



OPEN

A remaining useful life prediction method based on PSR-former

Huang Zhang^{1,2}, Shuyou Zhang^{1,2}, Lemiao Qiu^{1,2}✉, Yiming Zhang^{1,2}, Yang Wang^{1,2}, Zili Wang^{1,2} & Gaopeng Yang^{1,2}

The non-linear and non-stationary vibration data generated by rotating machines can be used to analyze various fault conditions for predicting the remaining useful life(RUL). It offers great help to make prognostic and health management(PHM) develop. However, the complexity of the mechanical working environment makes the vibration data collected easily affected, so it is hard to form an appropriate health index(HI) to predict the RUL. In this paper, a PSR-former model is proposed including a Phase space reconstruction(PSR) layer and a Transformer layer. The PSR layer is utilized as an embedding to deepen the understanding of vibration data after feature fusion. In the Transformer layer, an attention mechanism is adopted to give different assignments, and a layer-hopping connection is used to accelerate the convergence and make the structure more stable. The effectiveness of the proposed method is validated through the Intelligent Maintenance Systems (IMS) bearing dataset. Through analysis, the prediction accuracy is judged by the parameter RMSE which is 1.0311. Some state-of-art methods such as LSTM, GRU, and CNN were also analyzed on the same dataset to compare. The result indicates that the proposed method can effectively establish a precise model for RUL predictions.

Mechanical vibration is a common phenomenon in the operation of industrial equipment. With the increase of equipment service time, various wear and defects will gradually form on internal parts, resulting in equipment performance degradation, thus affecting the service life of the equipment. Bearing, as the core component of most rotating machines, once fails, the equipment will be affected or even collapsed, making it difficult to maintain the prognostic and health management (PHM)¹ of the equipment which is a task for real-time equipment operation monitoring. This information including the equipment operation status will be reflected in the vibration data through abnormal fluctuations. Therefore, PHM needs to analyze the bearing vibration data and predict the remaining useful life(RUL) that enables the service time of the equipment to be described quantitatively² which is conducive to reducing costs and improving production efficiency.

PHM driven by data to predict RUL is inseparable from time analysis methods³. The numerical calculation method⁴, data-driven method^{5,6}, and data-model combination method^{7,8} are the common methods for RUL prediction in time analysis. Among the methods using the modeling approach for prediction, Gabelli et al.⁹ predicted the RUL based on the rolling contact fatigue theory study by expressing the survival probability of the raceway surface as the basic life equation. In data-model combination, Qin et al.⁷ estimated the size of the defects by measuring the signals to reveal the evolution law of the defects in time and formed the bearing dynamics model. Since the model-based and data-model-based approaches require an explicit and comprehensive analytical model to reflect the bearing degradation process, it is difficult to achieve in practical production, making the physical model hard to describe complex systems. However, the data-driven approach allows the model to learn degradation patterns directly from the data collected by sensors to reduce the dependence on the physical knowledge background. Meanwhile, as the era of big data has led to the development of data-driven approaches¹⁰, a large amount of industrial data can be collected, so data-driven approaches are gradually becoming dominant¹¹.

The Recurrent Neural Network (RNN) has a unique recursive structure, which is conducive to extracting information from time series, so it is gradually applied to RUL prediction¹². However, with long-term iteration, the information will be blurred gradually, and the gradient will disappear or explode. To solve the problem, Long short-term memory (LSTM) was proposed. LSTM¹³ adds the "gate" structure to enhance the ability of the model to learn the series information. Then the Gate Recurrent Unit (GRU) method¹⁴ simplifies the LSTM by merging the input and forget gates as update gates. It simplifies the structure but also reduces the ability to express the complexity of the model. Convolutional Neural Network (CNN)¹⁵ is mostly used in the field of image processing

¹The State Key Laboratory of Fluid Power and Mechatronic System, Zhejiang University, Hangzhou 310027, China. ²Engineering Research Center for Design Engineering and Digital Twin of Zhejiang Province, Hangzhou 310027, China. ✉email: qiuml@zju.edu.cn

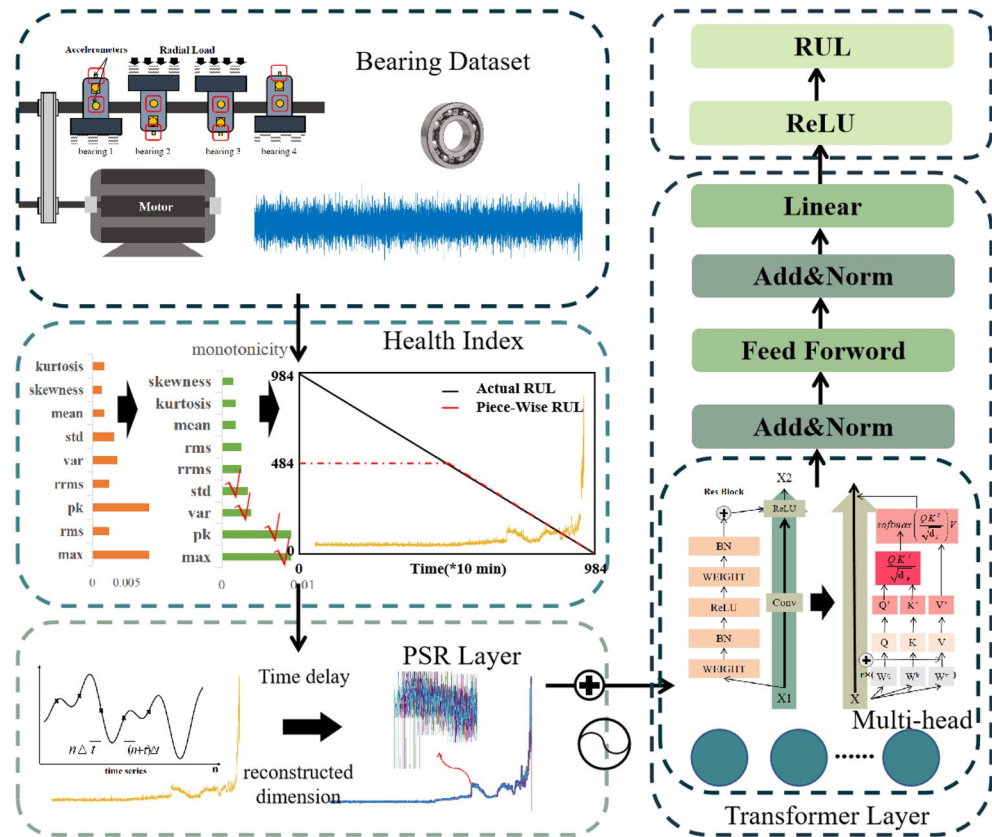


Figure 1. Flowchart of RUL prediction process.

but it can also be applied to RUL prediction using a one-dimensional structure. Nonetheless, these methods have inherent sequential properties, which hinder the parallelization of training samples.

With the publication of "Attention is all you need"¹⁶, the attention mechanism has been gradually applied to time series, inspired by the human visual attention mechanism. It relies on attention to model the whole series without considering the specific position in the sequence, so in this way, parallel computation can be carried out. The attention mechanism takes different weighting factors to make sure the input sequence gets different attention assignments. The Transformer model uses encoder-decoder architecture. In recent years it has been used more and more for time series prediction in the industry. Ding et al.¹⁷ designed a new tokenizer and encoder module to extract features from the time-frequency of vibration data and then used Transformer to diagnose the faults. Alexakos¹⁸ proposed an image classification transformer used to diagnose the vibration images after a short time Fourier transform. Unfortunately, Transformer has not been well mined in the RUL prediction field¹⁷, so the advantages of Transformer in avoiding recursion, parallel computation, and reducing performance degradation are not well utilized in RUL prediction.

Based on the above problems, a PSR-former model is proposed using the strong ability of the Transformer in global sequence modeling. The PSR layer further deepens the understanding of vibration features after feature fusion since the features are easily polluted by noise and it is difficult to form HI for prediction. The Transformer layer receives the enhanced vibration features as input to predict the RUL. Bearings are the most important and easily damaged part of rotating machines. Therefore, the bearings are analyzed as an example, and the RMSE value is used as a measurement tool to show the RUL accuracy to illustrate the implementation effect that has achieved 1.0311. Some state-of-art methods are also adopted on the same dataset to compare the result with PSR-former. However, PSR-former still shows good performance after the comparison.

