related features in time series. When the input time series is too long in the remaining

life prediction, the attention mechanism can assign different weights to the input features, using limited computational resources to screen out key information from a large amount of information, filter or weaken other redundant information, making the model pay more attention to features that have a great impact on performance degradation, thereby solving the problem of information loss caused by the time series being too long.

The model is mainly composed of an input layer, a spatial feature extraction layer, a temporal feature extraction layer, and an output layer. The sensor data of the mechanical equipment are input to the unidirectional residual module, the depth of the network is deepened through repeated convolution operations, the proposed multi-dimensional feature map is input to the fully connected layer to obtain the final output, and the extraction of deep features of the time series is completed. The BiLSTM module learns the rules of change of mechanical equipment with time from the proposed features, the attention module highlights the time period of attention concentration by introducing the weight probability  $P$ , and strengthens the connection between each segment to help the network capture the key time series information. Finally, the output layer makes an accurate prediction of the current remaining life.

1) INPUT LAYER

The input layer is responsible for inputting the historical sensor data of the mechanical equipment into the prediction model. The operating conditions and different types of sensor monitoring data are denoted as the related feature time series matrix  $X = (x_1, x_2, \dots, x_T) = (x_1^1, x_2^1, \dots, x_1^N, x_2^N, \dots, x_T^N)^T$ , which can be expanded to represent the Equation (8):

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^N \\ x_2^1 & x_2^2 & \dots & x_2^N \\ \vdots & \vdots & \dots & \vdots \\ x_T^1 & x_T^2 & \dots & x_T^N \end{bmatrix} \in \mathbb{R}^{T \times N} \quad (8)$$

In the equation:  $x_t = [x_t^1, x_t^2, \dots, x_t^N](1 \leq t \leq T)$  is the value sequence of  $N$  related variables at time  $t$ ,  $x^n = [x_1^n, x_2^n, \dots, x_T^n](1 \leq n \leq N)$  is the value sequence of the  $n$ -th related variable at  $T$  historical moments.

### 2) SPATIAL FEATURE EXTRACTION LAYER

The feature extraction layer is responsible for extracting the features of the input time series. By constructing a number of unidirectional residual modules to form an RCNN framework, and adding a maximum value pooling layer in the module to retain more data fluctuation information. After convolution and pooling operations, the original data is mapped to the hidden feature space, the ReLU activation function is selected for activation, a fully connected structure is built to convert it into a one-dimensional structure, and parameters are simplified to speed up training. The input is a single time series, and the structure of the 3-layer residual module is shown in Figure 7, and its calculation process is as follows:

$$R_1 = F(X \otimes W_1 + b_1) = \text{ReLU}(X \otimes W_1 + b_1) \quad (9)$$

$$P_1 = \max(R_1) + b_2 \quad (10)$$

$$R_2 = F(P_1 \otimes W_2 + b_3) = \text{ReLU}(P_1 \otimes W_2 + b_3) \quad (11)$$

$$R_3 = F(R_2 \otimes W_3 + b_4) = \text{ReLU}(R_2 \otimes W_3 + b_4) \quad (12)$$

$$H_o = \text{Sigmoid}(R_3 \times W_4 + b_5) \quad (13)$$

In the equations:  $R_1, R_2$ , and  $R_3$  are the outputs of residual convolution 1, 2, and 3, respectively;  $P_1$  is the output of the pooling operation;  $W_1, W_2, W_3$ , and  $W_4$  are the weight matrices;  $b_1 \sim b_5$  are bias vectors;  $\otimes$  is the convolution operation; the output of the feature extraction layer can be

denoted as  $H_o = [h_1^C, h_2^C, \dots, h_t^C, \dots, h_T^C]^T$ , where  $C$  is the number of extracted features, and  $T$  is the time length.

As can be seen from Figure 7, assuming that the size of the input data is  $14 \times 1$ , the number of convolution kernels is 2, after the residual convolution operation (Rconv)1, it becomes two feature maps of size  $2 \times 13 \times 1$ ; then after the maximum value pooling operation of size  $2 \times 1$ , the output of size  $2 \times 12 \times 1$  is obtained; after the residual convolution operations 2 and 3, the output of size  $4 \times 11 \times 1$  is obtained; then the output node is unfolded into a fully connected layer, and the final output of the feature extraction layer is calculated using the activation function, thereby realizing feature extraction and dimension reduction.

### 3) TEMPORAL FEATURE EXTRACTION LAYER

For the task of predicting the remaining life of mechanical equipment, the sensor data features as input are a series of continuous and highly correlated time series, which requires the network to have a certain “memory function”, by learning the differences in information before and after to judge the remaining life of the current equipment. The feature extraction layer can mine data features from different angles, but it is difficult to learn sequence information with obvious time correlation. Therefore, a BiLSTM module and an attention module were added after the feature extraction layer, allowing the network to learn time series information. The time series information learning layer shown in Figure 6 uses the output of the BiLSTM module as the input of the attention module, solving the problem of loss of long time series information.

1. BiLSTM module: BiLSTM learns the deep features extracted by URCNN from both forward and backward directions. As shown in Figure 6, the features at each moment  $t$  are input to the forward layer and the backward layer, the forward output  $h_{ft}$  and the backward output  $h_{bt}$  are obtained, and finally the output of the BiLSTM module  $h_t, t \in [1, T]$  is calculated by Equation (5).
2. Attention module: Let the remaining life label at time  $t$  be  $Y = (y_1, y_2, \dots, y_t, \dots, y_T)$ . First, the feature weight parameter  $W_t$  of  $h_t$  is calculated through a fully connected neural network, which is used to represent the relevance of the target value  $Y'$  obtained from  $h_t$  and  $Y$ , highlighting the time period of attention concentration, and then the Softmax function is used to normalize the weight, to obtain the weight probability  $P_t$  where the sum of all feature weights is 1, and finally the attention module output  $s_t$  is calculated by weighted sum of the input  $h_t$  based on the weight probability. The calculation process is as follows:

$$W_t = \tanh(w h_t + b) \quad (14)$$

$$P_t = \frac{\exp(W_t)}{\sum_{j=1}^t W_j} \quad (15)$$

$$s_t = \sum_{i=1}^T P_i h_i \quad (16)$$

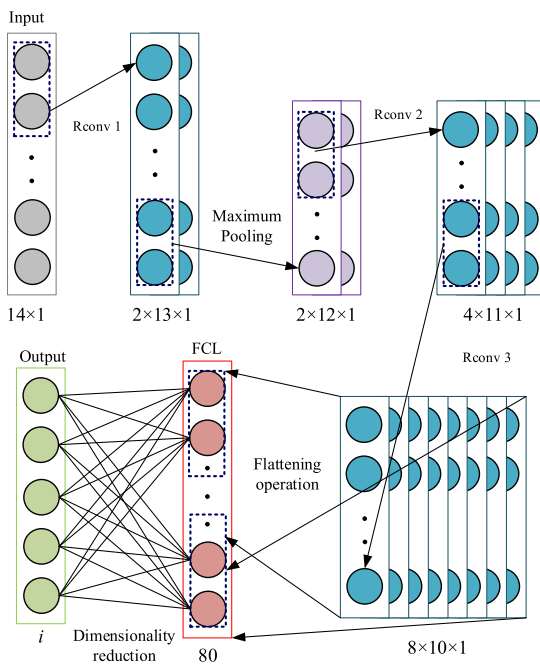


FIGURE 7. Feature extraction layer structure.

In the equations:  $w$  is the weight coefficient of the fully connected layer, and  $b$  is the bias vector.

4) OUTPUT LAYER

The output of the attention module is used as the input of the output layer, and the predicted value  $Y' = [y'_1, y'_2, \dots, y'_t, \dots, y'_T]^T$  of the remaining life of the equipment is calculated through the fully connected layer. The calculation equation can be represented as Equation (17):

$$y'_t = \text{Sigmoid}(w_o s_t + b_o) \tag{17}$$

In the equation:  $y'_t$  is the prediction value of the model at time  $t$ ;  $w_o$  is the weight matrix;  $b_o$  is the bias coefficient; the Sigmoid function is selected as the activation function.

