actuators

MDPI

Article

Multi-Head Attention Network with Adaptive Feature Selection for RUL Predictions of Gradually Degrading Equipment

Lei Nie *, Shiyi Xu and Lvfan Zhang

Hubei Key Laboratory of Modern Manufacturing Quantity Engineering, Hubei University of Technology, Wuhan 430068, China; xushiyi97@163.com (S.X.); zhanglvfan@163.com (L.Z.)
* Correspondence: leinie@hbut.edu.cn

Abstract: A multi-head-attention-network-based method is proposed for effective information extraction from multidimensional data to accurately predict the remaining useful life (RUL) of gradually degrading equipment. The multidimensional features of the desired equipment were evaluated using a comprehensive evaluation index, constructed of discrete coefficients, based on correlation, monotonicity, and robustness. For information extraction, the optimal feature subset, determined by the adaptive feature selection method, was input into the multi-head temporal convolution network–bidirectional long short-term memory (TCN-BILSTM) network. Each feature was individually mined to avoid the loss of information. The effectiveness of our proposed RUL prediction method was verified using the NASA IMS bearings dataset and C-MAPSS aeroengines dataset. The results indicate the superiority of our method for the RUL prediction of gradually degrading equipment compared to other mainstream machine learning methods.

Keywords: adaptive feature selection; multi-head attention; temporal convolutional network; bidirectional long short-term memory; remaining useful life

Citation: Nie, L.; Xu, S.; Zhang, L. Multi-Head Attention Network with Adaptive Feature Selection for RUL Predictions of Gradually Degrading Equipment. *Actuators* **2023**, *12*, 158. https://doi.org/10.3390/act12040158

Academic Editor: Giorgio Olmi

Received: 4 March 2023
Revised: 25 March 2023
Accepted: 2 April 2023
Published: 3 April 2023

1. Introduction

With the rapid development of the industrial economy, the maintenance cost of machinery and ensuring equipment safety and reliability is an important issue. An accurate prediction of the remaining useful life (RUL) can reduce equipment maintenance downtime, increasing productivity and lowering production costs. In general, the theoretical failure models for mechanical equipment are based on degradation due to performance failure. However, in practical engineering applications, it is often difficult to accurately predict the RUL of gradually degrading equipment.

From the research point of view, several models exist for RUL predictions. These can be classified into three main categories, namely physical [1], data-driven [2–4], and hybrid models [5]. Traditional prediction techniques require accurate theoretical and physical models to characterize the degradation process. However, often it is quite difficult to establish such universal models for the actual production of the equipment, resulting in a loss of time and labor. With the development of data storage technologies, there is an increasing demand to evaluate the health of machinery according to their historically available working status data. The data-driven model uses the sensor monitoring parameters to effectively mine information. It extracts useful feature information through data analysis and characterizes the health status of the equipment to achieve high-precision RUL prediction. The hybrid model offers better accuracy theoretically, owing to the combined advantage of both physical and data-driven models. However, to function properly, it requires a reasonably well-constructed physical model. This is often the most challenging task in RUL prediction, at times even impossible for complex mechanical systems. Therefore, data-driven models are the most extensively applied models in the present age.

Generally speaking, data-driven equipment RUL prediction for equipment consists of two important steps: (1) extracting features that can characterize the equipment's degradation state; and (2) building appropriate predictive models using machine learning methods, such as support vector machine (SVM) [6], grey forecasts model (GM (1, 1)) [7], hidden Markov model (HMM) [8], and so on.

It is necessary to extract features that conform to the equipment degradation trend to characterize the degradation process and predict the RUL. Guo et al. [9] used statistical features extracted from the original vibration signal of a rolling bearing to characterize its degradation state. However, some features unrelated to the bearing degradation process also tend to be extracted in practical applications. Such features need to be screened and rejected to simplify the prediction model and reduce errors, improving the calculation accuracy and efficiency. Mi et al. [10] proposed a double-layer feature selection method to screen out a subset from the candidate features and improve the sensitivity of the degradation trend by eliminating redundancy. Based on this method, it is necessary to build an appropriate model for the time-series RUL prediction. Saufi et al. [11] built a long short-term memory (LSTM) model by integrating the Laplacian score (LS), random search optimization, and LSTM to achieve an accurate RUL analysis for the rolling bearings. Behera et al. [12] proposed a novel RUL prediction method based on a multiscale deep bidirectional gated recurrent neural network (MDBGRU) for large and complex equipment. This method overcomes the pre-expertise requirement on multiple subcomponents of the system and realizes automatic learning both local and global information. Zhang et al. [13] proposed a dual-task network structure based on bidirectional gated recurrent unit (BiGRU) and multigate mixture-of-experts (MMoE), which simultaneously evaluates the HS and predict the RUL of aeroengines. In order to well mine the degradation trend in aeroengines in different levels, Xiang et al. [14] constructs a novel variant LSTM called multicellular LSTM (MCLSTM), which is used to extract the health indicators of aeroengines from raw data.

In order to accurately characterize the degradation process, the characteristic features of the equipment operation process should be extracted to the fullest extent. Feature extraction in practical applications can have several problems such as large dimensions, data redundancy, and high time costs due to manual screening. Consequently, a comprehensive evaluation index was constructed to screen multidimensional features. For improved information mining of the selected features, we propose a prediction method based on a multi-head attention mechanism for improved accuracy and enhanced generalization.

Initially, for multidimensional features, the noise was reduced through exponential smoothing. Subsequently, based on correlation, monotonicity, and robustness, a comprehensive evaluation index J was constructed using the discrete coefficients. The optimal feature subset was obtained by replacing manual screening with adaptive feature selection to improve the working efficiency. Finally, a temporal convolution network–bidirectional long short-term memory (TCN-BILSTM) network, based on the multi-head attention mechanism, was applied to effectively mine and extract information from the optimal feature subset. Different features were modeled individually to realize parallel processing and maximize data integrity preservation, while improving the network's computational efficiency. The results demonstrated higher prediction accuracy and better generalization of our proposed method compared to other mainstream machine learning methods.

The rest of this paper is organized as follows: Section 2 introduces the model construction and the related theory; Section 3 describes the experimental results and analysis; and Section 4 presents the conclusions.

2. Methodology

As mechanical equipment operates over time, depending on its health status, different sensor monitoring signals are produced. Different sensor signals can be regarded as different features. Based on the RUL analysis incorporated with the adaptive feature selection method, we constructed a multichannel prediction network to analyze

and process the different features and retain the useful information of the multidimensional features to the highest possible extent. The framework of the method is shown in Figure 1. This method is divided into two main steps: (1) data processing and (2) prediction model construction.

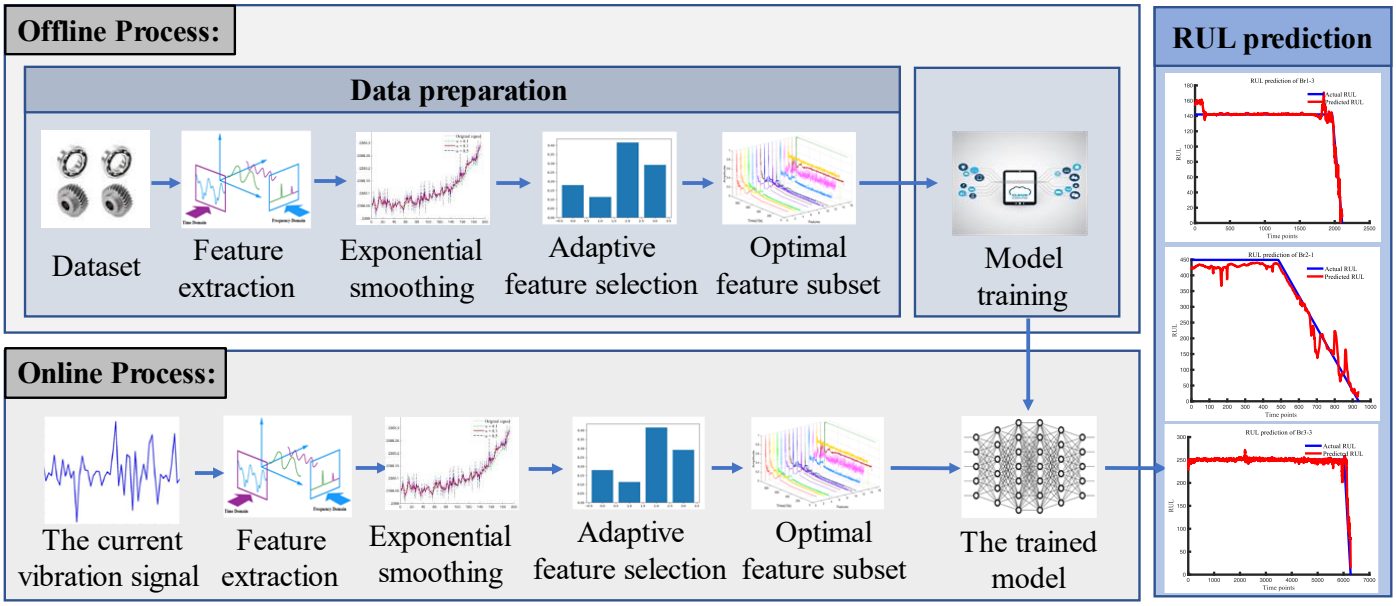


Figure 1. RUL prediction process.

2.1. Data Preprocessing

2.1.1. Exponential Smoothing

Since the correlation between the current parameters of the equipment features and their surrounding values decreases with the increase in the number of sample groups, it is feasible to use the exponential smoothing method for noise reduction [15]. The current value is expressed as the weighted average of the actual current value and the parameter value at the previous moment.

$$\begin{cases} S_t = \alpha \cdot y_t + (1-\alpha) \cdot S_{t-1}, t \geq 2 \\ S_1 = y_1 \qquad\qquad\quad , t = 1 \end{cases} \tag{1}$$

where S_t is the observed value at the t moment, S_t represents the observed value at the moment of $t-1$, y_t is the true value of the t moment, and α is a smoothing constant in the range of 0 to 1.

2.1.2. Adaptive Feature Selection

(I) Feature evaluation indices

The equipment degradation process is essentially a random process of continuous change. An excellent feature should basically meet three particular conditions:

- There exists a correlation between the features and the time series of the RUL; features change with the degradation of the equipment.
- Features should be monotonical owing to degradation being an irreversible process.
- Features should have good anti-interference properties against random noise.

In this study, correlation, monotony, and robustness [16] are taken as the feature evaluation indices for the adaptive feature selection process. With performance degradation being a random process, good degradation features can be broken down into the actual trend and the residual parts. The exponential weighted moving average method is used to

decompose the features into the stationary trend $f_T(t_k)$ and the random margin term $f_R(t_k)$:

$$f(t_k) = f_T(t_k) + f_R(t_k) \tag{2}$$

where $f(t_k)$ represents the feature value obtained at the time t_k, and $k = 1, 2, \ldots, K$ represents the instance of time.