The rest of the paper is organized as follows. "Proposed method" describes the method specifically. "Dataset" introduces the IMS bearing dataset. "Experiment results" processes the dataset using the proposed method and then compares the results with some deep learning algorithms and traditional machine learning methods. "Conclusions" summarizes the main contribution and looks forward to future work.

Proposed method

This section describes the PSR-former method, including two important components: the construction of the health index(HI)¹⁹ which reflects the health status of equipment in the form of values, and the construction of the PSR-former model. The model was conducted in Python3.8 with PyTorch. The graphics card model is NVIDIA GeForce RTX 2060. The flowchart of the RUL prediction process is shown in Fig. 1. Features of vibration data were extracted and the monotonicity index was chosen to select the features to form a new health index enhanced

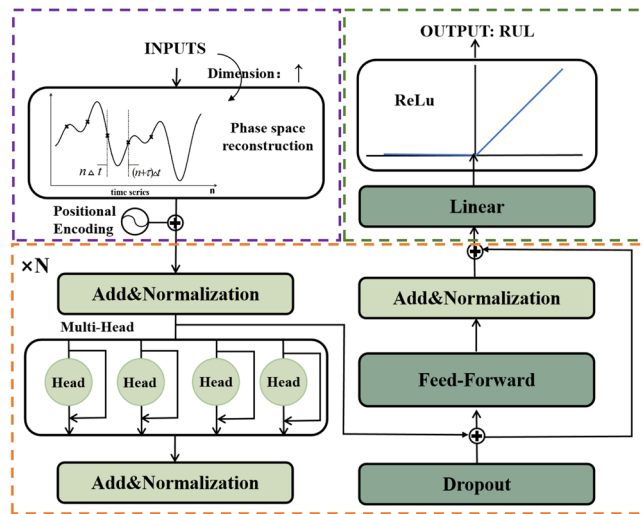


Figure 2. PSR-former model structure.

initially. Then it was input into the proposed PSR-former model to predict the RUL of the bearing. The specific description of each step is as follows.

A. Construction of the health index. Feature extraction is crucial in RUL prediction²⁰. In general, feature extraction is mainly analyzed from the time and frequency domain perspectives. The time domain analysis method is to analyze and discriminates the signals directly by analyzing the time scale parameters of the signals. It commonly includes statistical features such as mean, variance, kurtosis, deviation, and pulse indicators of the waveform.

The dimensionless time-domain feature is sensitive to the impact energy, but in the case of rolling bearings that wear slowly over time, there is a greater need for a feature that can describe the entire process of degradation.

The sensitivity, regularity, and monotonicity of the bearing fault differ from each other, it is difficult to quantify the degree of performance degradation. Based on this, this paper uses monotonicity²¹ as the discriminatory criterion for feature selection and accordingly considers that the selection of features should be as clear as possible to reflect the general trend of degradation.

$$\text{Monotonicity}(f_i) = \left| \frac{\text{Num}(\text{diff}(f_i) > 0)}{n-1} - \frac{\text{Num}(\text{diff}(f_i) < 0)}{n-1} \right| \quad (1)$$

where n is the number of sampled signals, f_i is the i th characteristic of signals, and $\text{diff}(f_i)$ is the difference of the i th feature of signals.

The features selected in this way can characterize the monotonicity trend of degradation. To better coalesce the features, PCA is used to reconstruct the features that meet the monotonicity discriminant to form new features as the health index of bearing deterioration.

B. PSR-former model. The most important part attention mechanism in Transformer was first proposed when using cosine functions to measure the similarity between reading and writing keyword vectors and working memory thus allocating different attention. The Transformer completely discards the RNN and CNN architectures and utilizes the attention mechanism to obtain powerful feature extraction and long-distance feature capture capabilities. Based on this, Transformer is increasingly used in temporal processing. However, Transformer is rarely applied to RUL prediction in the industry.

A PSR-former model based on the Transformer structure was then proposed. The PSR layer enhanced the ability to capture feature information by using the PSR layer in the encoder as the embedding to perform the dimension upgrading operation on the input. The self-attention structure in the encoder is connected as the residual block to enhance the weight of the feature matrix. The decoder layer is replaced by the linear layer to realize the regression problem. The model structure is shown in Fig. 2, where τ represents the delay time and N means having N identical structures. The purple part in Fig. 2 represents the PSR layer which is also the embedding part of the PSR-former model. The orange part in Fig. 2 represents the main structure of the Transformer structure. Features are further extracted and enhanced by the orange part. The model outputs features as RUL values through a linear layer in the green part of the PSR-former model. The key components of the model are described as follows:

B.1 PSR feature enhancing layer. In actual industry, the operation of equipment is a complex system. Although the operation of bearings is relatively simple, they can still be easily affected by the surrounding environment. The actual vibration data generated contains a large number of noise and variables making the time series chaotic. Phase Space Reconstruction is the key step of this system²². To further extract the useful infor-

mation in vibration data, the phase space reconstruction technique was used to map the low-dimensional time series to higher dimensions and thus further find the characteristic laws embedded within the system. In order to realize the mapping from low dimensional space to high dimensional space, the Takens theorem²³ was proposed to ensure that the phase space can be reconstructed from the one-dimensional time series in the same sense as the topological equivalence of the original dynamical system. The reconstruction of phase space is realized by coordinate delay. It needs two key parameters, dimension m and delay time τ . The common method to determine the delay time τ is mainly by the mutual information and autocorrelation function method. However, the autocorrelation function can only extract the linear correlation between time series, it is difficult to extend the time delay to higher dimensions. Therefore, the mutual information method is chosen in this paper. Firstly, the maximum time delay t for computing the mutual information is selected to obtain the time system $M:(m_1, m_2, \dots, m_t)$ and $N:(n_1, n_2, \dots, n_t)$ under different time delays where $n_i = m_i + \tau$. The entropies of information obtained from the two time systems are²³:

$$H(M) = - \sum_{i=1}^t P_i(m_i) \log_2 P_m(m_i) \tag{2}$$

$$H(N) = - \sum_{i=1}^t P_i(n_i) \log_2 P_n(n_i) \tag{3}$$

The probability of the occurrence of events m_i, n_i is constructed as a boundary statistic in terms of the most value of the sequence. The mutual information of system N is obtained by the time system M .

$$I(N, M) = H(M) - H(M|N) \tag{4}$$

$$I(N, M) = \sum_i \sum_j P_{mn}(m_i, n_i) \log_2 \left[\frac{P_{mn}(m_i, n_i)}{P_m(m_i)P_n(n_i)} \right] \tag{5}$$

where:

$$H(N|m_i) = - \sum_j [P_{mn}(m_i, n_i)/P_m(m_i)] \log [P_{mn}(m_i, n_i)/P_m(m_i)] \tag{6}$$

The first minimal value $I(N, M)$ calculated at different time delays is used as the optimal delay time τ which represents the greatest degree of uncorrelation between the reconstructed time series.

After determining the delay time τ , the dimension m is determined by Cao's method²⁴. A maximum dimension M is first determined and then the phase space is reconstructed using the delay time determined before.

$$\begin{cases} S(1) = [x(1), x(1 + \tau), \dots, x(1 + (m - 1)\tau)] & (7) \\ S(2) = [x(2), x(2 + \tau), \dots, x(2 + (m - 1)\tau)] & (8) \\ \dots & \\ S(k) = [x(k), x(k + \tau), \dots, x(k + (m - 1)\tau)] & (9) \end{cases}$$

For each dimension, the distance between the i th vector and other vectors is calculated, and for each phase vector $x(i)$, there is a nearest proximal point $x^N(i)$. The distance between them is:

$$R_d(i) = \|x(i) - x^N(i)\| \tag{10}$$

When the dimension of the phase space increases by one dimension, the distance between the phase points will change.

$$R_{d+1}^2(i) = R_d^2(i) + \|x(i + \tau d) - x^N(i + \tau d)\|^2 \tag{11}$$

If $R_{d+1}(i)$ is much larger than $R_d(i)$, then consider the adventitious point false. Define a in the following.

$$a(i, d) = \frac{R_{d+1}^2(i) - R_d^2(i)}{R_d(i)} = \frac{\|x(i + \tau d) - x^N(i + \tau d)\|^2}{R_d(i)} \tag{12}$$

then

$$a(i, d) = \frac{R_{d+1}^2(i) - R_d^2(i)}{R_d(i)} = \frac{\|x(i + \tau d) - x^N(i + \tau d)\|^2}{R_d(i)} = \frac{\|x(i + \tau d) - x^N(i + \tau d)\|^2}{\|x(i) - x^N(i)\|^2} \tag{13}$$

So $a(i, d)$ can be:

$$a(i, d) = \frac{\|x_{d+1}(i) - x_{d+1}^N(i)\|^2}{\|x_d(i) - x_d^N(i)\|^2} \tag{14}$$

$x_d(i)$ and $x_d^N(i)$ are the i th vector in the d -dimensional space and its most proximal point. By calculating $E_1(m)$ and $E_2(m)$, where $E_2(m)$ is a judgment criterion defined to better observe the change of $E_1(m)$, the embedding dimension m can be judged when $E_1(m)$ and $E_2(m)$ are constant.

$$E_i(m) = E_i^*(m + 1)/E_i^*(m) \tag{15}$$

where:

$$E_1^*(m) = \frac{1}{N - m\tau} \sum_{i=1}^{N-m\tau} a(i, m) \tag{16}$$

$$E_2^*(m) = \frac{1}{N - m\tau} \sum_{i=1}^{N-m\tau} |x(i + m\tau) - x^N(i + m\tau)| \tag{17}$$

When $E_1(m)$ is essentially stable, the dimension m at this point is the embedding dimension needed for the reconstruction. Similarly, the dimension m selected ensures the minimum correlation between sequences. The definition of $E_2(m)$ is based on the uncorrelation sequences. It always equals 1.