IV. VERIFICATION AND ANALYSIS

In this section, two datasets, namely the standard CMAPSS dataset and the full-life-cycle monitoring data for specialized transmission mechanism, are utilized to validate the efficacy of the proposed method in the precise prediction of the Remaining Useful Life (RUL) of mechanical systems. Section A details the selected benchmark datasets along with the experimental handling process of said datasets. Section B describes the experimental procedure for the specialized transmission mechanism test system. The model is developed utilizing the TensorFlow 2.0 framework. Under this framework, the model is capable of being trained via GPU. The experimental model is executed on a GeForce RTX 3090.

A. CASE1: BENCHMARK DATASET C-MAPSS

1) DATASET DESCRIPTION AND PREPROCESSING

NASA’s C-MAPSS aviation engine dataset is one of the most widely used public datasets in the field of remaining life prediction, which includes four sets of sensor monitoring data from normal operation to failure of turbofan engines under different fault modes and working conditions, and uses 21 types of sensors as typical indicators that can characterize the operation status of the engine. Each data set consists of a training set, a test set, and RUL labels. The training set contains all the data throughout the entire life cycle of the engine, the test set only contains part of the data from the initial state of the test engine, and the RUL label is the RUL at the last monitoring moment of the engine corresponding to the test set. The individual data files contain different numbers of engines, each with different degrees of initial wear, thus the sequence length of each engine’s monitoring data is also different. The specific information of the four sets of monitoring data is shown in Table 1.

2) VERIFICATION PROCESS AND RESULT ANALYSIS

a: RUL LABEL SETTING

In the early stage of engine operation, the performance is good and the degradation can be ignored. However, at the end of usage, the performance of the engine will decrease sharply over time. If the label of the monitoring data before

TABLE 1. Basic information of CMAPSS datasets.

	Dataset	Number of Engines	Total Data Volume	Operating Condition	Fault Mode
FD001	Training set	100	20631	1	1
	Testing set	100	13096		
FD002	Training set	260	53759	6	1
	Testing set	259	33991		
FD003	Training set	100	24720	1	2
	Testing set	100	16596		
FD004	Training set	249	61249	6	2
	Testing set	248	41214		

the rapid deterioration of the engine performance is set to the total running period minus the current running period, it will increase the lag of the RUL prediction result. Therefore, it can be considered that the RUL before the engine starts to degrade rapidly remains unchanged, that is, a threshold is set for the RUL label of the training set to make it a piecewise linear function. Studies have shown that setting the critical value of the sudden change in the training set RUL label to the 130th running cycle has a good prediction effect [32], and the result of the RUL label setting is shown in Figure 8.

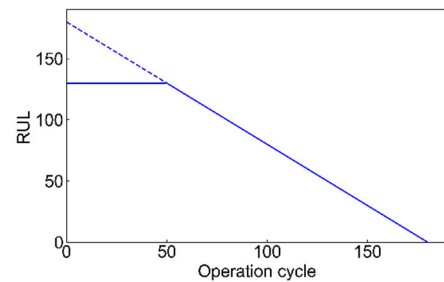


FIGURE 8. RUL label setting.

b: HYPERPARAMETER SETTING

The main hyperparameters involved in this model include: network structure, learning rate, batch size, number of iterations, Dropout rate, etc. The model hyperparameters have a significant impact on the performance of the model. Therefore, the model prediction error is minimized by adjusting a single parameter to obtain the optimal hyperparameter combination.

At the same time, in order to accurately reflect the error distance between the prediction value and the real value of the different hyperparameter models in the test set, the mean square error is used as the evaluation index to adjust the hyperparameters:

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2 \tag{18}$$



In the equation:  $y_i$  is the real value of the  $i$ -th input data;  $\hat{y}_i$  is the predicted value of the  $i$ -th input data;  $m$  is the number of samples.

Taking the batch size parameter as an example, the model hyperparameter adjustment process is shown in Figure 9. As can be seen from Figure 9, the MMSE is smallest when the batch size is 48, so 48 is the most reasonable batch size parameter for the model. Table 2 shows the optimal hyperparameter combination obtained through 30 repeated experiments. The methods for setting other hyperparameters in the table are all according to the above method.

TABLE 2. Optimal hyperparameter combination.

Parameter	Value	Parameter	Value
S of FD001 to FD004	30/20/30/15	Edge Padding	same
URCNN Convolution Kernel Size	2×1	Optimizer	Adam
URCNN Activation Function	tanh	Learning Rate	0.001
URCNN Layers	7	Batch Size	48
Number of LSTM Neurons	128	Iterations	100
Number of LSTM Layers	4	Dropout Rate	0.3

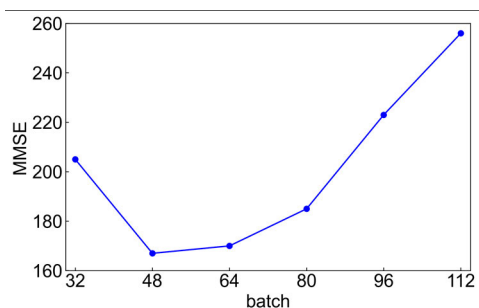


FIGURE 9. Different batch size test results. MMSE is the mean of MSE obtained from 30 trials.

c: MODEL TRAINING AND RESULT ANALYSIS

The preprocessed high-dimensional time series is input into the URCNN network for feature extraction. The original data has 17 features, which is reduced to 7 dimensions after 5 residual units and max pooling operations, and then the dimension reduction features are input into BiLSTM for time information learning. The dataset is divided into a training set and a test set at a ratio of 7:1, and if the error of the test set does not show a decreasing trend in 10 consecutive training epochs, training is stopped early to prevent model overfitting. The changes in test error and training error during the training process are shown in Figure 10. The evaluation indicators for training and testing gradually stabilize and become consistent as the number of training epochs increases, with the training error decreasing from 1852 to 234, and the test error gradually decreasing from 1794 to 263.

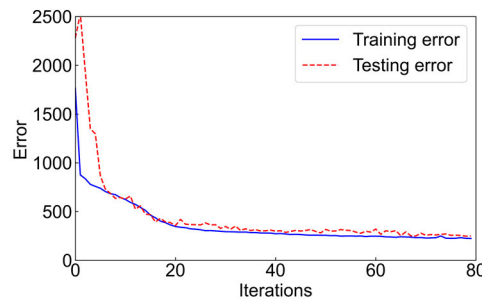


FIGURE 10. RUL prediction training error and test error curves.

The test set is input into the trained model, and the RUL prediction results are shown in Figure 11.

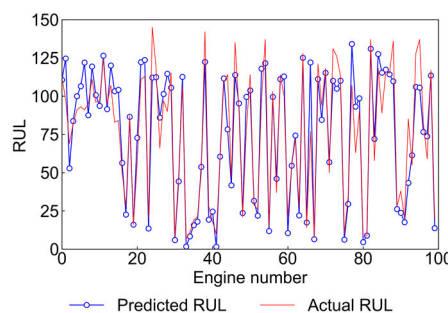


FIGURE 11. FD001 test set RUL prediction results.