Correlation, monotony, and robustness are calculated as shown in Equations (3)–(5). *Corr* represents the linearity between the measured feature and time; *Mon* represents the monotonous trend in the evaluated feature; and *Rob* represents the tolerance of the feature to outliers. The values of the aforementioned features evaluation indices range from [0,1]; the larger the index value, the better the performance of the feature.

$$Corr(F,T) = \frac{\left|K\sum_k f_T(t_k)t_k - \sum_k f_T(t_k)\cdot\sum_k t_k\right|}{\sqrt{\left[K\sum_k {f_T}^2(t_k) - \left(\sum_k f_T(t_k)\right)^2\right]\left[K\sum_k (t_k)^2 - \left(\sum_k t_k\right)^2\right]}} \tag{3}$$

$$Mon(F) = \frac{1}{K-1}\left|\sum_K \delta(f_T(k+1) - f_T(k)) - \sum_K \delta(f_T(k) - f_T(k+1))\right| \tag{4}$$

$$Rob(F) = \frac{1}{K}\sum_k \exp\left(-\left|\frac{f_R(k)}{f(k)}\right|\right) \tag{5}$$

where $f_T(t_k)$ represents the stationary trend, $f_R(t_k)$ represents the random margin term, $f(t_k)$ represents the feature value obtained at the time t_k, $k = 1, 2, \ldots, K$ represents the instance of time, and δ is a unit step function.

(II) Feature selection

Individual indices can only provide a one-sided measurement of the health of alternative features. In order to compute the three evaluation indices of each alternative feature, a linear combination can be constructed as the final comprehensive evaluation criterion

$$\underset{F \in \Omega}{J} = \omega_c CorrF, T + \omega_m Mon(F) + \omega_r Rob(F) \tag{6}$$

where J belongs to [0,1] represents a comprehensive criterion, Ω represents a set of alternative features, and ω_i represents the weight of each performance evaluation metric. The features with high J-values should be retained in order to effectively predict the RUL.

For the selection of weights in Equation (6), we used the dispersion coefficient (CV) to evaluate the degree of dispersion of the three features indices

$$CV = \frac{\sigma}{\overline{X}} \tag{7}$$

where CV is dispersion coefficient, σ is standard deviation, and $\overline{X}$ is the average value.

Using Equations (3)–(5), the weights of the three evaluation indices were calculated as

$$\omega_c = \frac{\overline{Corr}}{\overline{Corr} + \overline{Mon} + \overline{Rob}} \tag{8}$$

$$\omega_m = \frac{\overline{Mon}}{\overline{Corr} + \overline{Mon} + \overline{Rob}} \tag{9}$$

$$\omega_r = \frac{\overline{Rob}}{\overline{Corr} + \overline{Mon} + \overline{Rob}} \tag{10}$$

where ω_c, ω_m, ω_r, are the weights of *Corr*, *Mon* and *Rob* in Equation (6), and $\overline{Corr}$, $\overline{Mon}$ and $\overline{Rob}$ the mean values of the *CV*s for the *Corr*, *Mon* and *Rob*, respectively.

If the comprehensive evaluation index J of all the alternative features is sorted in decreasing order, and the first m number of features with the larger J-value are directly selected as the optimal feature subset, it may lead to feature redundancy. With increased redundancy, the RUL prediction model gradually becomes more complex with reduced prediction efficiency. Therefore, the adaptive optimal order search algorithm can be used to improve the search efficiency, and automatically determine the optimal feature subset. The algorithm screens out an optimal feature from k alternative features each time and introduces a random Gaussian white noise ε to keep the original number of features unchanged. This process reiterates until the screened optimal feature subset is nothing but the introduced Gaussian white noise (process explained in Section 2.1.3). The specific process is shown in Figure 2:

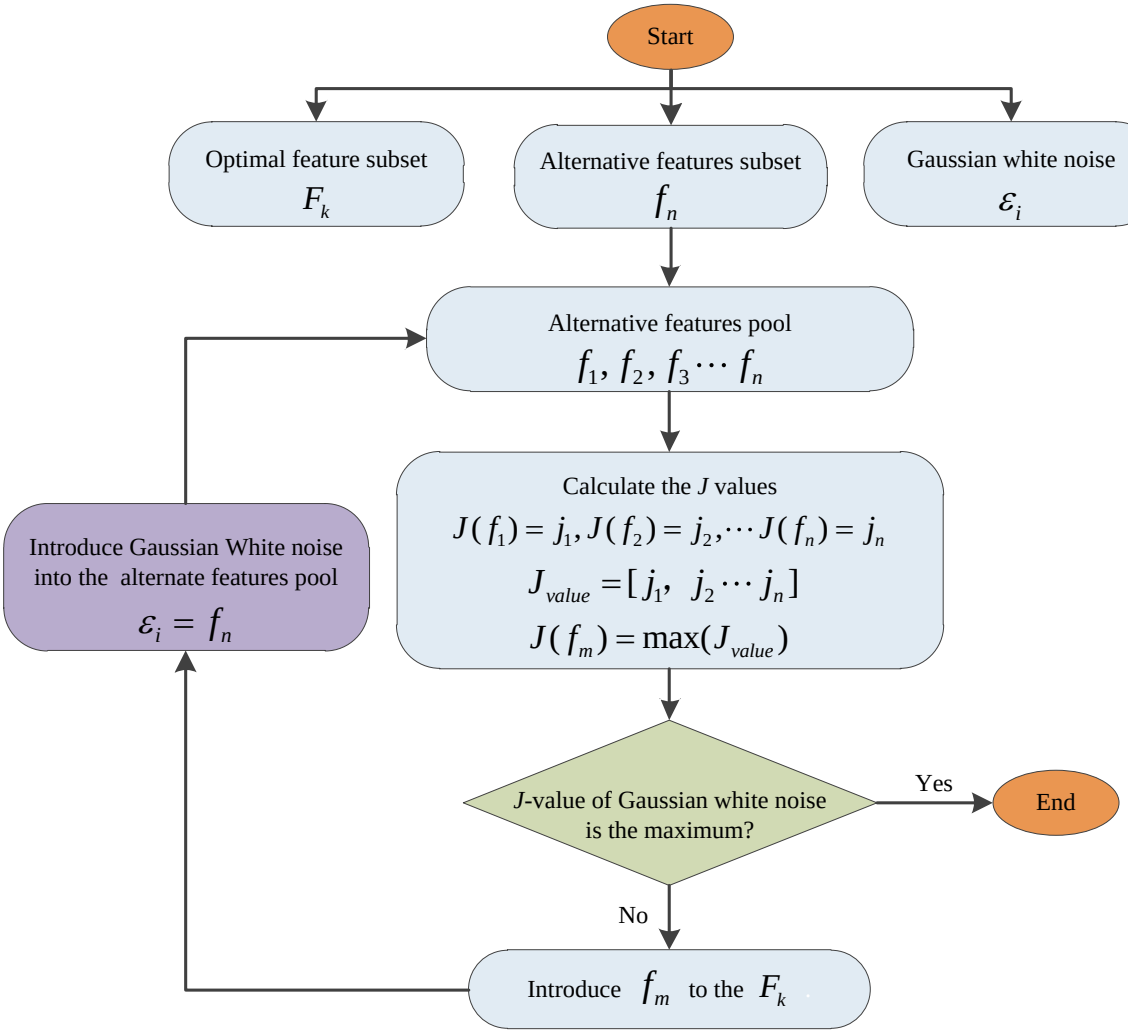


Figure 2. adaptive sequential optimal feature selection.

2.1.3. RUL Target Function

When mechanical equipment first begins to operate, due to negligible wear of its mechanical components, it is said to be in a healthy state. The RUL target function for the equipment is defined as follows:

$$RUL = \begin{cases} R, x < a - R \\ a - x, x \geq R \end{cases} \quad (11)$$

where x is the number of cycles of the measured point, a is the maximum number of cycles, and R is the critical degradation value.

2.2. Prediction Model Construction

2.2.1. Basic Theory

(I) Temporal Convolutional Network

TCN combines dilated causal convolution (DCC) and residual connections (RC) to solve timing problems [17]. Causal convolution is a time-constrained model that is unable to obtain future information. This increases the training costs while enhancing the memory of historical information on the network. On the other hand, by using dilated convolution with RC, the model enables a large, efficient receptive field. This field is able to accept more historical data while avoiding the problem caused by the very deep network. For the multidimensional features in this study, TCN can not only solve the problem that traditional convolutional neural networks (CNNs) are limited by (due to convolutional kernels), but also avoid the problem of gradient disappearance or explosion of recurrent neural networks (RNN). The RC is a constituent unit of the TCN and is illustrated in Figure 3.

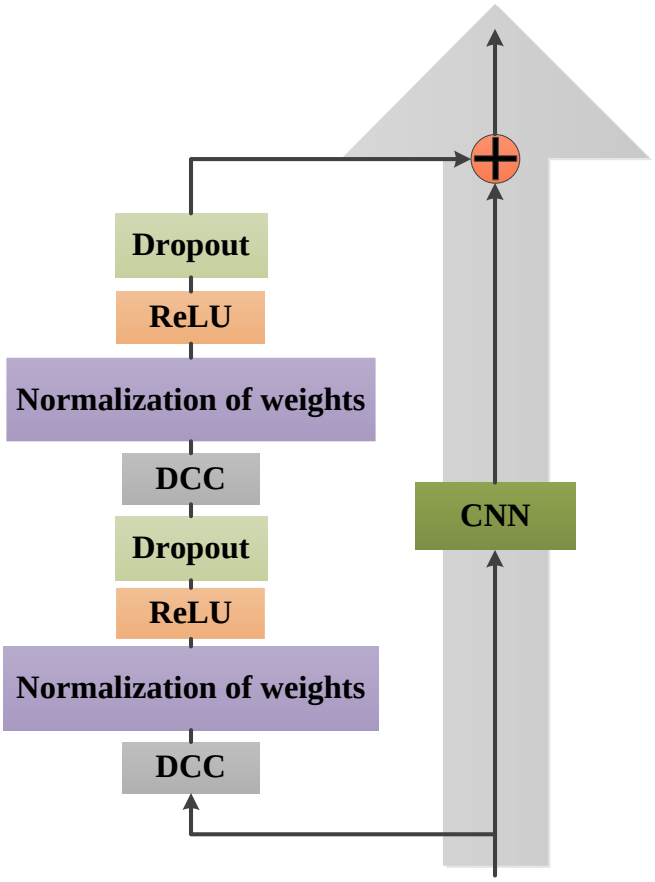


Figure 3. TCN residual block structure.

(II) Bidirectional Long Short-Term Memory

In order to extrapolate the health status of the equipment from the degradation law of the signal over time, the BILSTM algorithm is used to concatenate the hidden front and back layer vectors. This model fully considers the bidirectional information of the features, while improving the validity of time series prediction.

The BILSTM network (Figure 4) adds a reverse layer to the (LSTM) network, allowing fuller use of the effective information from the optimal feature subset. This network

combines the output of the front and back layers at each moment to obtain the final output as

$$h_t^R = f^R\left(w_1 x_t + w_2 h_{t-1}^R\right) \tag{12}$$

$$h_t^L = f^L\left(w_3 x_t + w_5 h_{t+1}^L\right) \tag{13}$$

$$h_t = f\left(w_4 h_t^R + w_6 h_t^L\right) \tag{14}$$

where x represents the input layer, h^R represents the forward layer, h^L represents the backward layer, h represents the output layer, w represents the weights, and f represents the function.