After getting the reconstructed dimensions to reconstruct the time series, Transformer uses the sliding window to divide the reconstructed time series data to get a sequence of time window data blocks. The data shape of each sample is: $[num, time, feature]$, where num is the number of time window sequences, $time$ is the length of the time window, and $feature$ is the number of reconstructed dimensions. The Transformer uses the global information but not the sequential information of the sequence. It needs to calculate the relative position of each sequence using position embedding. Its calculation formulas are:

$$PE(pos, 2i) = \sin(pos/10000^{\frac{2i}{d}}) \tag{18}$$

$$PE(pos, 2i + 1) = \cos(pos/10000^{\frac{2i}{d}}) \tag{19}$$

where d is the dimension of the sequence. The results of the sequence after feature enhancement and the result after position embedding are added to obtain the representation vector of the sequence as the input to the model.

B.2 Multi-head attention. The self-attention structure requires the query vector Q , the key vector K and the value vector V . Q, K , and V are obtained by multiplying the input matrix X by the matrix W^Q, W^K, W^V , where W^Q, W^K, W^V are trainable projection matrices. The multi-head attention layer uses the time series data to do the dot calculation. The output obtained is as follows:

$$Attention(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{20}$$

Inspired by the residual network, the residual unit can solve the degradation problem of the network and make convergence faster by connecting the input and attention fractions through layer hopping. Cao et al.²⁵ showed that the residual connection is an effective way to train the network to transfer information across layers and prevent the gradient from disappearing and exploding. Based on this, a layer-hopping was also performed in the calculation of Q, K, V so that the original Q, K, V becomes Q', K', V' . The structure is shown in Fig. 3. The projection matrix W^Q, W^K, W^V are multiplied twice by the coefficient r and the input matrix X to obtain Q, K and V , and then they are connected with X by layer hopping.

$$\begin{cases} Q = X \times rW^q \\ K = X \times rW^k \\ V = X \times rW^v \end{cases} \tag{21}$$

$$\begin{cases} Q' = X + X \times rW^q \\ K' = X + X \times rW^k \\ V' = X + X \times rW^v \end{cases} \tag{22}$$

The transformation matrix is multiplied by the coefficient r . The query vector Q is used as the analysis to show the loss back-propagation. The computation without taking a jump layer connection is:

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Q} \cdot \frac{\partial Q}{\partial X} = W^q \frac{\partial L}{\partial Q} \tag{23}$$

And when the idea of residuals is used for layer-hopping connections, the loss is.

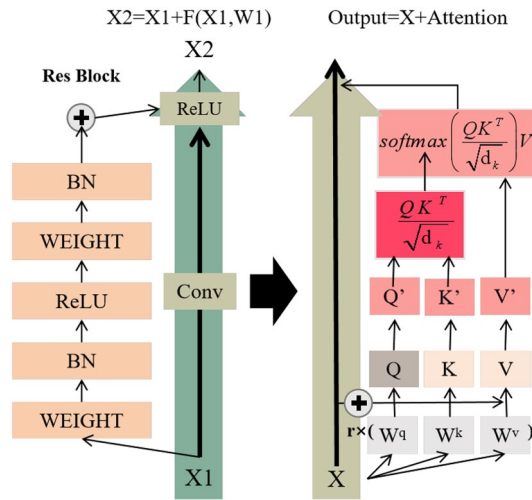


Figure 3. Layer-hopping connection mode.

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Q'} \cdot \frac{\partial Q'}{\partial X} = (1 + rW^q) \frac{\partial L}{\partial Q'} \tag{24}$$

It can be known that when this connection is used, the matrix has a greater range of variation during propagation and it is updated faster during gradient descent. Similarly, the Attention score is computed and then connected with the input X hopping layer to obtain $X + Attention$.

Each Attention value result is computed as a Head, and multiple Heads are copied and spliced together in the Multi-Head layer.

$$MultiHead(Q', K', V') = Concat(head_1, head_2, \dots, head_h)W^O$$

where $W^O \in R^{d_{model} * d_{model}}$, d_{model} is the input dimension of the sequence.

After Multi-Head Attention, the sequences need to go through the Add&Norm layer which consist of an add layer and a normalization layer. The add layer also takes a residual connection to make the network focus on the current difference part. The norm layer makes the output of each layer gets normalized to speed up the convergence.

$$LayerNorm(X + MultiHeadAttention(X))$$

$$LayerNorm(X + FeedForward(X))$$

Dataset

The data was provided by the Center for Intelligent Maintenance Systems (IMS), University of Cincinnati²⁶. The installation of bearings and sensors is shown in Fig. 4. The test rig is mainly composed of a motor, belt, shaft, bearing, sensors, and additional radial load. The sensors are high sensitivity PCB 253B33 Quartz ICP accelerometer located on the bearing. It contains three datasets, representing the whole process of bearing from normal operation to failure. There are four bearings in each dataset. The bearing type is Rexnord ZA-2115. The shaft speed of the test rig is 2000 rpm and the radial load is 6000 lbs. The sampling frequency is 20.48 kHz, the sampling time is 1 s, and the period is 10 min. The endurance duration of the three datasets is 828 h, 164 h, and 741.3 h. The fault location appeared in the inner ring, rolling element, and outer ring respectively in the different datasets which are specifically described in Table 1. The oil return pipe lubricated with oil is provided with a magnetic plug. When the debris adsorbed on the magnetic plug reaches the threshold value, it is considered that the bearing has been completely degraded, then the collection work stops. A description of the three different datasets is shown in Table 1.

Experiment results

This section provides the results of each step in this paper including feature reconstruction, PSR-former, and comparative experiments. The feature reconstruction step initially enhanced the degradation features of the bearing. PSR-former was the main structure in this paper. It combined the PSR layer with the Transformer layer. Then comparative experiments were conducted on the dataset.

A. Feature reconstruction. Dataset2 was used as an example to illustrate the experiment which describes the whole process of four bearings with only bearing 1 from normal to failure. The outer ring fails at last. The full RMS process of bearing 1 is shown in Fig. 5.

The expression of RMS is:

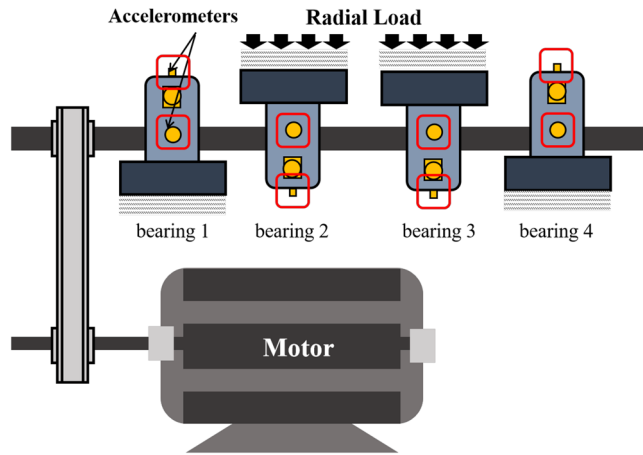


Figure 4. Installation diagram of bearings and sensors on IMS.

	Num of DOC	Num of channel	Duration	Fault location
Dataset 1	2156	8	828 h	Bearing 3: inner ring Bearing 4: rolling element
Dataset 2	984	4	164 h	Bearing 1: outer ring
Dataset 3	4448	4	741.3 h	Bearing 3: outer ring

Table 1. Description of three datasets.