To compare the impact of RUL size on model predictions, all engines in the FD001 test set are reordered from large to small according to the actual RUL values, as shown in Figure 12. When the RUL value is large, the engine is in a healthy state, indicating that the equipment is in good running condition. As shown in Figure 12, at the beginning of engine operation, the RUL prediction error is relatively large, with severe fluctuations and obvious lag. However, after a long period of operation, the predicted RUL converges around the actual RUL, significantly enhancing the prediction performance. Therefore, the more sufficient the equipment's historical information and the more obvious the performance

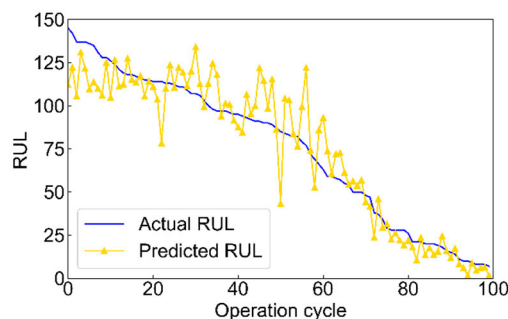
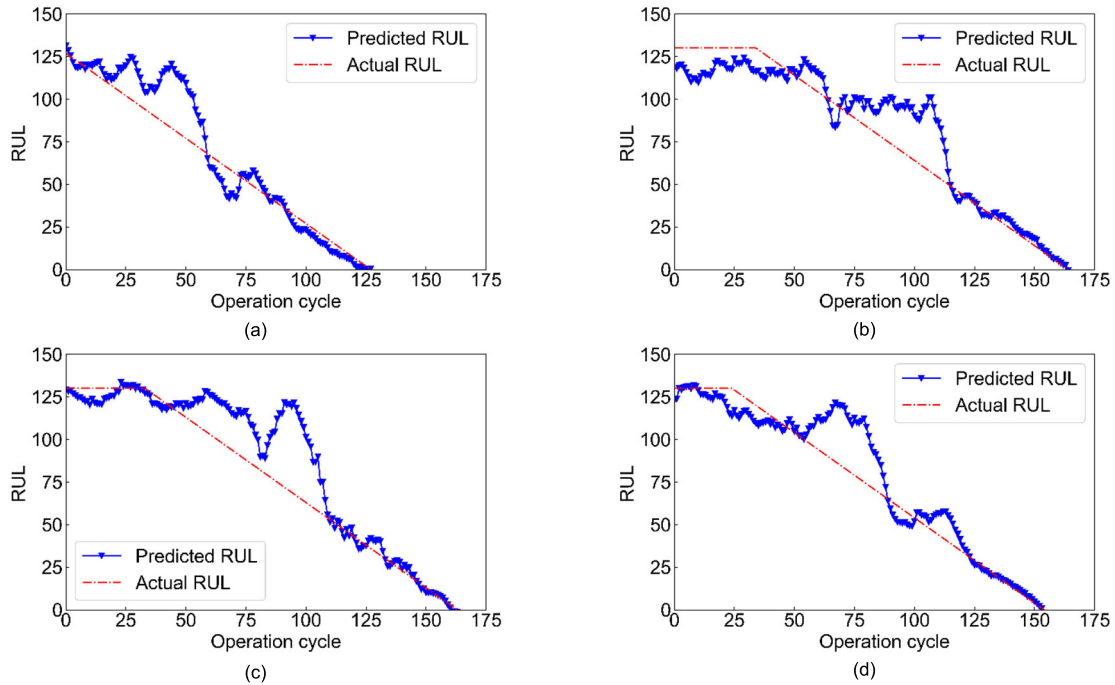


FIGURE 12. FD001 engine reordering prediction results.



**FIGURE 13.** Four engine RUL prediction results. (a) No. 19 engine test results; (b) No. 21 engine test results; (c) No. 30 engine test results; (d) No. 60 engine test results.

degradation information, the smaller the model’s prediction error.

Four random engines from the FD001 dataset are selected to show the continuous RUL prediction results, as shown in Figure 13. The figure shows that the deep convolution structure of the spatial-temporal feature extraction network can effectively extract deep features of engine degradation. Even when the engine just starts to operate and there is little historical data (0-50 cycles), making it difficult to predict the RUL size, the model’s predicted value is relatively close to the critical value of 130.

As the operating cycle increases, the amount of engine performance degradation gradually accumulates. The rapid development of engine losses suddenly intensifies before the model predicts them, causing the predictions to lag behind the actual equipment conditions. The above prediction lag phenomenon is reflected in Fig. 13 as the predicted RUL with peak shape is much higher than the actual value. Nevertheless, BiLSTM can learn the temporal relationship before and after the time series, while the attention mechanism adaptively selects the key time points of performance degradation. The model effectively improves the prediction accuracy over a longer period by combining the advantages of both in learning spatial and temporal features. As shown in Figure 13, in the middle and late stages of engine performance degradation, the model can fit the actual performance degradation curve well, and the RUL prediction results become increasingly accurate and stable. Therefore, the proposed model in this paper has strong capabilities for extracting spatial depth features and remembering long-term temporal features.

### 3) MODEL COMPARATIVE ANALYSIS

In order to objectively evaluate the accuracy and generalization ability of different models on the test set, the Root Mean Square Error (RMSE) and Scoring Function are used to evaluate the RUL prediction effect [18].

RMSE is used to measure the deviation between the predicted value and the real value, and its calculation equation is:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (19)$$

where  $y_i$  is the actual RUL value of the  $i$ -th engine;  $\hat{y}_i$  is the RUL predicted value of the  $i$ -th engine;  $m$  is the total number of engines.

Due to the high safety requirements of the engine, the cost caused by the catastrophic consequences of untimely maintenance is much greater than the cost of excessive maintenance resources. Therefore, the scoring function imposes a higher penalty on the overestimation of RUL. The lower the Score, the better the model’s predictive performance, and the calculation equation is as follows:

$$Score = \begin{cases} \sum_{i=1}^m e^{-\frac{\hat{y}_i - y_i}{13}} - 1 & \hat{y}_i - y_i < 0 \\ \sum_{i=1}^m e^{\frac{\hat{y}_i - y_i}{10}} - 1 & \hat{y}_i - y_i \geq 0 \end{cases} \quad (20)$$

To verify the effectiveness of the proposed spatial-temporal feature extraction network for engine RUL prediction, Support Vector Regression (SVR) [18], CNN [18], LSTM [20],

TABLE 3. Comparing the results of different forecasting models.

Model	FD001		FD002		FD003		FD004	
	RMSE	Score	RMSE	Score	RMSE	Score	RMSE	Score
SVR	20.96	1382	35.96	18900	21.64	2956	36.54	10023
CNN	18.45	1290	30.29	13600	19.82	1600	29.16	7890
GRU	16.45	327	23.37	3868	16.32	1447	27.61	5374
LSTM	16.14	338	24.49	4450	16.18	852	28.17	5550
CLSTM	16.13	303	20.46	3440	17.12	1420	23.26	4630
AE-BiLSTM	13.63	261	19.32	3560	14.21	285	22.45	4886
Transformer	12.31	252	15.35	1267	12.32	296	18.35	2120
Proposed model	12.54	239	17.63	2354	13.36	245	20.54	3326

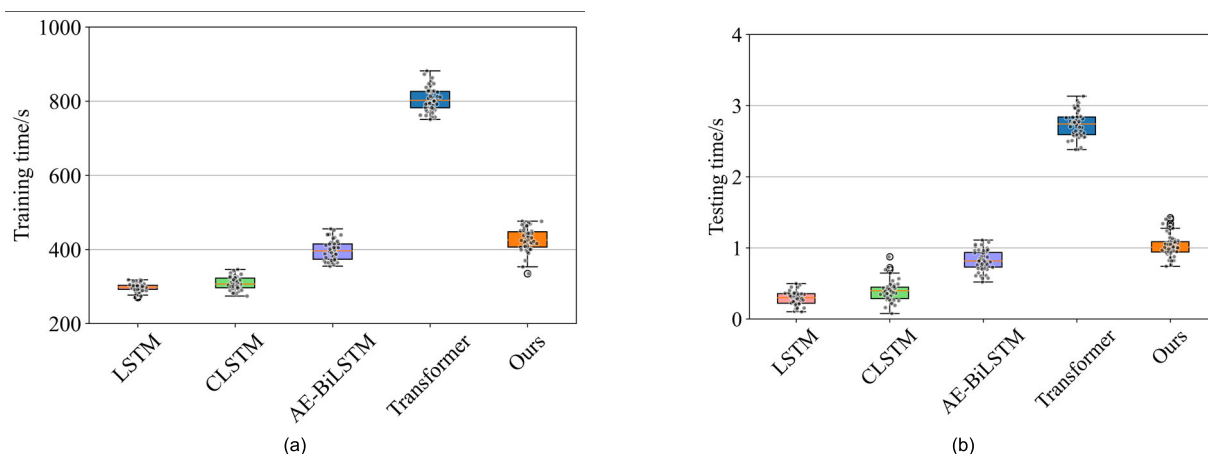


FIGURE 14. Training time and testing time for different models. (a) Training time; (b) Testing time.