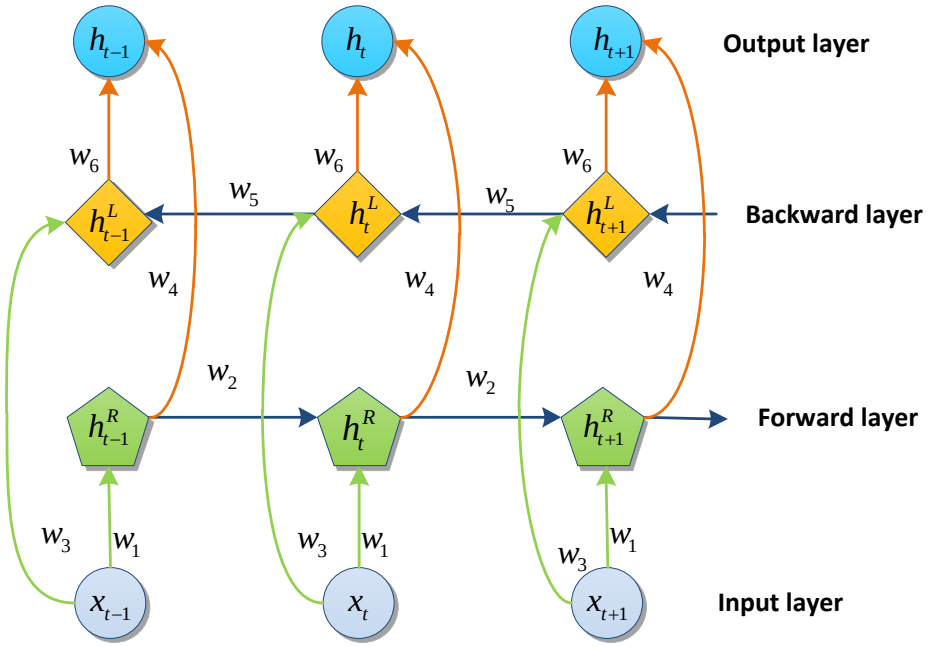


Figure 4. BILSTM loop structure diagram.

(III) Multi-head Attention

To account for the independent multidimensional equipment features, we utilized a multi-headed attention mechanism to process the different features parallelly (model structure shown in Figure 5).

The multi-head attention mechanism, based on the transformer, simultaneously attended to different parameters. The obtained results were stitched together to realize the final attention as

$$Multihead\ (Q, K, V) = Concat\ (head1, head2, \cdots, headn)W \tag{15}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{16}$$

where Q is a query matrix, K is a key matrix, and V is a values matrix, and W_i^Q, W_i^K, and W_i^V represent the weight matrices containing the weights of Q, K, and V in the ith attention head, respectively. The output of the multi-head attention mechanism is spliced by the merge layer.

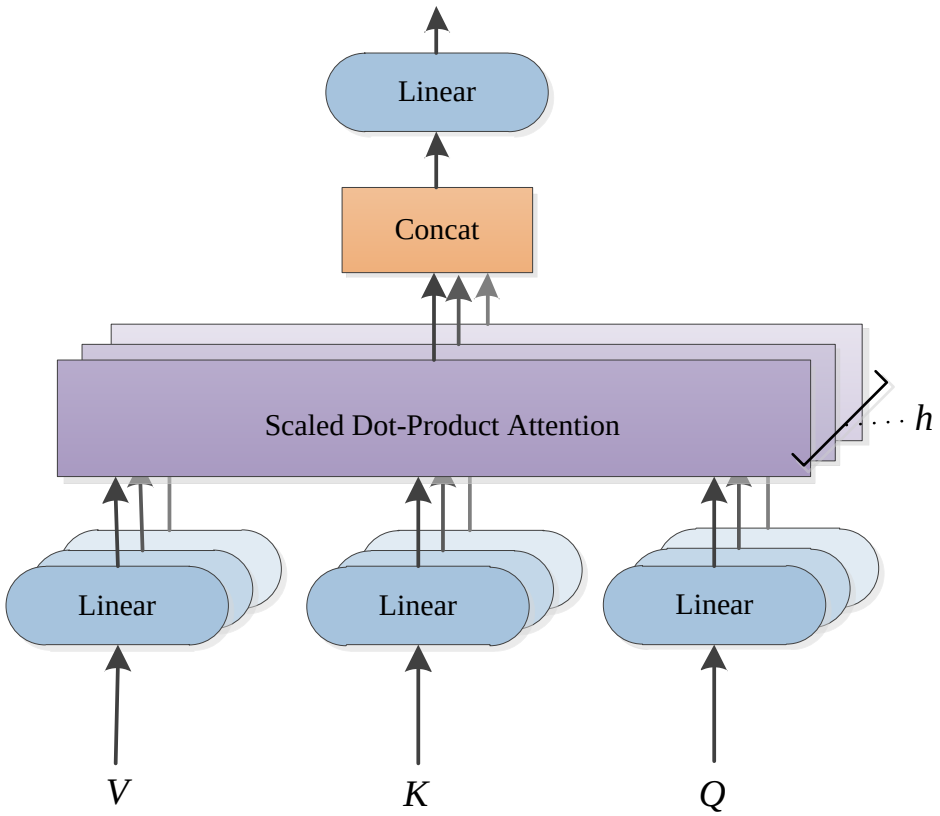


Figure 5. Multi-head attention model.

2.2.2. Metrics

In this study, we evaluate the proposed method using 3 indices, namely root mean square error (RMSE), mean absolute percentage error (MAPE), and SCORE:

- RMSE: It is a commonly used metric for evaluating prediction models in various fields, including machine learning, statistics, and engineering. It measures the differences between predicted values and actual values by computing the square root of the average squared difference between them. This metric provides a way to quantify the magnitude of the errors in the predictions and can be used to compare the performance of different prediction models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{17}$$

where y_i is the true value and $\hat{y}_i$ is the predicted value.

- MAPE: It is another commonly used evaluation metric for predicting RUL. MAPE measures the percentage difference between predicted values and actual values, which makes it useful for assessing the accuracy of predictions when the scale of the data varies widely.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \tag{18}$$

- SCORE: Early prediction is often more important and effective than later prediction for gradually degrading equipment, such as aircraft engines, bearings, etc., which experience gradual deterioration within their operational life cycle, and their failures typically develop gradually, causing progressive damage to the equipment over a period of time. By setting a penalty for later predictions compared to early predictions,

the score function can better capture the preference for early predictions. This is particularly useful for capturing the early warning signs of equipment degradation and preventing catastrophic failures.

$$score = \begin{cases} \sum_{i=1}^{n} \left(e^{-\frac{d_i}{13}} - 1 \right), d_i < 0 \\ \sum_{i=1}^{n} \left(e^{\frac{d_i}{10}} - 1 \right), d_i \geq 0 \end{cases} \tag{19}$$

$$d_i = R\hat{U}L_i - RUL_i \tag{20}$$

where $R\hat{U}L_i$ is the RUL prediction value at moment i and RUL_i is the RUL true value at moment i.

2.2.3. Proposed Model

In order to make full use of the optimal feature subset, we propose a multichannel prediction model of TCN-BILSTM based on the multi-head attention mechanism. This model introduces TCN to improve the computational efficiency of the network, while ensuring the integrity of the long-term sequence. Further, BILSTM is used to extract bidirectional sequence information, and achieve parallel processing of multiple features through the multichannel network structure. The specific structure is shown in the following Figure 6. The detailed model parameters are listed in Table 1.

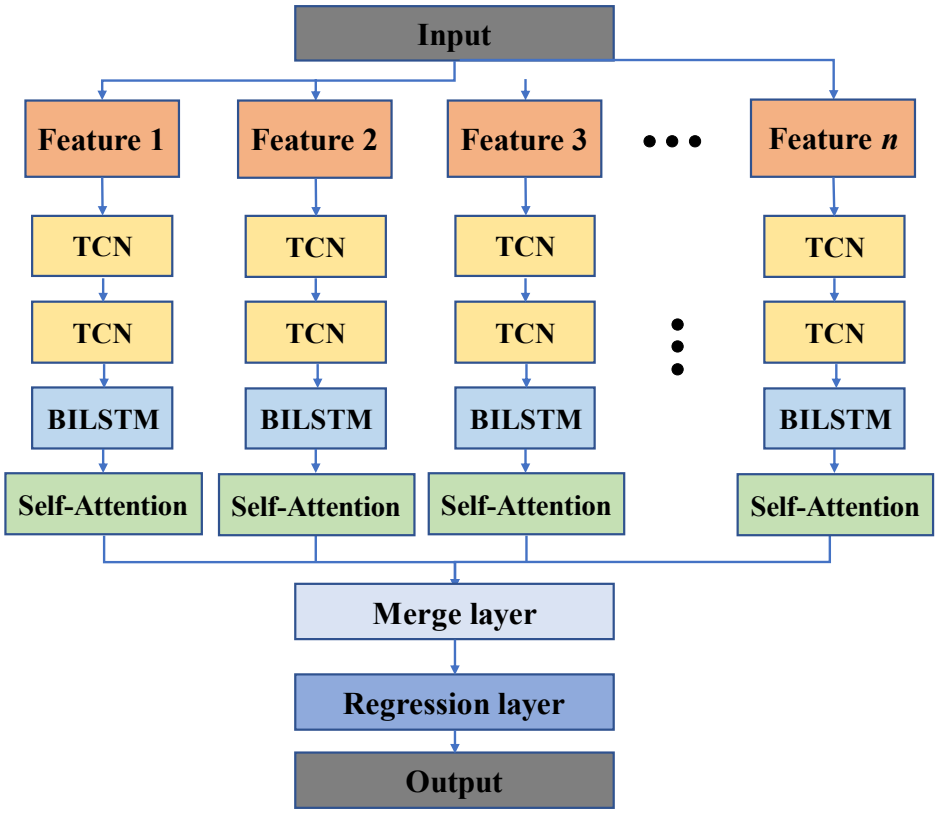


Figure 6. Network structure based on multi-head attention and TCN-BILSTM.

Table 1. Network parameters based on TCN-BILSTM and multi-head attention.

Type	Definition
Input Layer	The input layer
TCN	filters = 32, kernel size = 3
Batch Norm	Batch normalization
TCN	filters = 32, kernel size = 3
Batch Norm	Batch normalization
Bidirectional(LSTM)	Units = 32
Batch Norm	Batch normalization
SeqSelfAttention	Self-attentional layer
MaxPooling1D	The pooling layer
Flatten	Returns a 1D array
Concatenate	Merge the channels
Dense	Dense to 50
Dense	Dense to 1

3. Case Studies

3.1. Case Study1: Intelligent Maintenance System (IMS) Bearing Dataset

3.1.1. Dataset Description

The experimental data were procured from the Rolling Bearing Life Cycle Experiment of the Intelligent Maintenance System (IMS) at the University of Cincinnati, USA [18]. In the experimental device (shown in Figure 7), the DC motor drives the rotation of the four Rexnord ZA-2115 rolling bearings on the shaft. Each bearing is installed with a 353B33 high-sensitivity quartz acceleration sensor in both radial and axial directions. A constant radial load of 26.67 KN was applied to each bearing at a constant speed of 2000 r/min. The vibration signal was measured every 10 min during the experiment. The spindle speed was kept at a constant 2000 r/min. The sampling frequency was 20 kHz, and each sample contained 20,480 data points. Therefore, in order to maintain the internal information integrity of a single training sample, while greatly reducing the amount of computation, each sensor trains with an initial set of 10,240 data points, with the final set of 10,240 points being the actual test data. This dataset contains the sub-datasets of the three experiments (specific bearing data shown in Table 2).