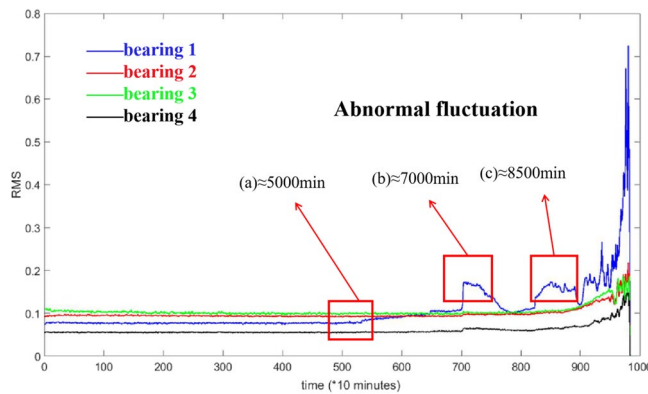


Figure 5. Process of bearing RMS for dataset 2.

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2} \tag{25}$$

where x_i denotes the i th point of the data.

It can be seen from Fig. 5 that the whole process of bearing degradation can be approximately divided into four stages according to the change of amplitude. The abscissa in Fig. 5 is the time point of bearing operation and the ordinate is the RMS value of bearing according to the time point. The four stages were roughly determined: (a) At about 5000 min, the curve shows a small upward fluctuation, indicating that the bearing starts to fail here. (b) At about 7000 min, the curve shows a large step, and then it slowly declines with similar fluctuation amplitude after (a) point. (c) At about 8500 min, the RMS shows a large oscillation, indicating that the bearing has reached a serious failure at this time. Therefore, the degradation process of the bearing can be roughly divided into four categories, 0–(a) for normal operation; (a)–(b) for mild failure; (b)–(c) for moderate failure; and (c)–last for severe failure. Hence, it can be considered that after point (a), the bearing began to deteriorate gradually.

In order to better discriminate the RUL of the bearing, the RUL of the data was truncated, and the highest value of RUL was set to 484×10 min according to (a) point. i.e., the RUL values of the time cycle from 0 to

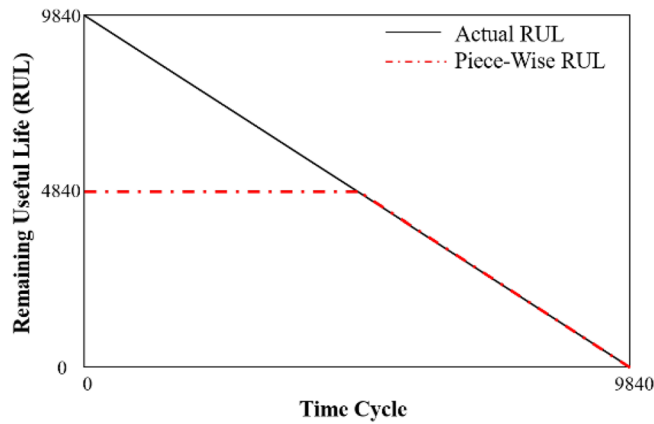


Figure 6. Piece-wise RUL.

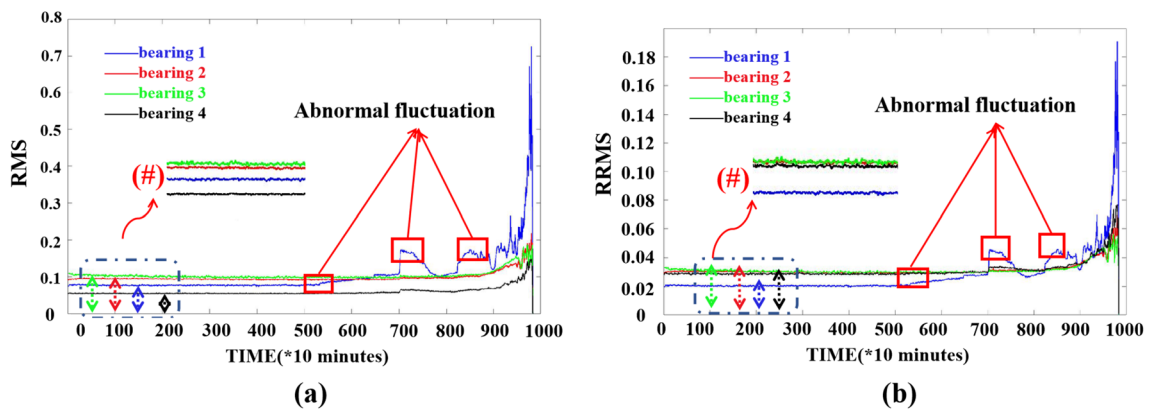


Figure 7. Plots of RMS vs. RRMS (a) RMS; (b) RRMS.

$Mean = \frac{1}{N} \sum_{i=1}^N x_i$	$Pk = \max\{ x_i \}$	$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$
$RRMS(i) = \frac{RMS(i)}{\frac{1}{N} \sum_{i=1}^N RMS(i)}$	$Max = \max\{x_i\}$	$Std = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$
$Kurtosis = \frac{E[(x-\mu)^4]}{\sigma^4}$	$Skewness = \frac{E[(x-\mu)^3]}{\sigma^3}$	$Var = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$

Table 2. Time domain features.

5000 min are 484×10 min and it is shown in Fig. 6. This operation mapped the HI to the RUL label. The accuracy of artificial division will not affect the judgment of the model on the degradation process since 484×10 min is located in the health stage. It did not contain the characteristics of bearing wear.

To show the degradation trend of four bearings in the same dataset, a new feature Reconstructed Root Mean Square (RRMS) was defined based on Root Mean Square (RMS). The definition of RRMS is:

$$RRMS = \frac{RMS}{RMS_{norm}}, \quad RMS_{norm} = \frac{1}{k} \sum_{i=1}^k RMS(i) \quad (26)$$

By comparing the RMS and RRMS of bearings in Fig. 7, it can be seen from (#) in Fig. 7 that RRMS is more beneficial to reduce individual differences. Since there are four bearings in each dataset in IMS bearing data, RRMS can more intuitively distinguish the degradation process of different bearings, although it has the same trend of change as RMS. Hence the RRMS was chosen as one of the features in the selection of RMS and RRMS. The obtained time domain features are shown in Table 2. Where x is the series of vibration data, N is the number of the data point, μ is the mean value of the whole vibration data and σ is the standard deviation.