GRU [18], CLSTM [33], AE-BiLSTM [23] and Transformer [33] were respectively built as comparison models, and the results are shown in Table 3. As can be seen from Table 3, compared with shallow machine learning methods (SVR) and single-layer deep learning models (CNN, LSTM, GRU), the CLSTM proposed in literature [34] performs slightly worse than LSTM on the subset FD003, but it predicts better on the complex multiple fault mode conditions subsets FD002 and FD004. The AE-BiLSTM proposed in reference [20] employs an autoencoder as a feature extraction tool, while leveraging BiLSTM to capture bidirectional long-term dependency characteristics. Compared with the CLSTM model, the RMSE on datasets FD001 and FD003 decreased by 2.5 and 2.91 respectively, and the Score decreased by 42 and 1135 respectively, further improving the RUL prediction results. However, the Score on datasets FD002 and FD004 increased by 120 and 256 respectively, showing obvious prediction lag. However, the spatial-temporal feature extraction network considers the superiority of the attention mechanism in extracting crucial degradation information from lengthy time series, and the prediction accuracy on all test sets is significantly improved. Compared with the CLSTM model, the proposed model reduced the RMSE on datasets FD001 and FD003 by 22.25% and 21.96%

respectively, and the Score by 21.12% and 82.75% respectively; compared with the AE-BiLSTM model, the RMSE decreased by 8.0% and 5.98% respectively, and the Score by 8.43% and 14.04% respectively. For the more complex and variable datasets FD002 and FD004, compared with RMSE, the performance indicator Score of the proposed model decreases more significantly, indicating that the proposed model still has the advantage of overcoming prediction lag under complex conditions.

Compared with the latest proposed model Transformer, the RMSE and Score of the model on datasets FD001 and FD003 are not much different, indicating that the model performs equally well on stable data. However, on datasets FD002 and FD004, which feature more complex simulated operating conditions, the accuracy of the proposed model does not match that of Transformer, and it exhibits a marginally reduced capacity to mitigate prediction lag. This suggests that larger models such as Transformer are better suited to deep data mining tasks. When working conditions grow more intricate, the advanced larger models demonstrate their prowess in surmounting prediction lag.

However, a significant drawback of deploying large models is the substantial increase in time required for training and inference, which complicates their use in online

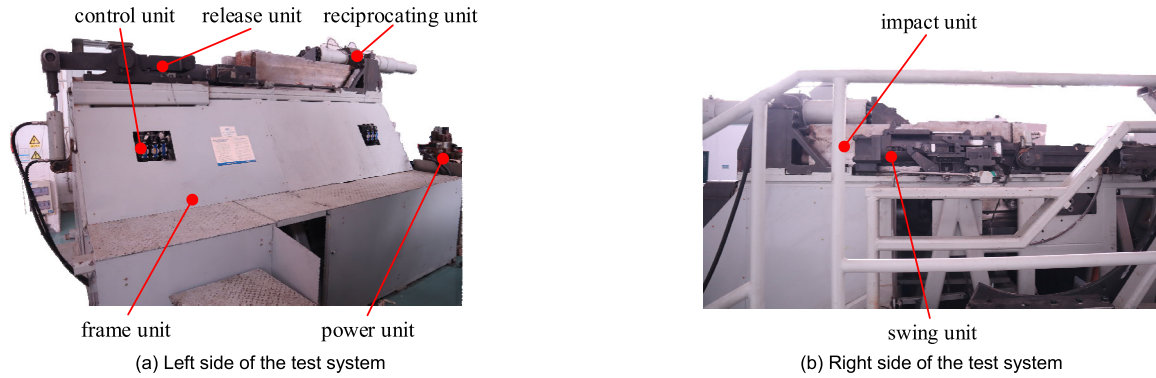


FIGURE 15. Filling mechanism test system.

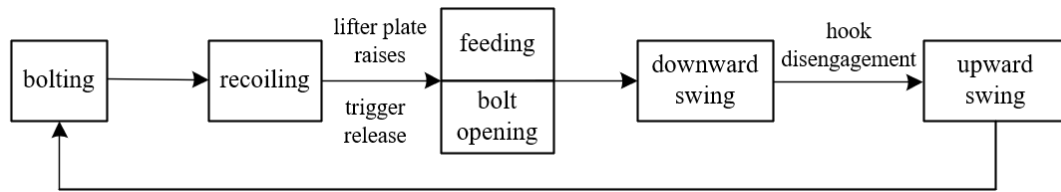


FIGURE 16. Action sequence of test system.

applications. Figure 14 portrays the box plots of training and testing times obtained for different models after 60 repeated trials. The horizontal axis represents the model type, sequentially listing LSTM, CLSTM, AE-BiLSTM, Transformer, and our proposed model, from left to right. The vertical axis depicts the training and testing times in seconds (s), spanning approximately 200 to 1000 seconds for training and varying from roughly 0 to 4 seconds for testing. As depicted in Figure 14, the LSTM base model exhibits the shortest training and testing times. In contrast, the Transformer model records the highest training and testing times, and its distribution is more spread out. Our model’s training and testing times do not significantly deviate from those of the enhanced LSTM model, yet it offers greater accuracy (as indicated in Table 3). Collectively, our proposed model proves more advantageous for real-world deployment, satisfactorily meeting accuracy requirements, particularly in situations with limited datasets. Therefore, we contend that the joint learning model “URCNN+BiLSTM” exhibits greater pragmatism.

**B. CASE2: SPECIAL TRANSMISSION MECHANISM DATASET**

**1) DATASET DESCRIPTION**

We have designed a specialized transmission mechanism test system that operates under reciprocating impact vibrations, as depicted in Figure 15. The test system primarily consists of a frame unit, a power unit, a reciprocating unit, a release unit, a control unit, an impact unit, and a swing unit, among other elements. The system is capable of replicating complex operating conditions characterized by high transients, strong impacts, and severe vibrations that occur during

reciprocation. It supports a maximum reciprocating speed of  $\geq 3.5$  m/s and a maximum reciprocating mass of  $\geq 5$  t, with an adjustable reciprocating distance ranging from 530 mm to 770 mm. The system allows for the full monitoring of the lifecycle of specialized transmission mechanism components under laboratory conditions, providing comprehensive lifecycle data that can be used to validate the proposed intelligent prediction methods. The system utilizes six travel switches to sense the position information and operational statuses of various mechanisms and controls six hydraulic cylinder solenoid valves to orchestrate the mechanisms’ sequential operations, which include bolting, recoiling, release, feeding, bolt opening, downward swing, and upward swing. Consequently, the action sequence of various units is as shown in Figure 16.

The swing unit is capable of performing downward and upward swing motions to transmit the load from the vertical to the horizontal orientation and is the main actuator of the specialized transmission mechanism. The core parts, the slide plate and the ratchet, directly influence the swing unit’s ability to correctly execute the swinging actions and are the most vulnerable life-limited components of the swing unit. The engagement of brand-new components is illustrated in Figure 17. During the reciprocation of the reciprocating unit, the slide plate engages with the ratchet, thereby compressing the downward swing spring to achieve a 90° swing of the load from the horizontal to the vertical direction. The compatibility between the slide plate and the ratchet directly affects the normal operation of the swing unit, which in turn impacts the proper functioning of the specialized transmission mechanism.





FIGURE 17. Ratchet and slide plate and its engagement relationship.