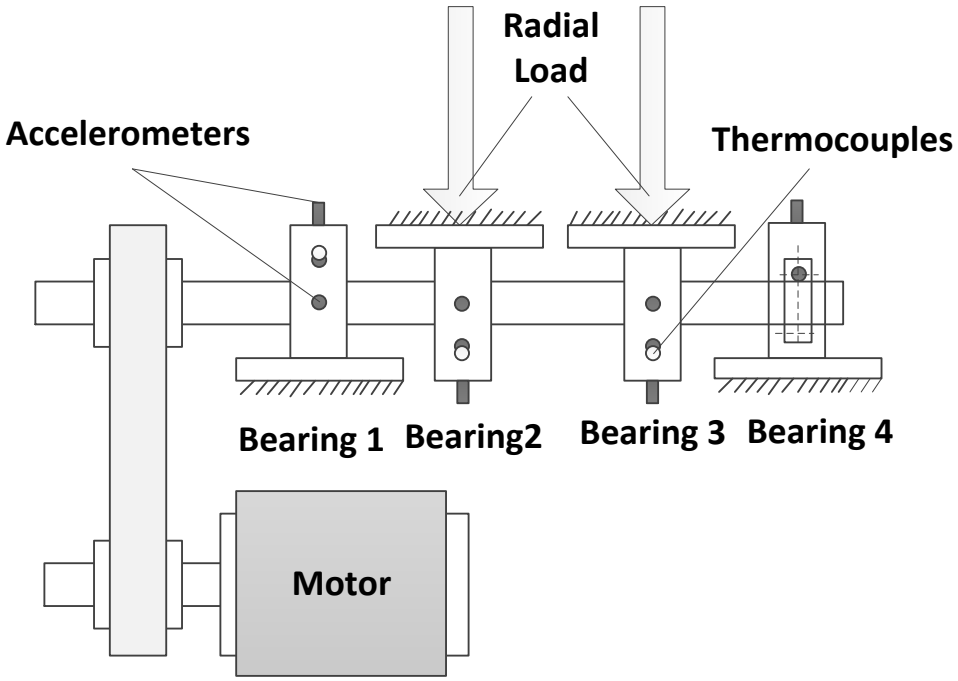


Figure 7. Experimental device diagram of bearing life cycle.

Table 2. Description of IMS datasets.

Test Number	File Number	Fault Bearing	Symbol	Fault Bearing
Test 1	2156	Bearing 3	Br1-3	inner race
Test 2	984	Bearing 1	Br2-1	outer race
Test 3	6324	Bearing 3	Br3-3	outer race

3.1.2. Data Preparation

In order to find the degradation law for the bearings and to improve the RUL prediction accuracy, a variety of features were extracted on the basis of the original bearing vibration signal. Fourteen features including the maximum, standard deviation, and root mean square (RMS) values were extracted from the time domain range. Simultaneously, four features, namely center of gravity frequency, average frequency, RMS frequency, and frequency standard deviation, were extracted from the frequency domain range. As shown in Figure 8, using the adaptive feature selection method, these 18-dimensional features were extracted for three types of bearings: Br1-3, Br2-1, and Br3-3.

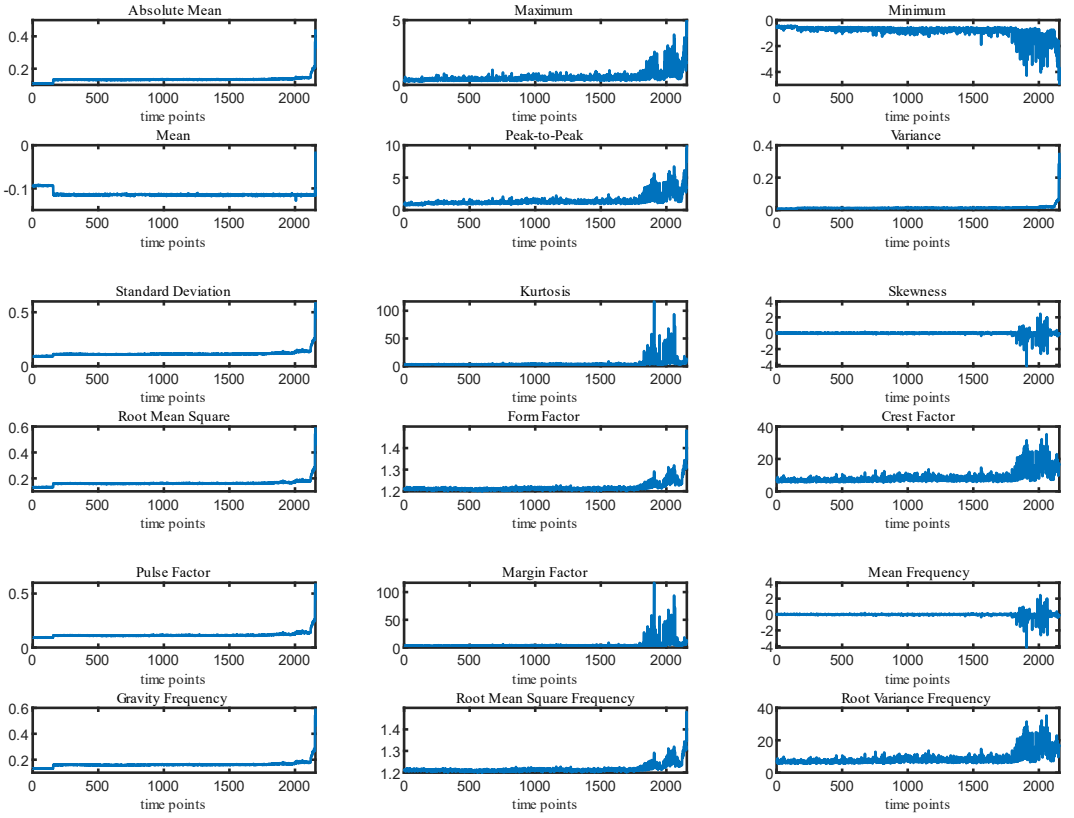
The value of α in the exponential smoothing method determines the degree of smoothing. The larger the α, the greater the influence of the recent data on the prediction results; the smaller the α, the smoother the data. When the data remain unchanged in the long term, α can be taken within 0.1 to 0.5 [19]. We selected α-values of 0.1, 0.3, and 0.5 to treat the 18-dimensional features of the three bearings separately. A random feature of each bearing was selected to demonstrate the exponential smoothing (Figure 9). For all three bearings, the effect of the different α-values on the RUL prediction results are compared and displayed in Figure 10.

According to several studies, the early degradation points of Br1-3, Br2-1, and Br3-3 were determined to be 2015, 536, and 6074 [20], respectively. The critical degradation values of R were found to be 142, 449 and 251, respectively. Figure 10 and Table 3 display the significant impact of varying the α-value on the RUL predictions.

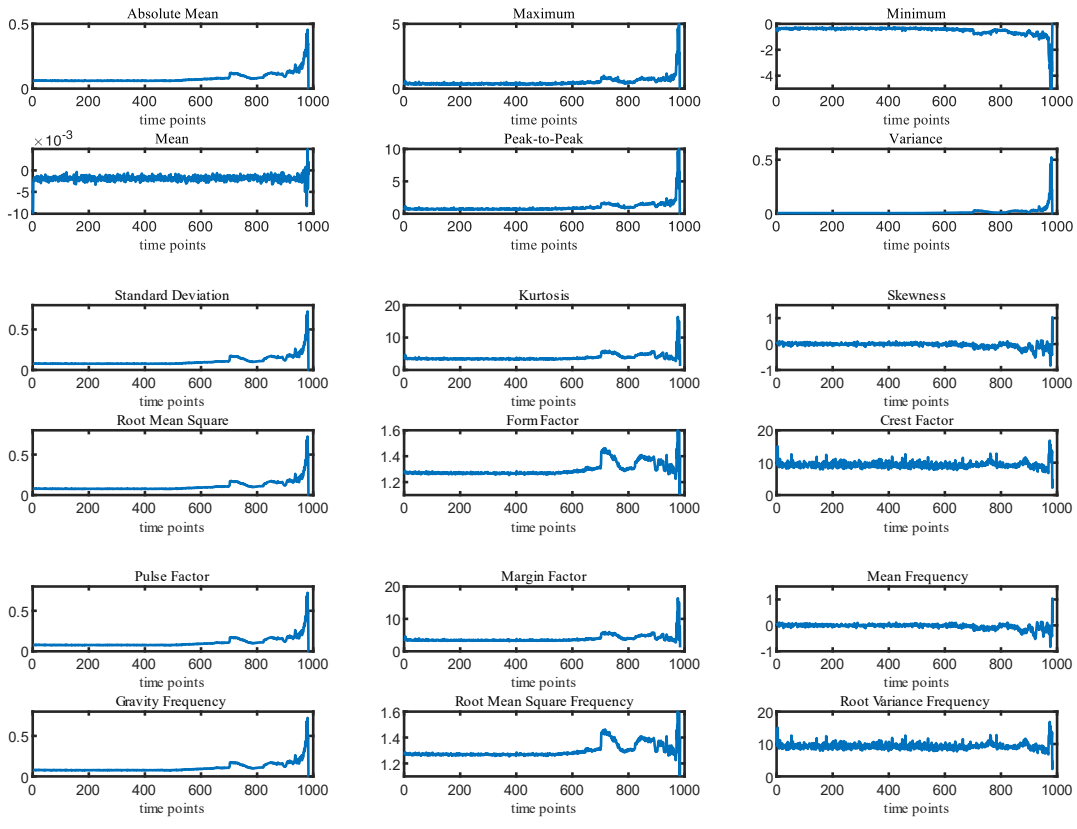
As shown in Figure 10, the use of different smoothing constants has varying effects on the RUL prediction of bearings. The selection of smoothing constant significantly impacts the prediction performance of Br1-3. When α = 0.1, a considerable amount of noise is removed, but this also affects the network's prediction accuracy, suggesting that excessive noise removal can reduce the network's robustness and, consequently, the prediction accuracy. On the other hand, when α = 0.5, excessive redundant noise is retained, leading to a decrease in the prediction accuracy. However, when α = 0.3, RMSE, SCORE, and MAPE were optimized by 43.23%, 73.69%, and 57.14%, respectively.

For the prediction results of Br2-1, the original data contain a large number of environmental interference factors, and the introduction of exponential smoothing can effectively improve the prediction performance. The optimal prediction performance is achieved when α = 0.3, with RMSE, SCORE, and MAPE optimized by 73.59%, 97.19%, and 58.33%, respectively.

Regarding the prediction results of Br3-3, it can be observed that the original data perform well in predicting the bearing in a healthy state, while the prediction performance of the bearing in a degraded state is poor. However, the introduction of exponential smoothing significantly improves the prediction performance for the bearing in the degraded state. The prediction performance for the Br3-3 in the healthy state decreases, whether α = 0.1 or α = 0.5. Nonetheless, the optimal prediction performance is achieved when α = 0.3, with RMSE, SCORE, and MAPE optimized by 57.71%, 99.99%, and 70.57%, respectively.