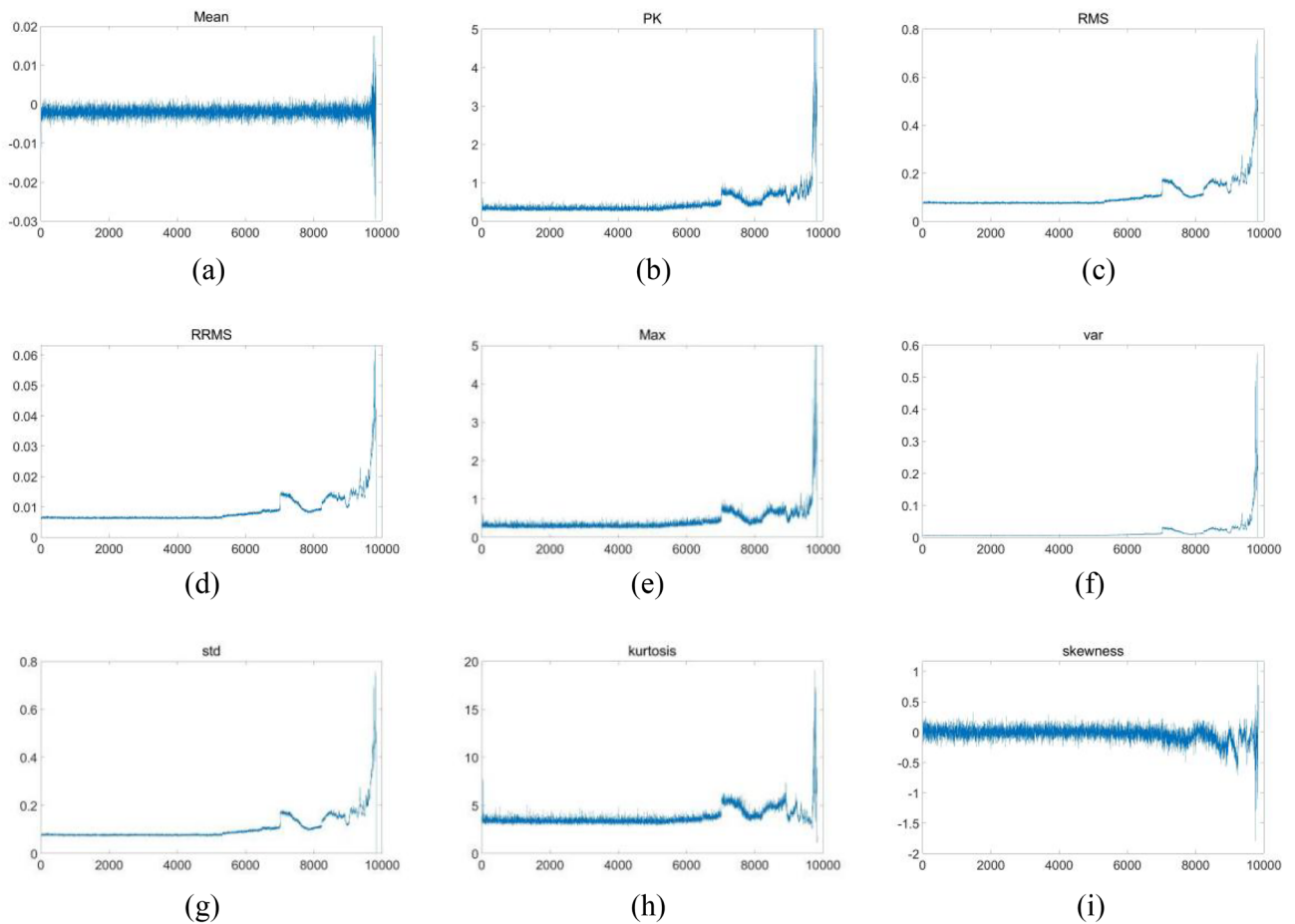


Figure 8. Time-domain features of dataset2 bearing 1 (a) mean; (b) peak; (c) RMS; (d) RRMS; (e) max; (f) var; (g) std; (h) kurtosis; (i) skewness.

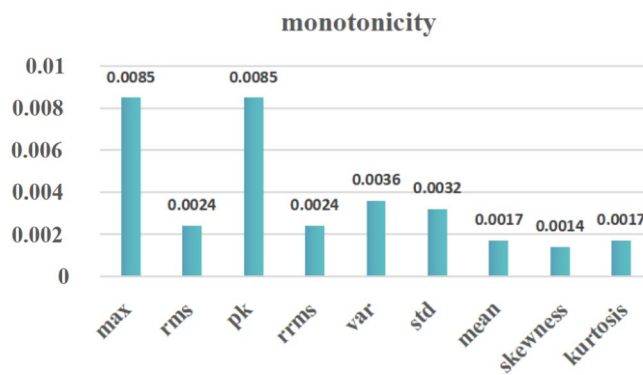


Figure 9. Monotonicity of different features.

Other features of the full process curve of dataset2 bearing1 from the time domain perspective were extracted in Fig. 8, which shows nine features of bearing in dataset2 including mean, peak value(*PK*), RMS, RRMS, max value, variance(*Var*), standard deviation(*Std*), kurtosis, and skewness.

The purpose of feature screening is to select clear features that reflect the general trend of degradation as much as possible. The extracted features were selected according to monotonicity. The monotonicity value of each feature is shown in Fig. 9. The top four features were selected. However, the Peak value and the max value have the same trend, and the RMS and the RRMS have the same trend, so one of them was selected. Then *PK*, *RRMS*, *Var*, and *Std* were selected as the four features that represent the time series features finally. The feature reconstruction of the four features by PCA found that the first principal component accounted for 96% of the total, thus the new principal component features were reconstructed to obtain.

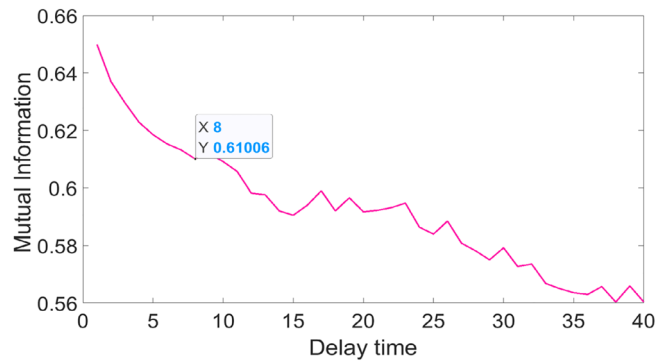


Figure 10. Time delay curve.

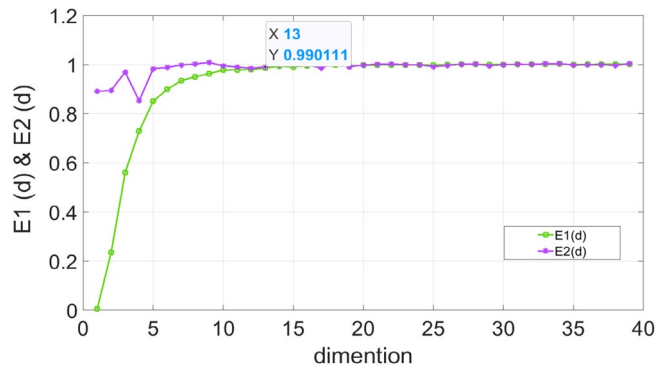


Figure 11. Dimensional determination chart.

$$principle1 = 0.5014 \times PK + 0.5031 \times RRMS + 0.4923 \times Var + 0.5031 \times Std \tag{27}$$

After the feature reconstruction step, the features were initially enhanced.

B. PSR-former. The reconstructed time series features are then input into the PSR-former model established to judge the RUL of the bearing. First, the PSR layer of the PSR-former reconstructs the time series. The maximum time delay t is determined according to the mutual information method. The curve of mutual information with delay time was plotted to obtain the first minimal value point in Fig. 10. It shows the determined time delay $t=8$.

After determining the time delay t , the phase space was initially reconstructed according to the time delay, and the magnitude of $E_1(m)$ and $E_2(m)$ values were calculated according to the Cao's method²⁴, and the variation curves of $E_1(m)$ and $E_2(m)$ in different dimensions are plotted as shown in the following Fig. 11.

Through the curves, it can be observed that when the dimensions $m \geq 13$, $E_1(m)$ and $E_2(m)$ converge, there is no fluctuation after a certain point. Then the characteristic dimension of the time series was extended to 13 through the PSR layer by Eqs. (7)–(9). The degradation characteristics were enhanced after the PSR layer shown in Fig. 12. The time series after the enhancement is then input into the Transformer, and the parameters of the model were evaluated according to its prediction results for RUL. The highest value of RUL is set to 484×10 min by truncation. For the Encoder layer, the dimensionality of encoder d_model is set to 14, the total dimension of the time series. The number of layers N is 4. For multi-head attention, the *head* is 7 and the jump connection coefficient r is 1, the *dropout* is set to 0.4, the optimizer is Adam optimizer and the *learning rate* is 0.001. Set the test data to the last ten percent of the total data to predict the RUL to show the service life of the bearing before the final collapse. The training set and the validation is 4:1. The dimension in the Feed-Forward layer is 64. The batch size is 64, the epoch is 1000 and the step is 1. The dataset2 was trained and tested. To speed up the training process, the input data were normalized so the interval range becomes $[-1, 1]$. The RMSE of the test set was calculated, and the results of the model were compared by RMSE. The definition of RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (y_i - \bar{y}_i)^2} \tag{28}$$