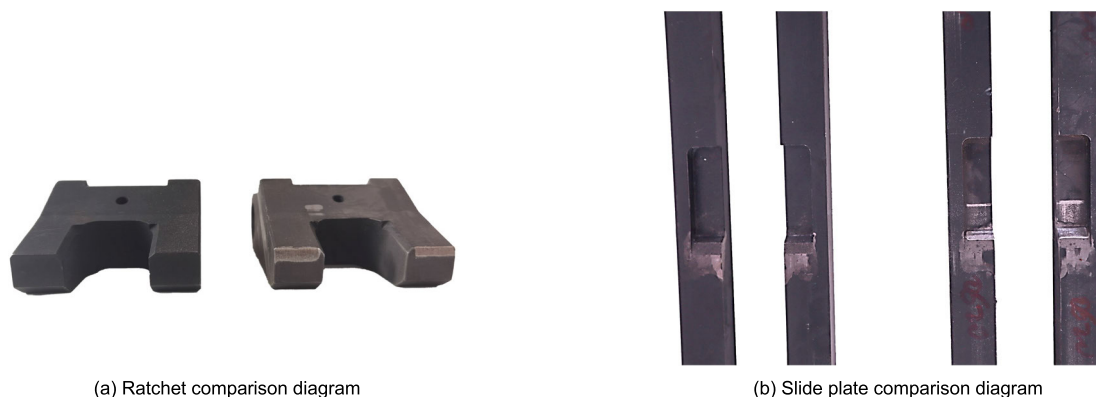


FIGURE 18. Comparison diagram of the ratchet and slide plate (the left side is the health state, the right side is the failure state).

Towards the end of its service life, significant impact wear appears on the meshing surfaces of the slide plate and ratchet, preventing normal engagement and causing the swing unit to cease functioning. At this point, the slide plate and ratchet are deemed to have reached their life threshold and enter a failed state. A comparison between a healthy slide plate and ratchet and one that has experienced significant impact wear after testing is shown in Figure 18. The worn meshing surfaces of the failed slide plate and ratchet are smooth. When the reciprocating slide plate engages with the ratchet, the frictional force in the direction of meshing action cannot sustain the pushing force of the slide plate, resulting in the disengagement of the ratchet. The reciprocating unit is then unable to drive the ratchet to compress the downward swing spring, severely affecting the normal operation of the transmission mechanism.

2) DATA ACQUISITION AND PROCESSING

The experimental signal acquisition system utilized a 32-channel LMS signal collector, as depicted in Figure 19(a). The type of vibration acceleration sensor employed was an ICP acceleration sensor, with the sampling frequency set to 10.24 kHz. During the operating process of the experimental system, transient vibrations produced were quite severe.

The vibration sensors were connected using wired methods, and they were affixed to the surface of the specimen under test using a liquid adhesive, as shown in Figure 19(b).

The mounting position of the vibration acceleration sensors should be as close to the component under test as possible to minimize the noise introduced by the vibration transmission path. Six vibration acceleration sensors, labeled V1 to V6, were arranged near the slide plate as shown in Figure 20(a), and two more, labeled V7 and V8, were positioned near the rollers as illustrated in Figure 20(b). The distances between the different sensor measurement points were suitably moderate, which is beneficial for the acquisition of multisource monitoring data richer in degradation information.

Taking the sensors V3 and V6, which are closest to the life-limited components, as examples: The complete cyclic vibration acceleration signals captured at the beginning and end of monitoring at the V3 measurement point are depicted in Figure 21(a) and 21(b), respectively, and those for the V6 measurement point at the beginning and end of monitoring are shown in Figure 21(c) and 21(d), respectively. It can be discerned from Figure 21 that the vibration acceleration signals obtained during the reciprocating motion of the transmission mechanism represent a typical near-periodic signal with sub-

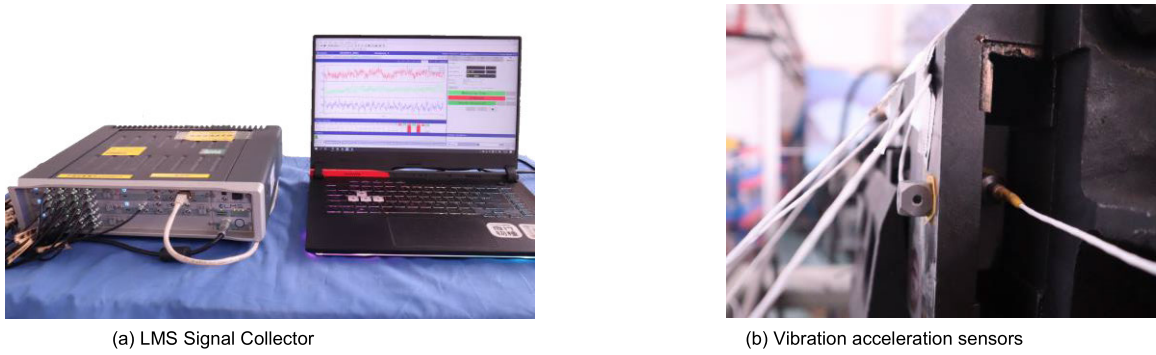


FIGURE 19. Sensor signal acquisition system.

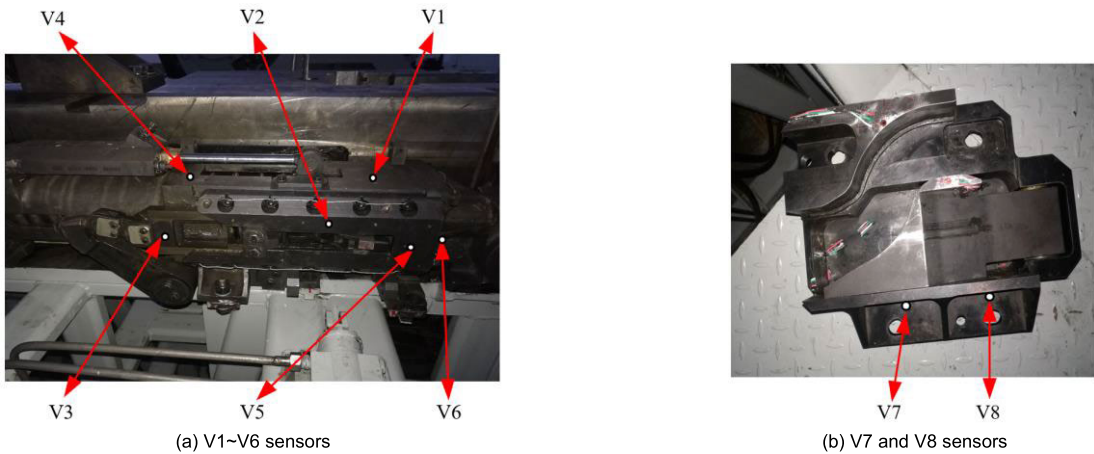


FIGURE 20. Sensor deployment schematic.

stantial transient shock vibrations. Moreover, the peak values of the vibration signals at the inception of monitoring are smaller compared to those at the conclusion, suggesting that as the lifespan of critical components nears its threshold, the performance of the loading mechanism continues to degrade, consequently leading to intensified vibrations of the test apparatus.

A unit cycle signal of the aforementioned near-periodic signal is extracted, with a duration of approximately 25.8 seconds, as shown in Figure 22. The unit cycle signal can be divided into five motion processes: recoiling, feeding, downward swing, upward swing, and bolting. Each motion process causes transient changes in the statistical characteristics of the signal, characterizing it as a typical nonlinear and non-stationary signal.

3) HYPERPARAMETER SETTINGS

The determination of hyperparameters was carried out using the same trial and error method as described in 1) of Section A. Table 4 presents the optimal hyperparameter combination obtained through 30 repeated experiments.

TABLE 4. Optimal hyperparameter combination.