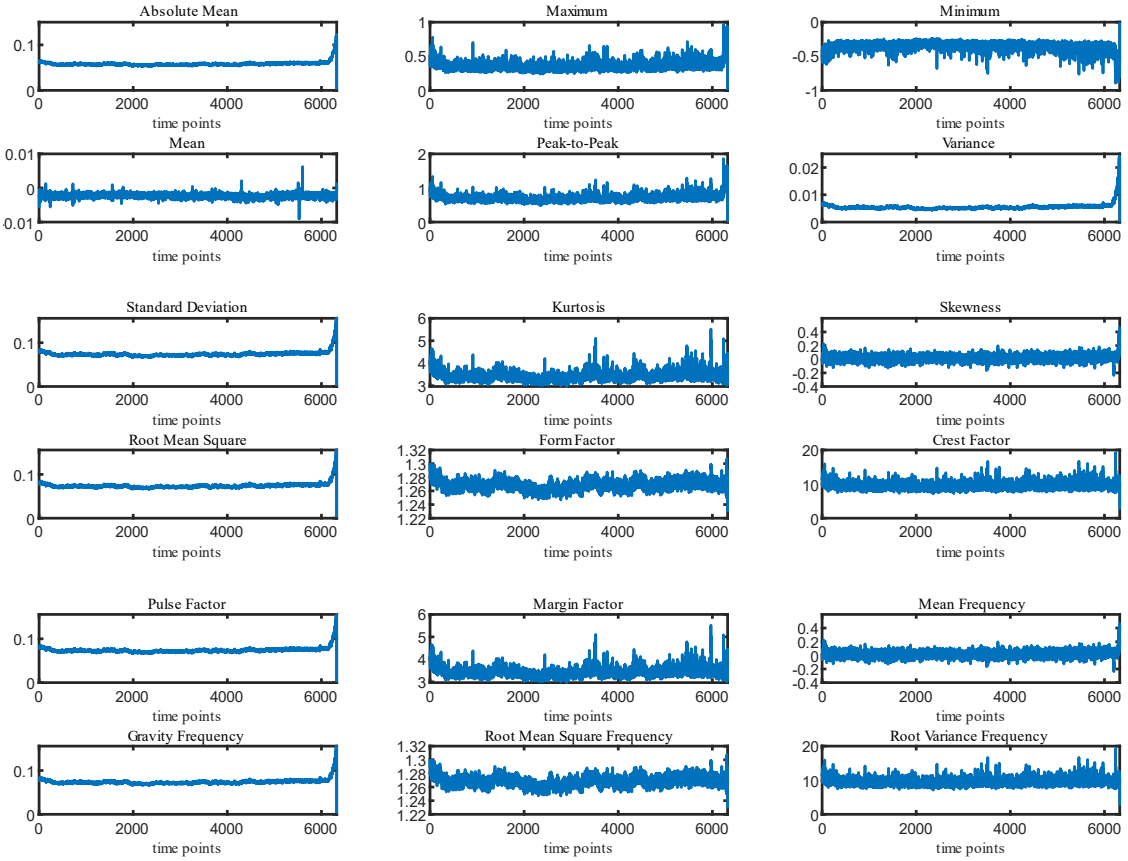


(**a**)



(**b**)

Figure 8. *Cont.*



(**c**)

Figure 8. (**a**) Feature extraction results of Br1-3; (**b**) feature extraction results of Br2-1; (**c**) feature extraction results of r3-3.

3.1.3. Optimal Feature Subset

The *Corr*, *Mon*, and *Rob* of the 18-dimensional features of the three bearings were obtained using Equations (3)–(5), and the weights of the three evaluation indices were calculated to be $\omega_c = 0.15$, $\omega_m = 0.64$, and $\omega_r = 0.21$, using Equations (8)–(10) (Table 4).

Based on the adaptive selection method explained in Section 2.1.2, using Equation (6), the 18-dimensional features of the three bearings were screened down to an 11-dimensional optimal feature subset. To test the validity of this subset, we compared the effects of the screened 11-dimensional features (subset 1), original 18-dimensional features (subset 2), and the first 11-dimensional features with a large *J*-value (subset 3), on their bearing RUL predictions (Figure 11 and Table 5).

From Figure 11, it can be seen that subset 1 shows a better RUL prediction for the three bearings, displaying the highest and lowest impact on the prediction for bearings Br2-1 and Br1-3, respectively. The quantitative analysis (Table 6) shows a superiority in the prediction performance of subset 1 over that of subsets 2 and 3 for the three evaluation indices. This indicates that the adaptive feature selection method can efficiently screen the features of the bearings, and the optimal feature subset can effectively reduce prediction errors.

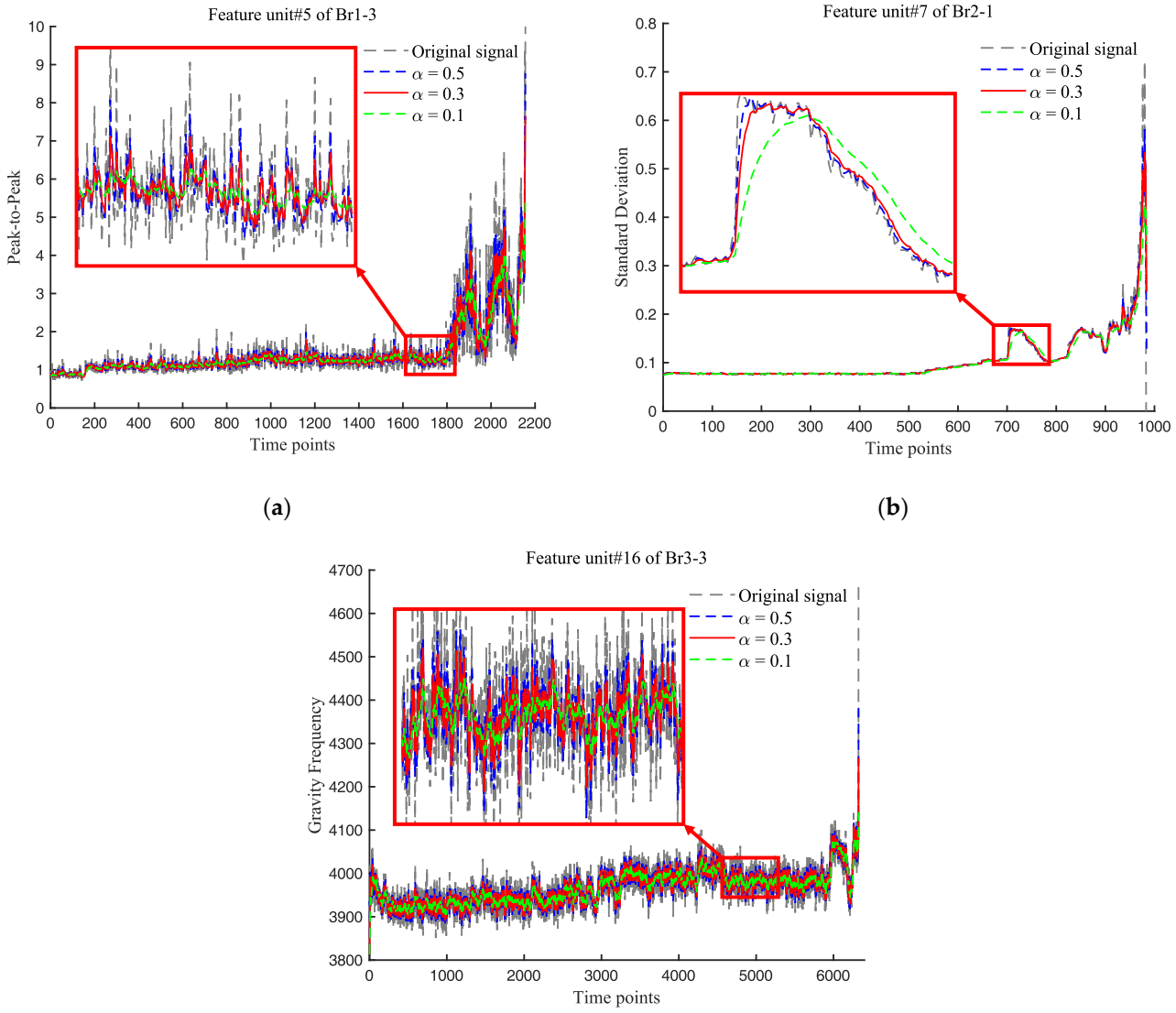


Figure 9. (**a**) Performance of ES on peak-to-peak of Br1-3; (**b**) performance of ES on standard deviation of Br2-1; (**c**) 3; (**b**) performance of ES on gravity frequency of Br3-3.

3.1.4. Discussion and Comparison

In order to verify the superiority of our model, we constructed several commonly used deep learning models: CNN, TCN, LSTM, BILSTM, BIGRU, CNN-LSTM, TCN-BIGRU, and TCN-BILSTM. To validate the comparison, the optimal prediction results of each model were selected, implying possible variations in the datasets of the different models. The prediction results of the different models are shown in Figure 12, and the error analysis is displayed in Table 7.

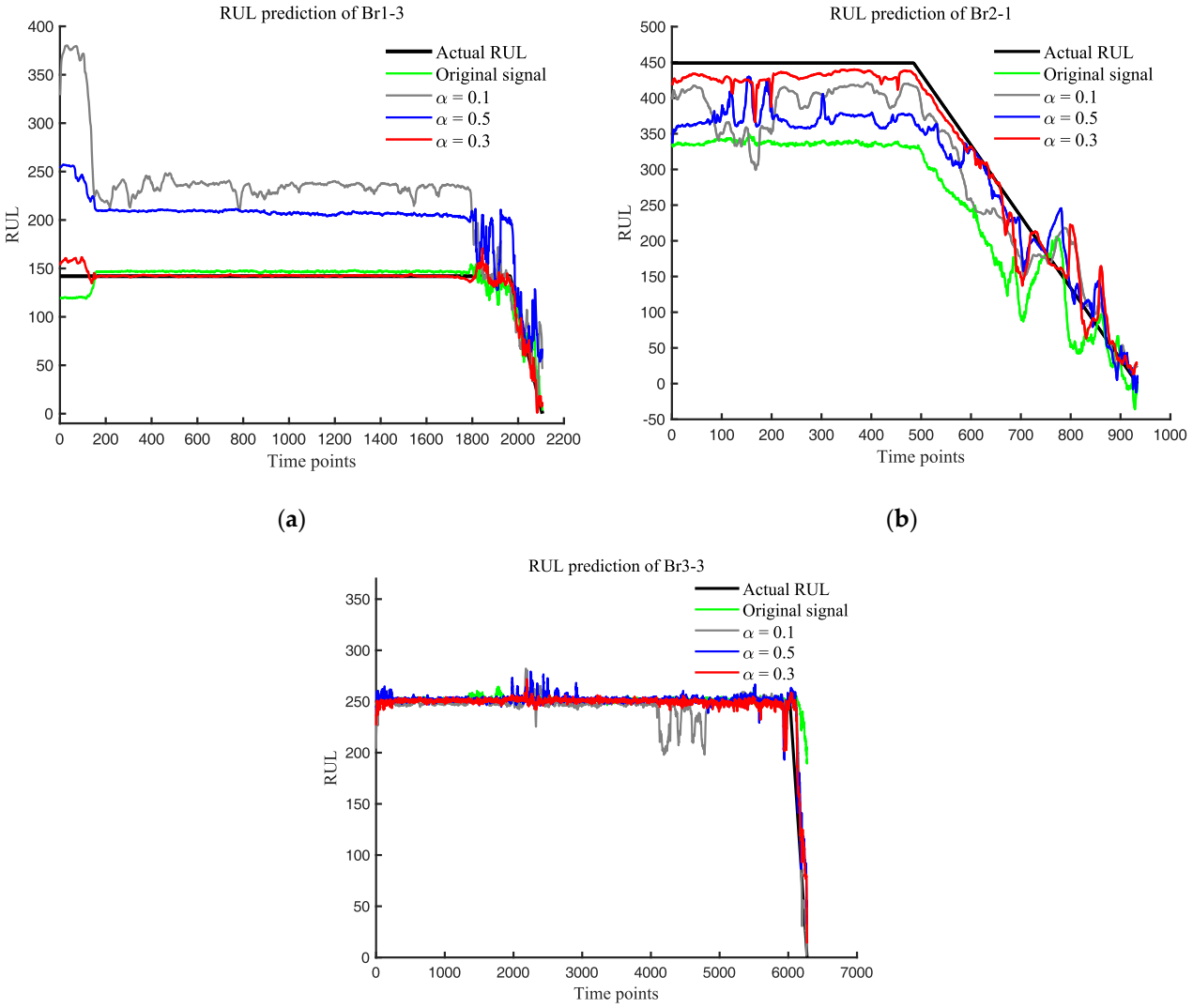


Figure 10. (**a**) RUL prediction of different α on Br1-3; (**b**) RUL prediction of different α on Br2-1; (**c**) RUL prediction of different α on Br3-3.