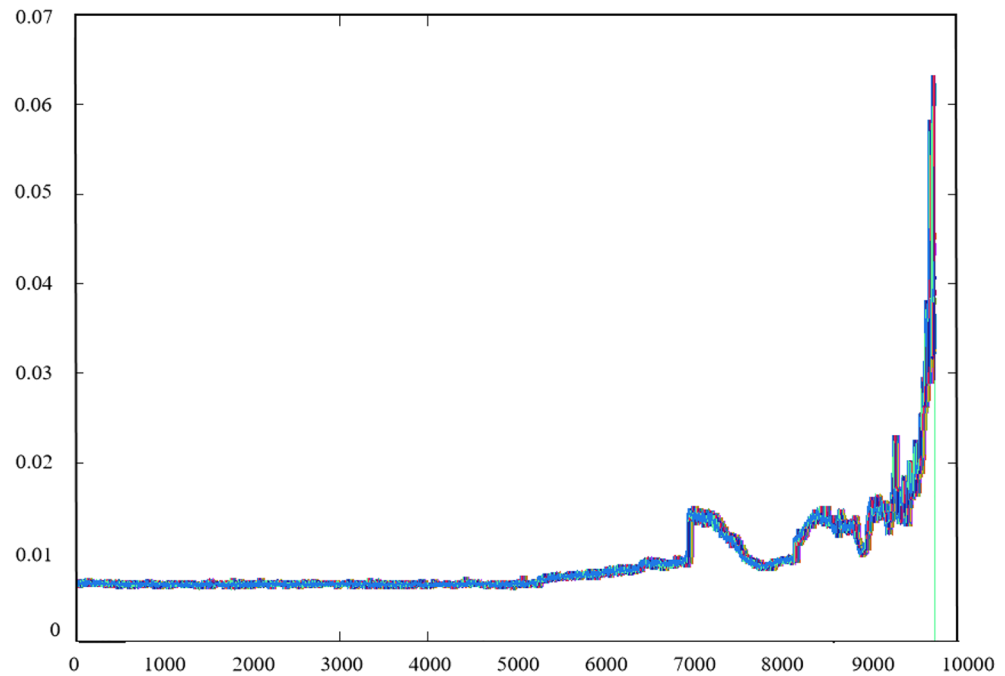


Figure 12. Result after the PSR layer.

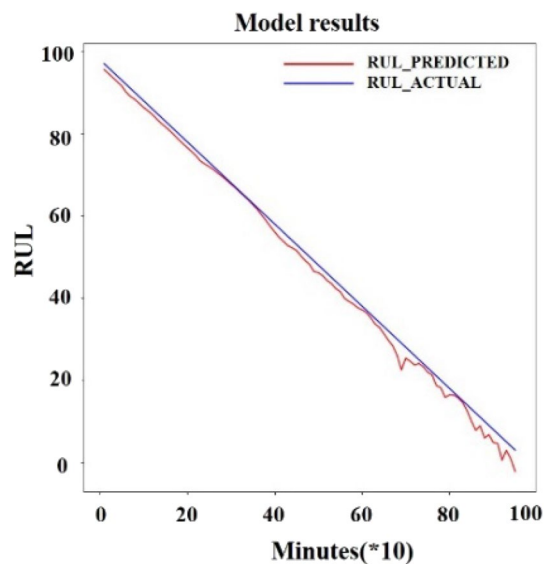


Figure 13. Test result in dataset 2.

where y_i is the true RUL value and \bar{y}_i is the predicted RUL value. The results obtained from the test set is showed in Fig. 13.

The RMSE value obtained from the test set after the normalization of the data is 0.0014. The RMSE value after inverse normalization is 1.0311. To make the prediction more reliable, uncertainty estimation was added to the forecast. Uncertainty estimation combined with deep learning is mainly applied to images to estimate the boundaries. Gal²⁷ divided uncertainty into two main categories, aleatoric uncertainty, and Epistemic uncertainty. Aleatoric uncertainty mainly originates from the data itself. Epistemic uncertainty mainly measures the uncertainty of the estimated parameters of the model during the training process. The model fusion approach²⁸ was adopted to estimate the uncertainty of the model. Use the mean as the predicted value and variance as the uncertainty to do epistemic uncertainty estimation. The results are obtained as shown in Fig. 14.

From the results, it can be seen that the predicted results are accurate to some extent and the uncertainty interval is distributed evenly around the actual RUL. Dataset1 operation is completed with bearing3 having inner

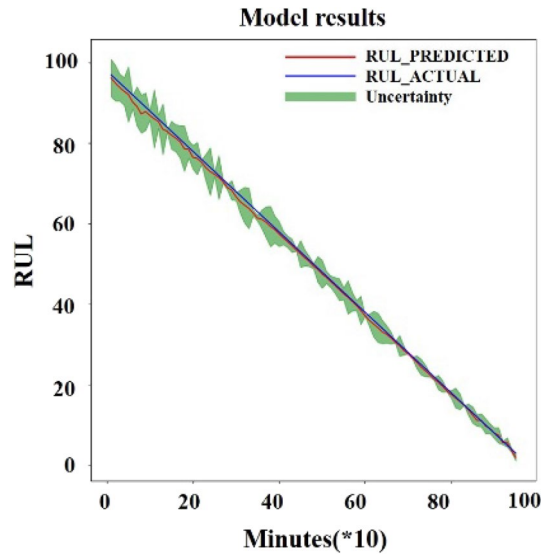


Figure 14. Test result with uncertainty in dataset 2.

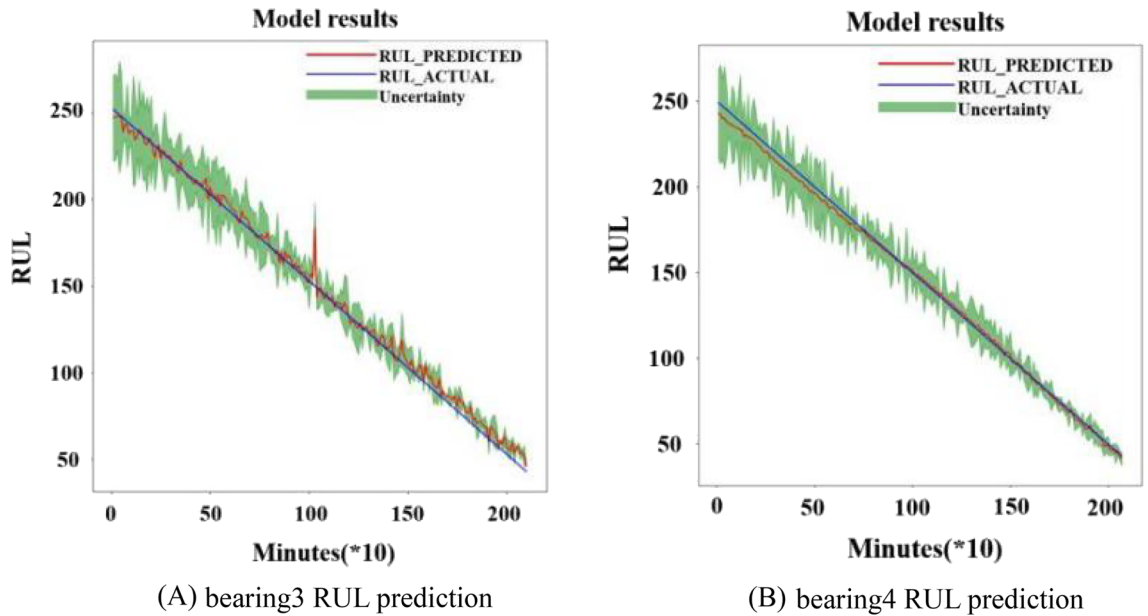


Figure 15. Test result with uncertainty in dataset 1; (A) bearing3; (B) bearing4.

ring failure and bearing4 having rolling element failure. The parameters of the PSR are determined with delay time $t=2$ and dimension $m=15$ after analysis on bearing3 and bearing4. Then the RUL of the two bearings of dataset1 were predicted based on the same training parameters shown in Fig. 15 where the average RMSE value of bearing3 is 3.7641 and the average RMSE value of bearing4 is 1.9729.

The prediction for bearing3 fluctuates more than bearing4 in Fig. 15. There are more unusual fluctuations in the middle of the test set and the predicted value is slightly larger than the accurate value in the second half of the test set of bearing3. The prediction for bearing4 shows that the predicted value is slightly smaller than the accurate value in the first half of the test set. Overall, the predictions are accurate to some extent.