Parameter	Value	Parameter	Value
S of data	30	Edge Padding	same
URCNN Convolution Kernel Size	2×1	Optimizer	Adam
URCNN Activation Function	tanh	Learning Rate	0.0008
URCNN Layers	10	Batch Size	56
Number of LSTM Neurons	256	Iterations	300
Number of LSTM Layers	8	Dropout Rate	0.3

4) MODEL TRAINING AND RESULT ANALYSIS

Ten experiments were repeated, with the slide plate and the ratchet failing on average at the 1586th cycle, thereby setting 1586 as the Remaining Useful Life (RUL) of the brand new mechanism. The vibration data from the eight measurement points was used as the dataset. Since each measuring point can measure vibration data in the x, y, and z directions, the dimensionality of the dataset is 24. Compared to the C-MAPSS dataset, this dataset is higher-dimensional and more nonlinear. The URCNN performs feature extraction on this 24-dimensional data. After passing through

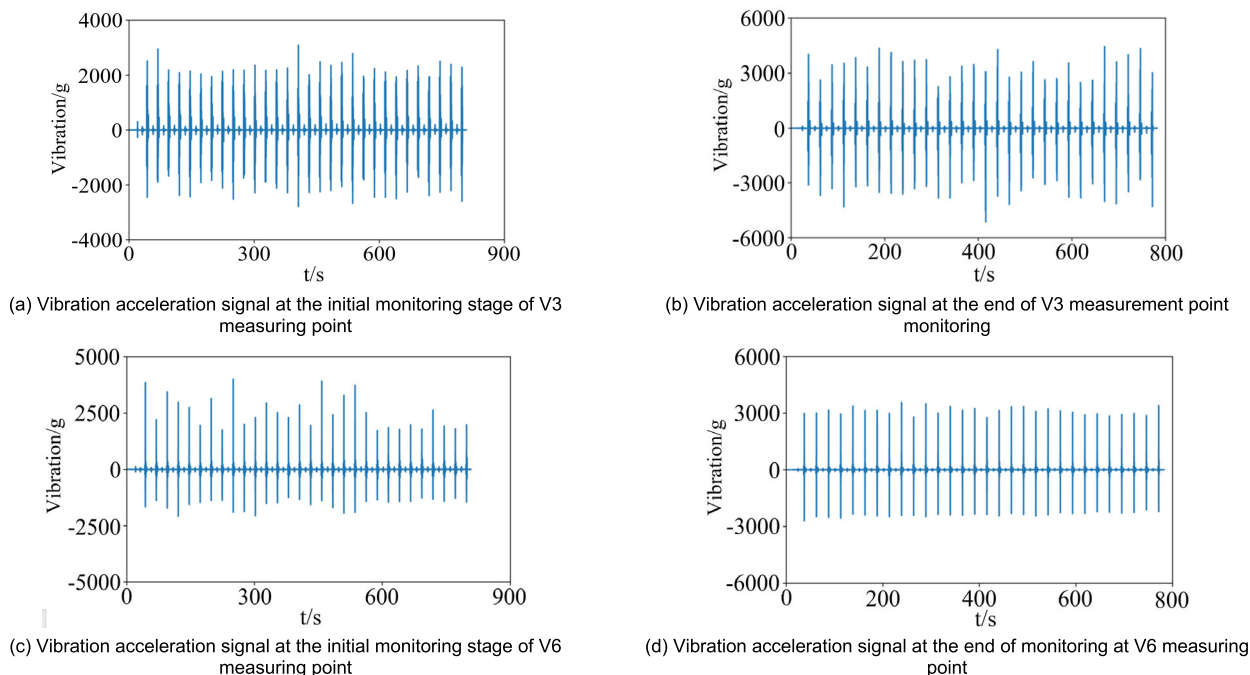


FIGURE 21. Original vibration acceleration signal.

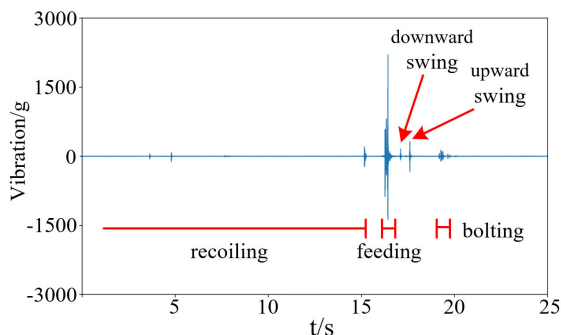


FIGURE 22. Unit period vibration acceleration signal.

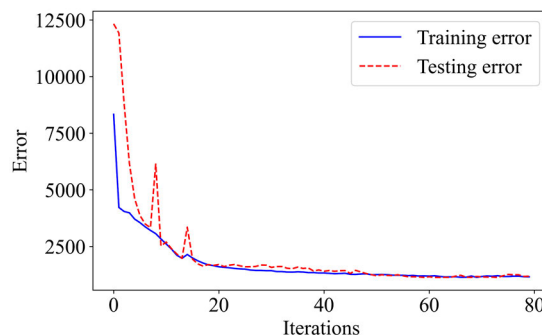


FIGURE 23. Error change curve.

10 residual units and a max-pooling operation, the dimensionality is reduced to 7 dimensions, and then the reduced-dimensionality features are input into ABiLSTM for temporal information learning. The data is split into a training set and a test set using a 8:2 ratio, meaning 8 sets of data are used for training, and 2 sets of data are for testing. The proposed model is trained using the training data, and the variation of testing error and training error during the training process is illustrated in Figure 23. As the number of training epochs increases, the training and testing evaluation metrics gradually stabilize and converge, with the training error decreasing from 8332 to 1156, and the testing error gradually reducing from 12329 to 1182.

When the test set is input into the trained model, the obtained RUL interval prediction results at a 95% confidence level are shown in Figure 24. The pink area in the figure represents the RUL interval at a 95% confidence level, where the

parameters of the RUL probability distribution are obtained through fitting the results of the six sets of test data for each cycle to a normal distribution. The RUL interval represents the uncertainty associated with the model's RUL prediction.

As indicated by Figure 24, during the initial stages of operation, the predicted RUL values significantly deviate from the true values. This discrepancy may arise because the performance degradation of the transmission mechanism is not sufficiently apparent at the start of operation. Additionally, transient shocks and vibrations produced during the operational process heavily influence the prediction model, resulting in predicted RUL values that are less than the actual values. However, as the operating cycles increase, the performance degradation of the transmission mechanism accumulates, enabling the model to capture the degradation information from nonlinear monitoring data across the time series. The point prediction results for RUL align closely with

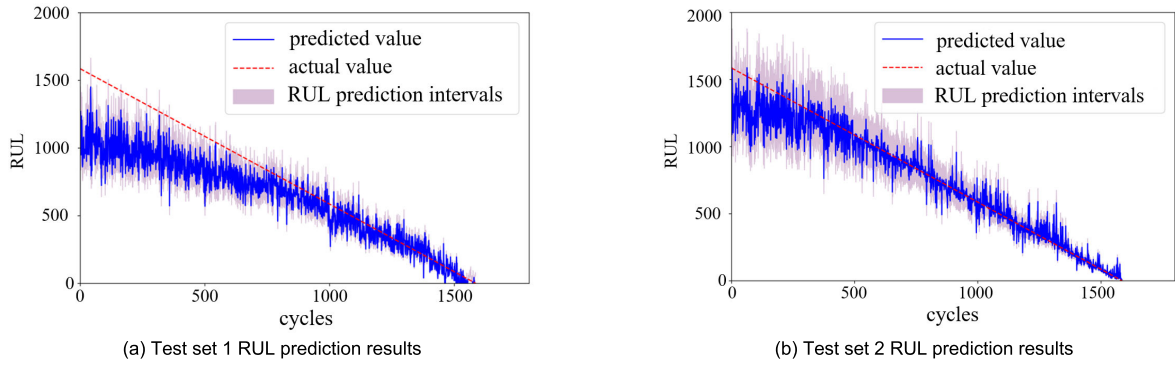


FIGURE 24. Predicted results for the RUL interval.