Table 3. Evaluation indexes for different α.

Bearings	α	RMSE	SCORE	MAPE
Br1-3	0	9.30	3.55×10^3	0.07
	0.1	99.85	1.42×10^{12}	0.74
	0.3	5.28	9.34×10^2	0.03
	0.5	66.28	8.45×10^6	0.56
Br2-1	0	99.13	4.34×10^6	0.36
	0.1	56.41	1.27×10^6	0.23
	0.3	26.18	1.22×10^5	0.15
	0.5	62.36	3.46×10^5	0.19
Br3-3	0	25.61	8.01×10^9	0.17
	0.1	14.41	2.40×10^5	0.04
	0.3	10.83	1.23×10^5	0.05
	0.5	11.45	1.65×10^5	0.05

Table 4. Dispersion Coefficient.

Bearings	*Corr*	*Mon*	*Rob*
Br1-3	8.46	39.34	8.28
Br2-1	5.33	13.11	9.31
Br3-3	5.17	29.93	10.16

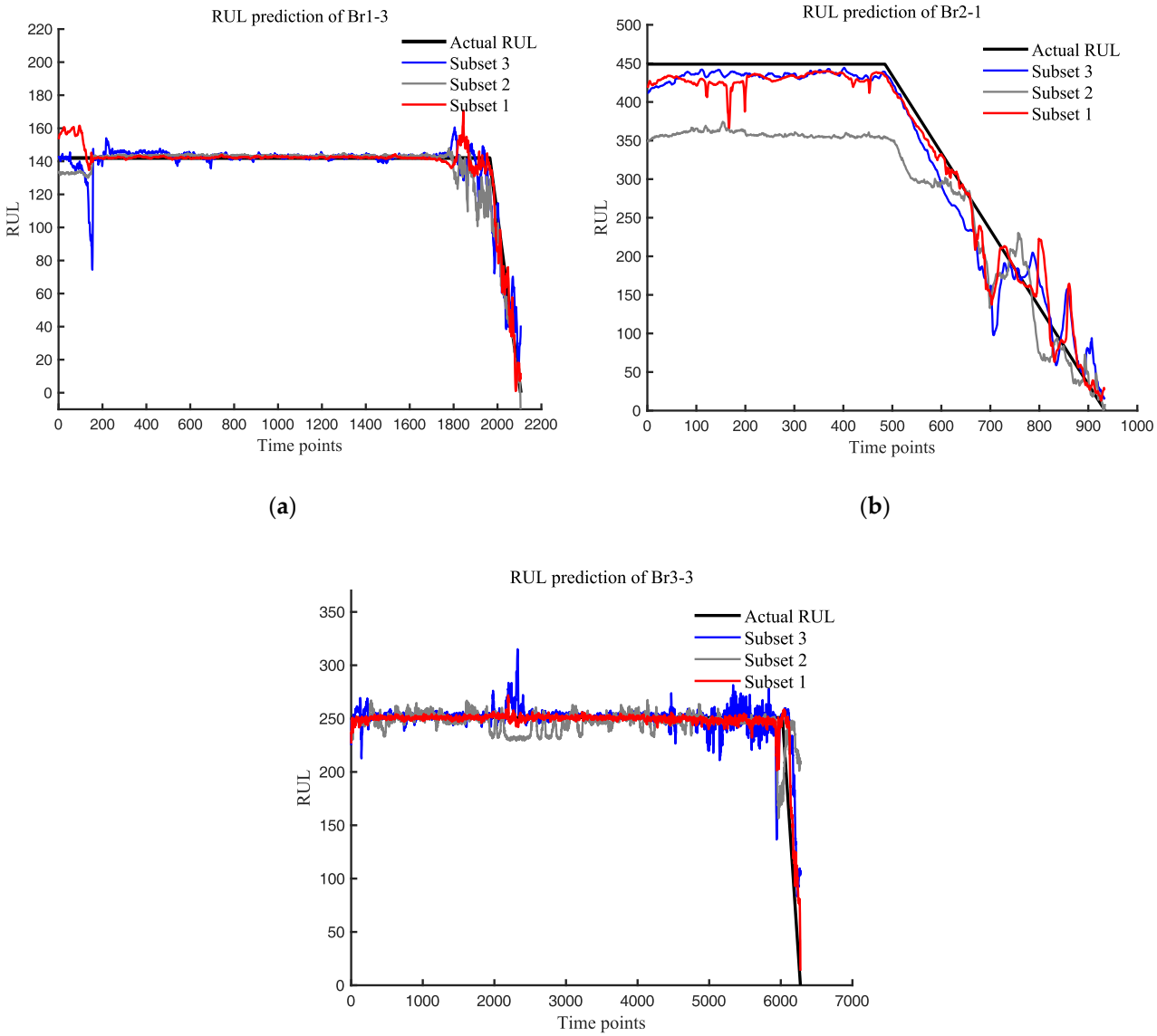


Figure 11. (**a**) RUL prediction of different feature subsets on Br1-3; (**b**) RUL prediction of different feature subsets on Br2-1; (**c**) RUL prediction of different feature subsets on Br3-3.

Table 5. Different feature subsets.

Feature Subsets	Features
Subset 1	1,10,4,15,6,7,5,8,2,3,9
Subset 2	1–18
Subset 3	1,10,7,15,16,18,6,3,17,5,14

Table 6. Evaluation indexes for different feature subsets.

Bearings	Subset	RMSE	SCORE	MAPE
Br1-3	1	5.28	9.34×10^2	0.03
	2	7.13	1.33×10^3	0.04
	3	8.25	2.63×10^3	0.07
Br2-1	1	26.18	1.22×10^5	0.15
	2	74.07	6.44×10^5	0.21
	3	31.11	1.74×10^4	0.19
Br3-3	1	10.83	1.23×10^5	0.05
	2	27.55	1.34×10^{10}	0.19
	3	17.85	3.38×10^6	0.10

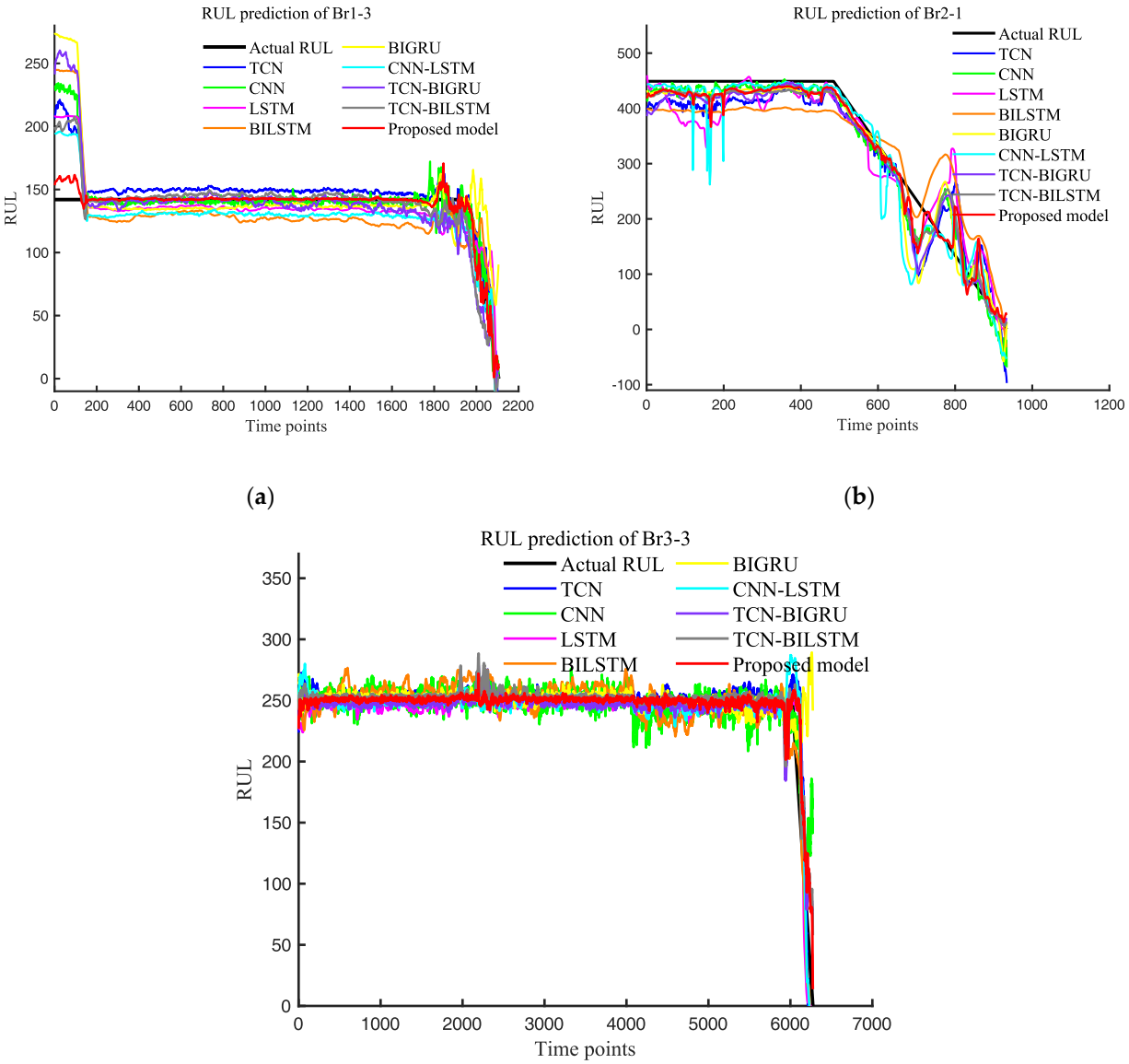


Figure 12. (**a**) RUL prediction of different feature models on Br1-3; (**b**) RUL prediction of different feature models on Br2-1; (**c**) RUL prediction of different feature models on Br3-3.

Table 7. Evaluation indexes for different models.