C. Comparative experiments. TO verify the effectiveness of the method, comparative experiments were conducted to compare the results of the PSR-former with and without the PSR layer and the results of the analysis of dataset 2 using different methods, respectively. The data used in comparative experiments are data in data-

set2 without the upgrade dimension process. At present, the state-of-art methods in RUL prediction are mainly based on machine learning. In order to compare the methods in this paper with those adopted in the current field, the algorithms are classified into two categories: the method based on deep learning and the method based on traditional machine learning. In deep learning methods, Transformer²⁹, RNN³⁰, LSTM³¹, GRU³², 1D-CNN³³, and CNN combined with LSTM³⁴ were chosen under the same data settings as proposed method in this paper to analyze and they were all connected in three layers whose hidden dimension is 14 and 64 to be as close as possible to the PSR-former model. The other parameters of the networks such as batch size and learning rate were adjusted to the best. Random Forest(RF)³⁵, Support Vector Machine(SVM)³⁶, Linear Regression(LR)³⁷, Logistic Regression(LOR)³⁸, and K-Nearest Neighbours(KN)³⁹ were chosen as traditional machine learning algorithms to analyze the data. In addition, Stacked-autoencoder-LSTM(SAE-LSTM) model⁴⁰ and Bidirectional (BiLSTM) model⁴¹ in the class of deep learning were tested through similar data in this paper. The two models were trained according to the literature which were built on the same IMS bearing dataset. The SAE-LSTM model structure was 5–20–5 which is the size of different hidden layers of SAE-LSTM. However, the parameters in the BiLSTM were not fully recorded, to maintain consistency, the unrecorded parameters are the same as the SAE-LSTM and the model structure was also 5–20–5. Then the two models were tested using the same test data as PSR-former model.

If the predicted RUL deviates greatly from the actual RUL, the RMSE value will be too large. At this time, the prediction result is not ideal, which is meaningless for analysis. Therefore, a threshold is chosen for RMSE which is set to 100, then the predicted result of RMSE less than 100 is plotted in Fig. 16. In Fig. 16, (A) is the result of Transformer, (B) is the result of RNN, (C) is the result of LSTM, (D) is the result of GRU, (E) is the result of one-dimensional CNN (1D-CNN), (F) is the result of CNN-LSTM, the combination of CNN and LSTM, (G) is the result of RF, (H) is the result of SVM, (I) is the result of SAE-LSTM, and (J) is the result of BiLSTM. The blue line in the graph represents the real RUL of the bearing, the red line is the predicted RUL, and the green interval represents the uncertainty of the prediction results. The interval obtained from multiple predictions can make the errors between real and predicted RUL more intuitive. The specific values of RMSE results are recorded in Table 3. In Table 3, (1) is Transformer, (2) is RNN, (3) is LSTM, (4) is GRU, (5) is 1D-CNN, (6) is CNN-LSTM, (7) is SAE-LSTM, (8) is BiLSTM, (9) is RF, (10) is SVM, (11) is LR, (12) is LOR, (13) is KN, and the last one is the method proposed in this paper. In addition the trained model was also validated through bearing3⁴¹. The last 1/3 prediction result is shown in Fig. 17. From the result, it can be seen that the model can accurately predict the final stage of bearing failure.

From Fig. 16, it can be seen that when using deep learning methods to analyze the dataset, the predicted RUL value deviates slightly from the true value and oscillates up and down around. It can be known from the RMSE results in Table 3, the deep learning model is much better for the analysis of RUL than the results obtained using traditional machine learning methods. Basically, the RMSE value of all those who use the deep learning methods to predict the RUL is less than 10. However, the results obtained by traditional machine learning methods in the same data settings are not satisfactory. The Transformer model has more oscillations, but the overall trend is around the actual RUL value. LSTM, RNN, and GRU have similar prediction results which are larger than the actual RUL value with GRU's prediction for the test set fluctuating more sharply in the second half. Among these three models used most frequently in RUL which are RNN, LSTM, and GRU, LSTM is the best. The result of RUL prediction using the one-dimensional CNN method is worse than LSTM, but the effect is slightly better when CNN is combined with LSTM whose RMSE is 7.0081 and the RMSE of LSTM is 7.1399. The BiLSTM used to test data has the best result which is 4.1996. After comparison, the PSR-former method is the most effective approach for bearing RUL prediction among those methods under the same data settings.

Conclusions

Due to the complexity of the mechanical working environment, the data collected can be easily affected. A PSR-former remaining useful life prediction method was proposed in response to the situation. The contributions of this paper are summarized below:

- (1) To better reflect the degradation trend of the bearings, features chosen by the monotonicity were used to form a new enhanced HI index.
- (2) A PSR-former model was proposed including a PSR layer and a Transformer layer. PSR layer was used as an embedding to deepen the understanding of the characteristics. The Transformer was used to discriminate the RUL. A new layer-hopping was adopted in the self-attention structure in the PSR-former model to speed up the propagation and make the structure more stable.
- (3) The method was verified using IMS bearing dataset and compared with other deep learning methods and traditional machine learning algorithms. The effective implementation of the methods provides a theoretical basis for the data analysis in PHM.

However, the applicability of the model in other datasets needs further verification since this paper only analyzed the IMS bearing dataset. Time domain features were extracted in this manuscript to initially enhance the degradation trend and form a new health index. Other features in the time domain or frequency domain may also be extracted to build the degradation model, so it needs to be further explored in the future.

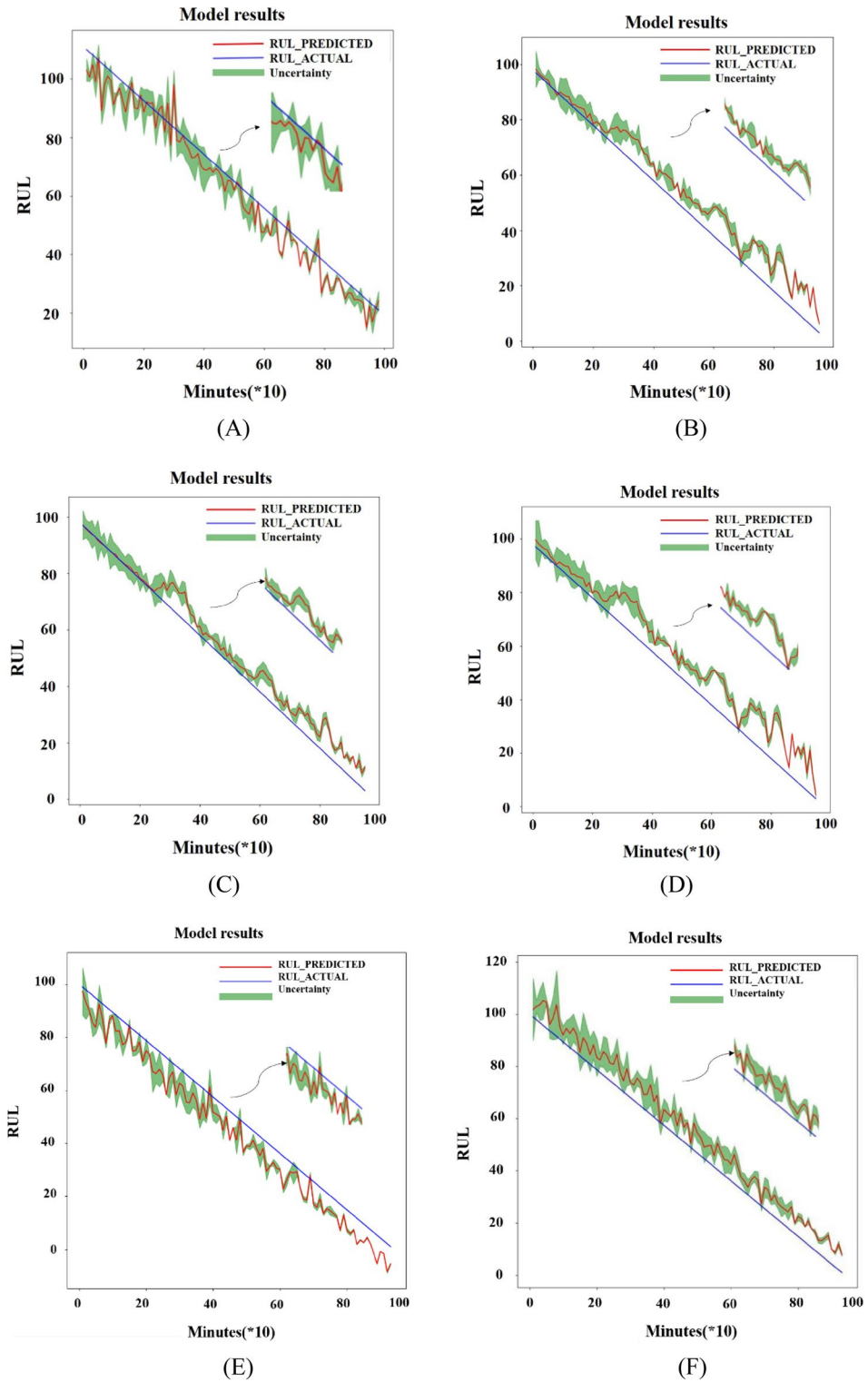


Figure 16. Visualization of RUL prediction results of different comparison algorithms. (A) Transformer; (B) RNN; (C) LSTM; (D) GRU; (E) 1-D CNN; (F) CNN-LSTM; (G) RF; (H) SVM; (I) SAE-LSTM; (J) BiLSTM.