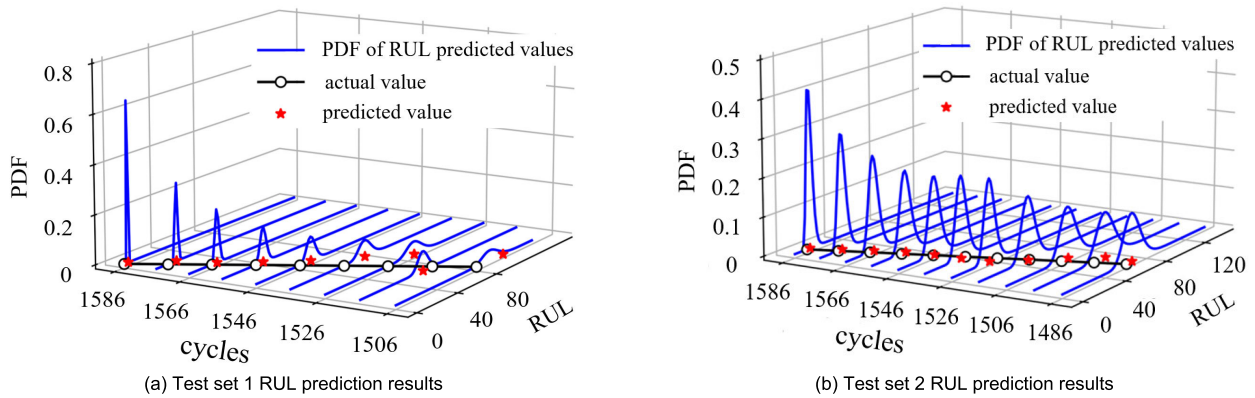


FIGURE 25. Predicted results of RUL at the end of transmission mechanism operation.

the actual values, and the 95% confidence interval predictions generally cover the true RUL values. Additionally, the width of the prediction interval expands during periods of fluctuation in point estimates caused by degradation, but gradually narrows as the actual RUL decreases, signifying that the uncertainty in RUL predictions diminishes with increasing monitoring time.

The RUL interval prediction results and the Probability Density Function (PDF) curve for the last 80 cycles of the swing mechanism operation are depicted in Figure 25. As shown in the figure, towards the end of the transmission mechanism’s lifespan (monitoring cycles 1506-1586 and 1486-1586), the predicted RUL values (denoted by the red pentagrams in Figure 25) closely approach the true RUL values (indicated by the black hollow circles in Figure 25). Furthermore, during the monitoring of the swing mechanism, the peak values of the predicted PDF curve gradually increase, indicating that the point predictions of RUL are more trustworthy towards the end of the monitoring period. Consequently, this enhances the persuasiveness of maintenance strategies based on predictive maintenance that are established subsequent to the predictions.

5) MODEL COMPARATIVE ANALYSIS

To investigate the prognostic performance of the method for RUL prediction, the same evaluation metrics as in 3) of

TABLE 5. Comparing the results of different forecasting models.

Model	Test set			
	RMSE	Score	Training Time(s)	Testing Time(s)
SVR	80.20	6091	1097	0.6
CNN	77.49	6510	1351	0.9
GRU	56.45	5213	1290	0.9
LSTM	53.14	5432	1472	1.1
CLSTM	49.28	4852	1499	1.3
AE-BiLSTM	43.25	3890	1532	1.4
Transformer	37.28	2940	4640	2.6
Proposed model	40.37	3379	1710	1.7

section A are employed to objectively assess the accuracy and generalization ability of different models on the test set. Table 5 compares the training effects on the test set for various models (as with those in Section A).

Compared to the seven other methods, the proposed method’s average RMSE and Score metrics are the smallest, with the exception of the Transformer. Further analysis indicates that although the GRU’s training performance is slightly inferior to that of the LSTM, the training time is reduced by about 12%, which is advantageous for the deployment of



real-time prediction. Moreover, the AE-BiLSTM leverages an autoencoder for feature extraction, thereby more easily capturing the bidirectional long-term dependencies, resulting in a reduction of approximately 17% and 26% in RMSE and Score respectively compared to single network models. Additionally, compared to the C-MAPSS dataset, the full-life-cycle data of the specialized transmission mechanism exhibits stronger nonlinearity and noise, and the performance of the proposed method is marginally inferior to that of the larger Transformer model. However, during the offline training phase, the proposed method averages a runtime of 1710s, and only 1.7s during the online testing phase, which is 63% and 34% less time-consuming than the Transformer model, respectively. Additionally, compared with other methods, the average testing time of the proposed method when applied to the dataset differs by only 0.7 seconds, yet it offers higher prediction accuracy, compensating for the temporal losses associated with complex models. Therefore, the aforementioned analysis reveals that a vertical comparison of the prediction results indicates that the current method possesses commendable predictive precision and generalization capability, while also fulfilling real-time requirements.

## V. CONCLUSION

In response to the characteristics of mechanical equipment monitoring data being nonlinear, multi-dimensional, large-scale, and with key degradation information easily submerged in long time series, this paper proposes a method for predicting the remaining useful life of mechanical equipment based on a spatial-temporal feature extraction network. This method combines RCNN and BiLSTM to obtain spatial-temporal features and uses the CMAPSS dataset and special transmission mechanism dataset to perform multi-faceted validation and testing of the model. The experimental results show that:

1. The more sufficient the historical information of the equipment, the more apparent the performance degradation information, and the smaller the prediction error of the model.
2. The spatial-temporal feature extraction network deeply mines the degradation features of multi-dimensional monitoring data through the RCNN model, while ABiLSTM can learn the temporal relationship before and after the time series, and uses the attention mechanism to adaptively select the key time points of performance degradation. The model effectively improves the prediction accuracy over longer periods by integrating the advantages of both in learning spatial and temporal features.
3. Compared with shallow machine learning methods, single-layer deep learning models, and multi-layer deep learning models, the use of a spatial-temporal feature extraction network to process multi-dimensional mechanical equipment monitoring data with complex operating conditions and variable fault modes can accurately locate

degradation time points and enhance the safety of equipment use.

4. Spatial-temporal feature extraction network performs slightly worse than integrated large model on complex datasets with complicated working conditions. However, it is more favorable for rapid deployment and more practical value under the requirement of accuracy.