Bearings	Model	RMSE	SCORE	MAPE
Br1-3	TCN	17.51	1.17×10^5	0.11
	CNN	21.06	6.61×10^5	0.07
	LSTM	20.06	9.68×10^4	0.12
	BILSTM	30.05	3.03×10^6	0.23
	BIGRU	34.24	4.06×10^7	0.26
	CNN-LSTM	18.47	2.54×10^4	0.19
	TCN-BIGRU	28.27	7.41×10^6	0.11
	TCN-BILSTM	16.93	5.12×10^4	0.08
	Proposed model	5.28	9.34×10^2	0.03
Br2-1	TCN	45.63	5.24×10^6	0.42
	CNN	28.48	1.07×10^5	0.25
	LSTM	52.98	1.28×10^9	0.22
	BILSTM	61.66	1.98×10^8	0.36
	BIGRU	40.48	1.91×10^6	0.22
	CNN-LSTM	44.93	9.44×10^6	0.34
	TCN-BIGRU	36.31	3.53×10^5	0.14
	TCN-BILSTM	29.62	4.16×10^5	0.12
	Proposed model	26.18	1.22×10^5	0.15
Br3-3	TCN	20.59	5.81×10^9	0.19
	CNN	16.06	2.51×10^8	0.13
	LSTM	11.64	8.09×10^4	0.04
	BILSTM	13.24	1.56×10^4	0.08
	BIGRU	29.75	1.18×10^{13}	0.22
	CNN-LSTM	12.49	2.48×10^5	0.06
	TCN-BIGRU	12.06	1.00×10^5	0.07
	TCN-BILSTM	11.66	1.28×10^5	0.07
	Proposed model	10.83	1.23×10^5	0.05

Based on the prediction results, the proposed model can accurately predict the remaining useful life (RUL) of the three bearings. Among them, the prediction performance of Br1-3 is the best, and all three evaluation indices show good performance. For Br3-3, it can be accurately predicted in both healthy and degraded stages, and the prediction performance is good. As shown in Table 2, the life cycle of Br2-1 is the shortest among the three bearings, which means that compared with the other bearings, there are less full-life cycle data available for training on Br2-1. Therefore, the model has learned less relevant information on Br2-1, resulting in relatively poor prediction performance for Br2-1.

3.2. Case Study2: C-MAPSS Aeroengines Dataset

3.2.1. Dataset Description

In this section, we discuss the use of NASA's C-MAPSS dataset to generalize the prediction effect of our proposed method for aeroengines. The dataset contains four sub-datasets, each containing the training set, the test set, and the RUL true value. Both the training and test datasets contain 21-dimensional feature parameters. For each engine, the process goes through the initial healthy state to the final failure state. The test dataset contains 100 engines that run a certain number of cycles before they fail [21,22]. The work in this section is mainly based on the dataset under the FD001 operating conditions (shown in Tables 8 and 9).

Table 8. FD001 data description.

Dataset	FD001	
	Training Set	Testing Set
Engines	100	100
Sensor measurements	21	21
Operation conditions	H = 0 kft Ma = 0 TRA = 100°	
Fault modes	Fault of high-pressure compressor	

Table 9. C-MAPSS outputs to measure system response.

No.	Symbol	Description	Units
1	T2	Total temperature at fan inlet	(°)
2	T24	Total temperature at LPC outlet	(°)
3	T30	Total temperature at HPC outlet	(°)
4	T50	Total temperature at LPT outlet	(°)
5	P2	Pressure at fan inlet	Pa
6	P15	Total pressure in bypass-duct	Pa
7	P30	Total pressure at HPC outlet	Pa
8	Nf	Physical fan speed	r/min
9	Nc	Physical core speed	r/min
10	epr	Engine pressure ratio (P50/P2)	-
11	Ps30	Static pressure at HPC outlet	Pa
12	Phi	Ratio of fuel flow to Ps30	pps/psi
13	NRf	Corrected fan speed	r/min
14	NRc	Corrected core speed	r/min
15	BPR	Bypass ratio	-
16	FarB	Burner fuel–air ratio	-
17	htBleed	Bleed enthalpy	-
18	Nf_dmd	Demanded fan speed	r/min
19	PCNfR_dmd	Demanded corrected fan speed	r/min
20	W31	HPT coolant bleed	lbm/s
21	W32	LPT coolant bleed	lbm/s

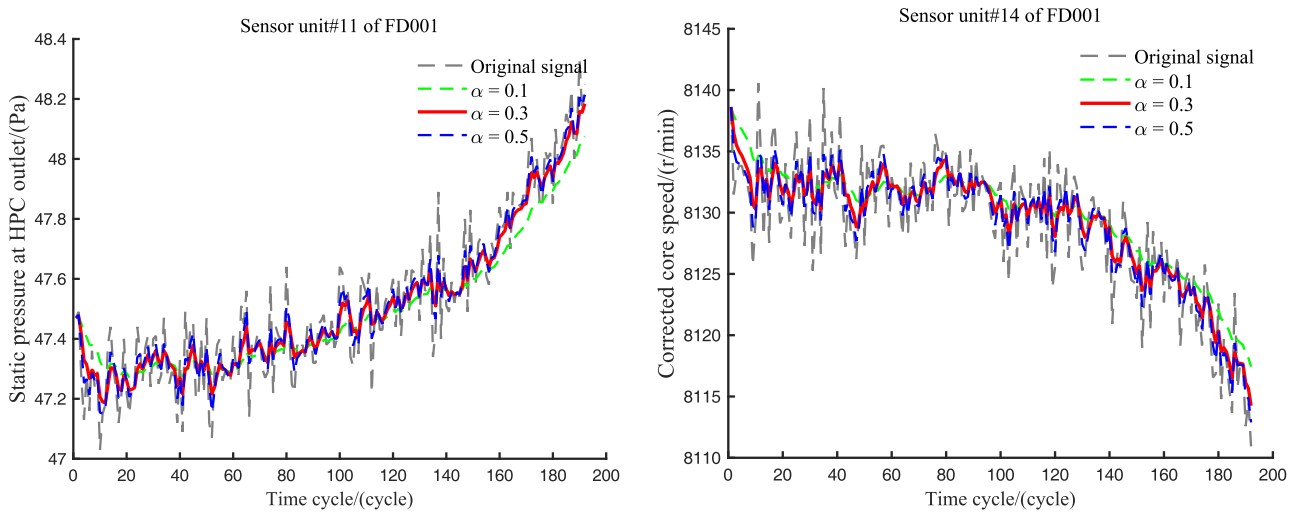
3.2.2. Data Preparation

We chose 21 sensor detection features for the 100 engines. These were processed with exponential smoothing coefficient (α-values of 0.1, 0.3, and 0.5, and two sensor detection datasets were randomly selected for plotting. The processed data are shown in Figure 13.

From the above figure, we can observe that when (i) α = 0.1, it is not possible to perform a good trend fitting in the later periods of the cycle; (ii) α = 0.5, the environmental noise is not completely eliminated; and α = 0.3, the environmental noise interference is avoided to the highest extent, while ensuring a good fit of the degradation curve of the cycle. Therefore, when α = 0.3, we have the most optimal smoothing effect on the sensor detection features of the aeroengine.

3.2.3. Optimal Feature Subset

Based on Equations (3)–(5) and (8)–(10), the weights of ω_c, $\omega_{m,}$ and ω_r of the C-MAPSS data were calculated as 0.29, 0.43, and 0.28, respectively. The 21-dimensional sensor detection features of the C-MAPSS data were filtered using the adaptive feature selection method (Section 2.1.2), and finally, an optimal feature subset consisting of 12-dimensional features was obtained. As shown in Table 10, in order to verify the effectiveness of the optimal feature subset, two other feature subsets were constructed (in the same manner as Section 3.1.3) for the comparative testing of aeroengine RUL prediction (Figure 14).



(**a**) (**b**)

Figure 13. (**a**,**b**) Performance of ES on sensors #11 and #14 of FD001.

Table 10. Different feature subsets.

Feature Subsets	Features
Subset 1	2,3,4,7,8,9,11,12,13,15,20,21
Subset 2	4,7,8,9,11,12,13,14,15,17,20,21
Subset 3	1–21

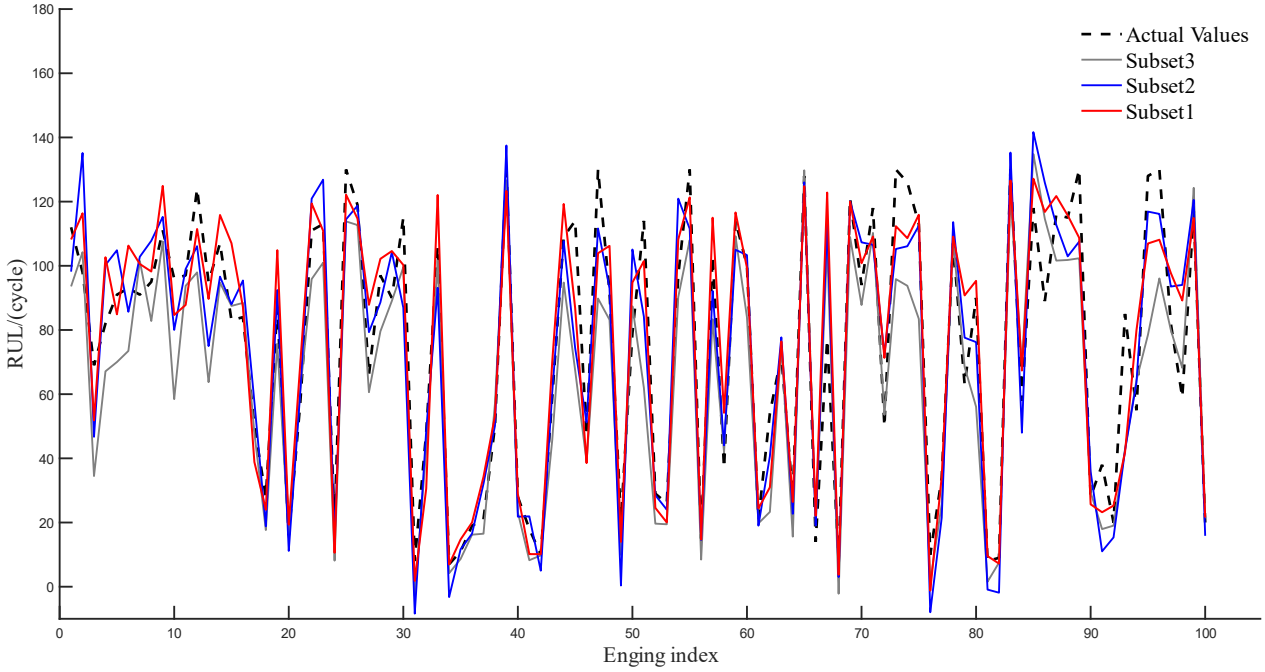


Figure 14. RUL prediction of different feature subsets.

From Figure 14, we can observe that subset 1 achieves the most accurate prediction of the aeroengine RUL. The comparison with the other subsets is based on two commonly used evaluation indices of C-MAPSS, i.e., RMSE and SCORE. The prediction effect of subset 1 is seen to be significantly better than the other two subsets on the two evaluation indices (Table 11).

Table 11. Evaluation indexes for different feature subsets.

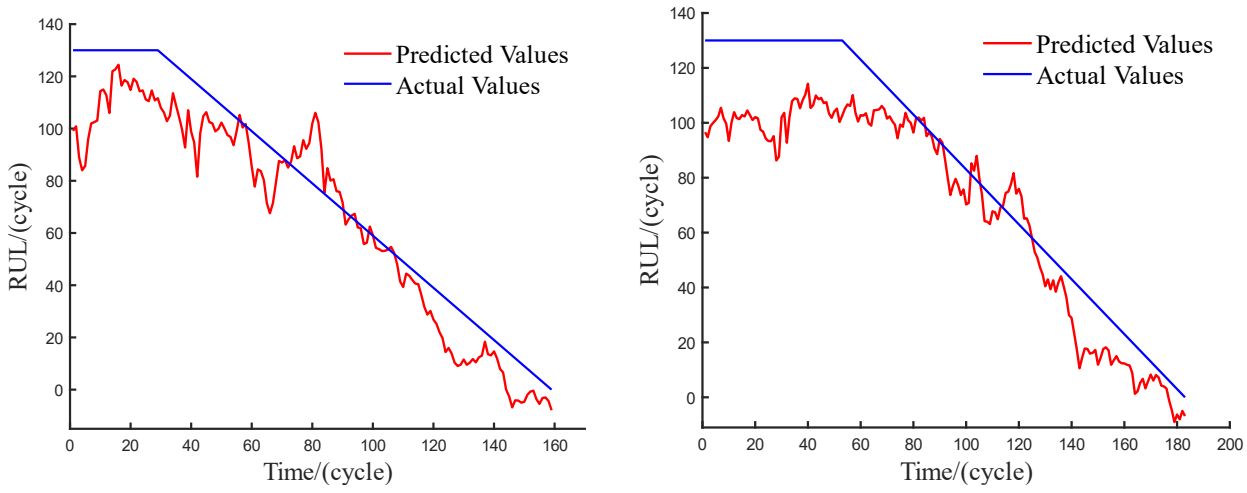
Subset	RMSE	SCORE
1	13.99	313
2	15.42	411
3	17.85	969

3.2.4. Discussion and Comparison

The NASA C-MAPSS data are a widely used public dataset for RUL prediction, by which many researchers achieved exciting results. Thakkar et al. [23] developed a deep layer recurrent neural network (DL-RNN) model to predict the RUL of aeroengines and the results showed this DL-RNN model achieved high prediction accuracy. As to our model, it aims to realize relatively satisfactory RUL prediction for various types of gradually degrading equipment, with a focus on enhancing the generalization and versatility of the prediction model. In order to verify the effectiveness of our proposed method for the RUL prediction of aeroengines, we compared it with the mainstream machine learning methods (Table 12). Our method improves on and vastly outperforms other methods tested on the C-MAPSS dataset, with both RMSE and SCORE values reduced by 6%. The RUL prediction comparison results are visually displayed by choosing 3 engines randomly out of the set of 100 engines (Figure 15).

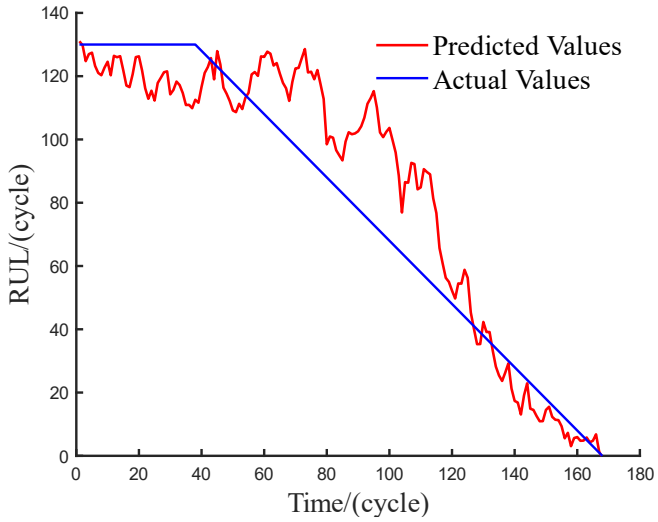
Table 12. Evaluation indexes for different models.

Model	RMSE	SCORE
LSTMBS [24]	14.89	481
DBN [25]	15.04	334
LSTM [26]	16.14	338
CNN [27]	18.45	1286
RVM [28]	23.80	10,502
Proposed model	13.99	313



(**a**) (**b**)

Figure 15. *Cont.*



(**c**)

Figure 15. (**a**) RUL prediction results for engine units #52 of FD001; (**b**) RUL prediction results for engine units #81 of FD001; (**c**) RUL prediction results for engine units #100 of FD001.

4. Conclusions

Based on the multi-head attention mechanism, a novel RUL prediction method for gradually degrading equipment was proposed, which uses discrete coefficients to construct evaluation indices for obtaining the optimal feature subset. This method effectively eliminates redundant features and improves the prediction accuracy of the model. The TCN-BILSTM network based on the multi-head attention mechanism was used to perform individual feature mining to maximize information integrity and avoid information loss. In addition, the proposed method was verified using the IMS bearing dataset and the C-MAPSS aeroengines dataset. The results indicate that the proposed method has a higher prediction accuracy and greater generalization than other machine learning methods for different types of mechanical equipment. With multidimensional features being generally used to improve the RUL prediction accuracy, our method, in particular, demonstrates great advantages in the case of such input features, providing an effective guide for future mechanical equipment maintenance.

Author Contributions: Conceptualization, L.N.; methodology, L.N. and S.X.; investigation, L.Z. and S.X.; software, S.X.; validation, L.Z.; writing—original draft preparation, S.X.; writing—review and editing, L.N., S.X. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China Program (No. 51975191).

Data Availability Statement: Not applicable.

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

References

1. Lu, Y.; Li, Q.; Liang, S.Y. Physics-based intelligent prognosis for rolling bearing with fault feature extraction. *Int. J. Adv. Manuf. Technol.* **2018**, *97*, 611–620. [CrossRef]
2. Zhao, H.; Liu, H.; Jin, Y.; Dang, X.; Deng, W. Feature Extraction for Data-Driven Remaining Useful Life Prediction of Rolling Bearings. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3511910. [CrossRef]
3. Gao, T.; Li, Y.; Huang, X.; Wang, C. Data-Driven Method for Predicting Remaining Useful Life of Bearing Based on Bayesian Theory. *Sensors* **2021**, *21*, 182. [CrossRef] [PubMed]
4. Que, Z.; Jin, X.; Xu, Z. Remaining Useful Life Prediction for Bearings Based on a Gated Recurrent Unit. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3511411. [CrossRef]

5. Kai, G.; Celaya, J.; Sankararaman, S.; Roychoudhury, I.; Saxena, A. *Prognostics: The Science of Making Predictions*; CreateSpace Independent Publishing Platform: Scotts Valley, CA, USA, 2017.
6. Shen, F.; Yan, R. A New Intermediate-Domain SVM-Based Transfer Model for Rolling Bearing RUL Prediction. *IEEE/ASME Trans. Mechatron.* **2022**, *27*, 1357–1369. [CrossRef]
7. Wang, F.; Liu, X.; Liu, C.; Li, H.; Han, Q. Remaining Useful Life Prediction Method of Rolling Bearings Based on Pchip-EEMD-GM(1,1) Model. *Shock Vib.* **2018**, *2018*, 3013684. [CrossRef]
8. Zhu, J.; Chen, N.; Shen, C. A new data-driven transferable remaining useful life prediction approach for bearing under different working conditions. *Mech. Syst. Signal Process.* **2020**, *139*, 106602. [CrossRef]
9. Guo, L.; Li, N.; Jia, F.; Lei, Y.; Lin, J. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing* **2017**, *240*, 98–109. [CrossRef]
10. Mi, J.; Liu, L.; Zhuang, Y.; Bai, L.; Li, Y.-F. A Synthetic Feature Processing Method for Remaining Useful Life Prediction of Rolling Bearings. *IEEE Trans. Reliab.* **2022**, *72*, 125–136. [CrossRef]
11. Saufi, M.S.R.M.; Hassan, K.A. Remaining useful life prediction using an integrated Laplacian-LSTM network on machinery components. *Appl. Soft Comput.* **2021**, *112*, 107817. [CrossRef]
12. Behera, S.; Misra, R.; Sillitti, A. Multiscale deep bidirectional gated recurrent neural networks based prognostic method for complex non-linear degradation systems. *Inf. Sci.* **2021**, *554*, 120–144. [CrossRef]
13. Zhang, Y.; Xin, Y.; Liu, Z.-W.; Chi, M.; Ma, G. Health status assessment and remaining useful life prediction of aero-engine based on BiGRU and MMoE. *Reliab. Eng. Syst. Saf.* **2022**, *220*, 108263. [CrossRef]
14. Xiang, S.; Qin, Y.; Luo, J.; Pu, H.Y.; Tang, B.P. Multicellular LSTM-based deep learning model for aero-engine remaining useful life prediction. *Reliab. Eng. Syst. Saf.* **2021**, *216*, 107927. [CrossRef]
15. Zhang, N.; Wu, L.; Wang, Z.; Guan, Y. Bearing Remaining Useful Life Prediction Based on Naive Bayes and Weibull Distributions. *Entropy* **2018**, *20*, 944. [CrossRef] [PubMed]
16. Zhang, B.; Zhang, L.; Xu, J. Degradation Feature Selection for Remaining Useful Life Prediction of Rolling Element Bearings. *Qual. Reliab. Eng. Int.* **2016**, *32*, 547–554. [CrossRef]
17. Pan, M.; Hu, P.; Gao, R.; Liang, K. Multistep prediction of remaining useful life of proton exchange membrane fuel cell based on temporal convolutional network. *Int. J. Green Energy* **2022**, *20*, 408–422. [CrossRef]
18. Ding, H.; Yang, L.; Cheng, Z.; Yang, Z. A remaining useful life prediction method for bearing based on deep neural networks. *Measurement* **2020**, *172*, 108878. [CrossRef]
19. Niu, T.; Zhang, L.; Zhang, B.; Yang, B.; Wei, S. An Improved Prediction Model Combining Inverse Exponential Smoothing and Markov Chain. *Math. Probl. Eng.* **2020**, *2020*, 6210616. [CrossRef]
20. Lv, M.; Liu, S.; Su, X.; Chen, C. Early degradation detection of rolling bearing based on adaptive variational mode decomposition and envelope harmonic to noise ratio. *J. Vib. Shock* **2021**, *40*, 271–280. [CrossRef]
21. Cheng, Y.; Wu, J.; Zhu, H.; Or, S.W.; Shao, X. Remaining Useful Life Prognosis Based on Ensemble Long Short-Term Memory Neural Network. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3503912. [CrossRef]
22. Wu, J.Y.; Min, W.; Chen, Z.; Li, X.L.; Yan, R. Degradation-Aware Remaining Useful Life Prediction With LSTM Autoencoder. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3511810. [CrossRef]
23. Thakkar, U.; Chaoui, H. Remaining Useful Life Prediction of an Aircraft Turbofan Engine Using Deep Layer Recurrent Neural Networks. *Actuators* **2022**, *11*, 67. [CrossRef]
24. Liao, Y.; Zhang, L.; Liu, C. Uncertainty prediction of remaining useful life using long short-term memory network based on bootstrap method. In Proceedings of the 2018 IEEE International Conference on Prognostics and Health Management (ICPHM), Seattle, WA, USA, 11–13 June 2018; pp. 1–8.
25. Zhang, C.; Lim, P.; Qin, A.K.; Tan, K.C. Multiobjective Deep Belief Networks Ensemble for Remaining Useful Life Estimation in Prognostics. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2306–2318. [CrossRef] [PubMed]
26. Shuai, Z.; Ristovski, K.; Farahat, A.; Gupta, C. Long Short-Term Memory Network for Remaining Useful Life estimation. In Proceedings of the 2017 IEEE International Conference on Prognostics and Health Management (ICPHM), Dallas, TX, USA, 19–21 June 2017.
27. Sateesh Babu, G.; Zhao, P.; Li, X.-L. Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life. In Proceedings of the Database Systems for Advanced Applications, Cham, Switzerland, 25 March 2016; pp. 214–228.
28. Gu, Y.; Wylie, B.K.; Boyte, S.P.; Picotte, J.; Howard, D.M.; Smith, K.; Nelson, K.J. An Optimal Sample Data Usage Strategy to Minimize Overfitting and Underfitting Effects in Regression Tree Models Based on Remotely-Sensed Data. *Remote Sens.* **2016**, *8*, 943. [CrossRef]