## REFERENCES

- [1] F.-K. Wang, X.-B. Cheng, and K.-C. Hsiao, "Stacked long short-term memory model for proton exchange membrane fuel cell systems degradation," *J. Power Sources*, vol. 448, Feb. 2020, Art. no. 227591.
- [2] C. Zhang, S. Zhao, Z. Yang, and Y. Chen, "A reliable data-driven state-of-health estimation model for lithium-ion batteries in electric vehicles," *Frontiers Energy Res.*, vol. 10, Sep. 2022, Art. no. 1013800, doi: 10.3389/fenrg.2022.1013800.
- [3] C. Cheng, G. Ma, Y. Zhang, M. Sun, F. Teng, H. Ding, and Y. Yuan, "A deep learning-based remaining useful life prediction approach for bearings," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 3, pp. 1243–1254, Jun. 2020, doi: 10.1109/TMECH.2020.2971503.
- [4] C. Zhang, S. Zhao, and Y. He, "An integrated method of the future capacity and RUL prediction for lithium-ion battery pack," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2601–2613, Mar. 2022, doi: 10.1109/TVT.2021.3138959.
- [5] W. Yu, I. Y. Kim, and C. Mechefske, "Remaining useful life estimation using a bidirectional recurrent neural network based autoencoder scheme," *Mech. Syst. Signal Process.*, vol. 129, pp. 764–780, Aug. 2019.
- [6] A. Cubillo, "Physics-based integrated vehicle health management system for predicting the remaining useful life of an aircraft planetary gear transmission," *Int. J. Structural Integrity*, vol. 8, no. 4, pp. 484–495, Aug. 2017, doi: 10.1108/ijsi-01-2016-0003.
- [7] Z. Zhang, X. Si, C. Hu, and Y. Lei, "Degradation data analysis and remaining useful life estimation: A review on Wiener-process-based methods," *Eur. J. Oper. Res.*, vol. 271, no. 3, pp. 775–796, Dec. 2018, doi: 10.1016/j.ejor.2018.02.033.
- [8] Z. Fan, G. Liu, X. Si, Q. Zhang, and Q. Zhang, "Degradation data-driven approach for remaining useful life estimation," *J. Syst. Eng. Electron.*, vol. 24, no. 1, pp. 173–182, Feb. 2013, doi: 10.1109/JSEE.2013.00022.
- [9] Y. Li, X. Huang, P. Ding, and C. Zhao, "Wiener-based remaining useful life prediction of rolling bearings using improved Kalman filtering and adaptive modification," *Measurement*, vol. 182, Sep. 2021, Art. no. 109706, doi: 10.1016/j.measurement.2021.109706.
- [10] Y. Li, X. Huang, T. Gao, C. Zhao, and S. Li, "A wiener-based remaining useful life prediction method with multiple degradation patterns," *Adv. Eng. Informat.*, vol. 57, Aug. 2023, Art. no. 102066, doi: 10.1016/j.aei.2023.102066.
- [11] J. Zhao, Y. Zhou, Q. Zhu, Y. Song, Y. Liu, and H. Luo, "A remaining useful life prediction method of aluminum electrolytic capacitor based on Wiener process and similarity measurement," *Microelectron. Rel.*, vol. 142, Mar. 2023, Art. no. 114928, doi: 10.1016/j.microrel.2023.114928.
- [12] D. Wu, M. Jia, Y. Cao, P. Ding, and X. Zhao, "Remaining useful life estimation based on a nonlinear Wiener process model with CSN random effects," *Measurement*, vol. 205, Dec. 2022, Art. no. 112232, doi: 10.1016/j.measurement.2022.112232.
- [13] L. Feng, H. Wang, X. Si, X. Yang, and B. Wang, "Online prediction of remaining life based on semi-random filtering-expectation-maximization algorithm," *Acta Aeronautica ET Astronautica Sinica*, vol. 36, no. 2, pp. 555–563, Jan. 2015.
- [14] P. Kundu, A. K. Darpe, and M. S. Kulkarni, "Weibull accelerated failure time regression model for remaining useful life prediction of bearing working under multiple operating conditions," *Mech. Syst. Signal Process.*, vol. 134, Dec. 2019, Art. no. 106302, doi: 10.1016/j.ymssp.2019.106302.
- [15] Y. Zhang, Y. Jia, and T. Feng, "Remaining useful life prediction model of planetary carrier in helicopter main gear-box based on Gamma degradation process," *J. Vibrat. Shock*, vol. 31, no. 14, pp. 47–51, 2012.
- [16] Z. Wang, Y. Chen, and Z. Cai, "Equipment remaining useful lifetime online prediction based on accelerated degradation modeling with the proportion relationship," *Syst. Eng. Electron.*, vol. 43, no. 2, pp. 584–592, 2021.

- [17] Y. Tang, G. Dai, Y. Zhou, Y. Huang, and D. Zhou, "Conflicting evidence fusion using a correlation coefficient-based approach in complex network," *Chaos, Solitons Fractals*, vol. 176, Nov. 2023, Art. no. 114087, doi: [10.1016/j.chaos.2023.114087](https://doi.org/10.1016/j.chaos.2023.114087).
- [18] C. Zhang, L. Luo, Z. Yang, S. Zhao, Y. He, X. Wang, and H. Wang, "Battery SOH estimation method based on gradual decreasing current, double correlation analysis and GRU," *Green Energy Intell. Transp.*, vol. 2, no. 5, Oct. 2023, Art. no. 100108, doi: [10.1016/j.geits.2023.100108](https://doi.org/10.1016/j.geits.2023.100108).
- [19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [20] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, "Long short-term memory network for remaining useful life estimation," in *Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, Jun. 2017, pp. 88–95.
- [21] S. Zhao, C. Zhang, and Y. Wang, "Lithium-ion battery capacity and remaining useful life prediction using board learning system and long short-term memory neural network," *J. Energy Storage*, vol. 52, Aug. 2022, Art. no. 104901, doi: [10.1016/j.est.2022.104901](https://doi.org/10.1016/j.est.2022.104901).
- [22] S. Xiang, Y. Qin, J. Luo, H. Pu, and B. Tang, "Multicellular LSTM-based deep learning model for aero-engine remaining useful life prediction," *Rel. Eng. Syst. Saf.*, vol. 216, Dec. 2021, Art. no. 107927, doi: [10.1016/j.res.2021.107927](https://doi.org/10.1016/j.res.2021.107927).
- [23] Y. Song, T. Xia, and Y. Zheng, "Prediction of remaining life of turbofan engine based on autoencoder-BLSTM," *Comput. Integr. Manuf. Syst.*, vol. 25, no. 7, pp. 1611–1619, 2019.
- [24] J. Chen, D. Chen, and G. Liu, "Using temporal convolution network for remaining useful lifetime prediction," *Eng. Rep.*, vol. 3, no. 3, Mar. 2021, Art. no. e12305, doi: [10.1002/eng2.12305](https://doi.org/10.1002/eng2.12305).
- [25] P.-H. Kuo and C.-J. Huang, "A high precision artificial neural networks model for short-term energy load forecasting," *Energies*, vol. 11, no. 1, p. 213, Jan. 2018.
- [26] D. Yao, B. Li, H. Liu, J. Yang, and L. Jia, "Remaining useful life prediction of roller bearings based on improved 1D-CNN and simple recurrent unit," *Measurement*, vol. 175, Apr. 2021, Art. no. 109166, doi: [10.1016/j.measurement.2021.109166](https://doi.org/10.1016/j.measurement.2021.109166).
- [27] B. Wang, Y. Lei, T. Yan, N. Li, and L. Guo, "Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery," *Neurocomputing*, vol. 379, pp. 117–129, Feb. 2020, doi: [10.1016/j.neucom.2019.10.064](https://doi.org/10.1016/j.neucom.2019.10.064).
- [28] Y. Li, C. Wang, X. Wan, H. Dong, and S. Gong, "Spatiotemporally consistent video event recognition based on deep residual dual unidirectional DLSTM," *Chin. J. Comput.*, vol. 41, no. 12, pp. 2852–2866, 2018.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis.*, Las Vegas, CA, USA, 2016, pp. 770–778.
- [30] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 187–197, May 2015.
- [31] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *Proc. Int. Conf. Prognostics Health Manage.*, Denver, CO, USA, Oct. 2008, pp. 1–9, doi: [10.1109/PHM.2008.4711414](https://doi.org/10.1109/PHM.2008.4711414).
- [32] F. O. Heimes, "Recurrent neural networks for remaining useful life estimation," in *Proc. Int. Conf. Prognostics Health Manage.*, Oct. 2008, pp. 1–6, doi: [10.1109/phm.2008.4711422](https://doi.org/10.1109/phm.2008.4711422).
- [33] H.-K. Wang, Y. Cheng, and K. Song, "Remaining useful life estimation of aircraft engines using a joint deep learning model based on TCNN and transformer," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–14, Nov. 2021, doi: [10.1155/2021/5185938](https://doi.org/10.1155/2021/5185938).
- [34] Z. Kong, Y. Cui, Z. Xia, and H. Lv, "Convolution and long short-term memory hybrid deep neural networks for remaining useful life prognostics," *Appl. Sci.*, vol. 9, no. 19, p. 4156, Oct. 2019.



**XIAOJIA YAN** received the B.S. degree from the School of Mechanical and Electrical Engineering, Wuhan University of Technology, Wuhan, China, in 2020. He is currently pursuing the M.S. degree in weapon science and technology with the School of Weapon Engineering, Naval University of Engineering (NUE), Wuhan. His current research interests include deep learning, signal processing, and remain useful life prediction of complex systems.



**WEIGE LIANG** received the B.S. degree in weapons science and technology from Naval Engineering University, Wuhan, China, in 2005, and the Ph.D. degree in weapons science and technology in 2010. He is currently an Instructor with the School of Weapon Engineering, NUE. His research interests include machine learning, prognostic health management, and control theories and applications.



**SHIYANG SUN** received the Ph.D. degree in systems engineering from Naval Engineering University, Wuhan, China, in 2008. He is currently a Professor with the Naval University of Engineering (NUE). His research interests include deep learning, fault diagnosis, fuzzy mathematics, and systems engineering.

...