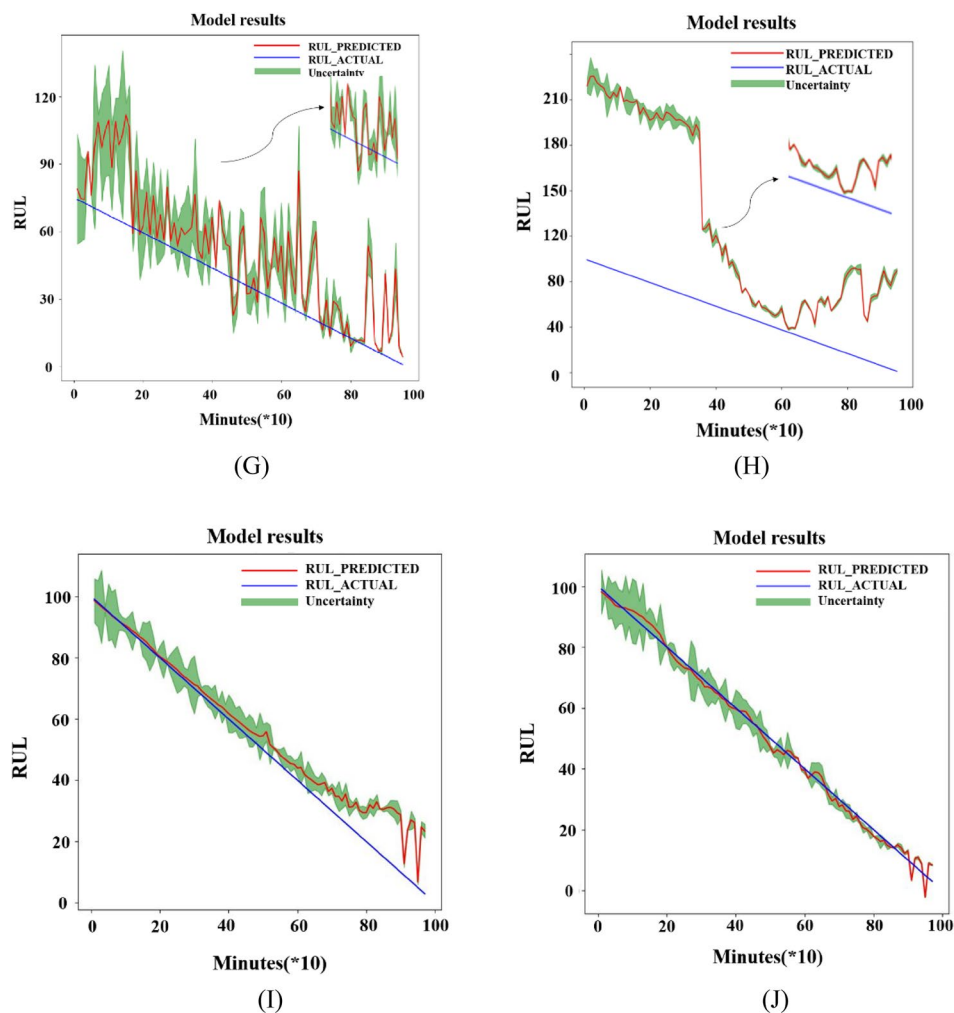


Figure 16. (continued)

Methods	(1) Transformer	(2) RNN	(3) LSTM	(4) GRU
Deep learning	6.9779	9.8762	7.1399	10.1544
	(5) 1D-CNN	(6) CNN-LSTM	(7) SAE-LSTM	(8) BiLSTM
	8.2647	7.0081	8.0087	4.1996
Traditional machine learning	(9) Random forest	(10) Support vector machine	(11) Linear regression	(12) Logistic regression
	50.6241	93.7156	1083.1913	3828.8760
	(13) K-nearest neighbours	-	-	-
(14) Proposed method	1.0311	-	-	-

Table 3. RMSE results of different methods under the same data settings. (1) Transformer; (2) RNN; (3) LSTM; (4) GRU; (5) 1D-CNN; (6) CNN-LSTM; (7) SAE-LSTM; (8) BiLSTM; (9) RF; (10) SVM; (11) LR; (12) LOR; (13) KN; (14) proposed method.

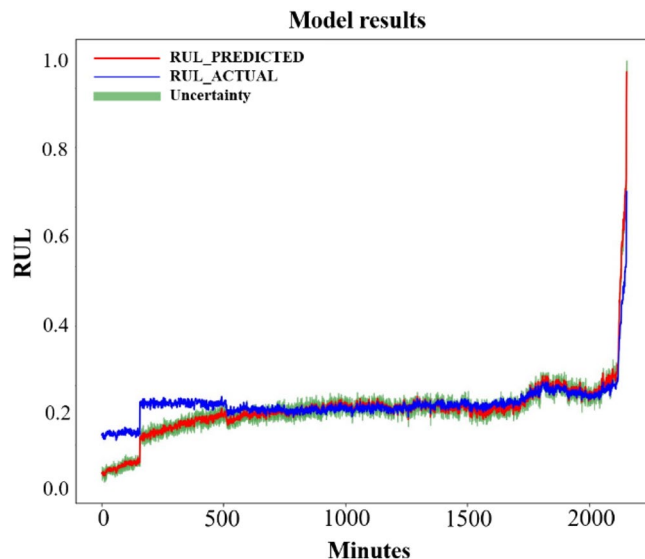


Figure 17. Visualization of retest results using bearing3 data.

Data availability

The data that support the findings of this study are openly available in NASA Intelligrnt Systems Division at <http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/>.

Received: 24 June 2022; Accepted: 21 October 2022

Published online: 25 October 2022

References

- Berghout, T. & Benbouzid, M. A systematic guide for predicting remaining useful life with machine learning. *Electronics* **11**(7), 1125. <https://doi.org/10.3390/electronics11071125> (2022).
- Lei, Y. *et al.* Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process.* **104**, 799–834. <https://doi.org/10.1016/j.ymssp.2017.11.016> (2017).
- Nan, X. A., Rui, K. B. & Yy, A. Tool condition prognostic model based on digital twin system. *Proc. CIRP* **93**, 1502–1507. <https://doi.org/10.1016/j.procir.2020.03.045> (2020).
- Qiu, J. *et al.* Damage mechanics approach for bearing lifetime prognostics. *Mech. Syst. Signal Process.* **16**(5), 817–829. <https://doi.org/10.1006/mssp.2002.1483> (2002).
- Guo, R. & Wang, Y. Remaining useful life prognostics for the rolling bearing based on a hybrid data-driven method. *Proc. Inst. Mech. Eng. Part I J. Syst. Control Eng.* **235**(3), 095965182094828. <https://doi.org/10.1177/0959651820948284> (2020).
- Ravi, V. R. *et al.* Remaining serviceable life indicator of bearing. *Mater. Today Proc.* **33**(7), 3951–3955. <https://doi.org/10.1016/j.matpr.2020.06.306> (2020).
- Qin, Y., Wu, X. & Luo, J. Data-model combined driven digital twin of life-cycle rolling bearing. *IEEE Trans. Industr. Inf.* **18**(3), 1530–1540. <https://doi.org/10.1109/TII.2021.3089340> (2022).
- Xu, G. *et al.* High-speed train wheel set bearing fault diagnosis and prognostics: A new prognostic model based on extendable useful life. *Mech. Syst. Signal Process.* **146**, 107050. <https://doi.org/10.3901/JME.2016.22.045> (2021).
- Gabelli, A. & Morales-Espejel, G. E. A model for hybrid bearing life with surface and subsurface survival. *Wear* **422–423**, 223–234. <https://doi.org/10.1016/j.wear.2019.01.050> (2019).
- Li, X., Zhang, W. & Ding, Q. Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliab. Eng. Syst. Saf.* **182**, 208–218 (2019).
- Liu, C., Zhang, L. & Wu, C. Direct remaining useful life prediction for rolling bearing using temporal convolutional networks. *IEEE Sympos. Ser. Comput. Intell.* **2019**, 2965–2971. <https://doi.org/10.1109/SSCI44817.2019.9003163> (2019).
- Cui, Q., Li, Z., Yang, J. & Liang, B. Rolling bearing fault prognosis using recurrent neural network. in *2017 29th Chinese Control and Decision Conference (CCDC), 2017*. 1196–1201. <https://doi.org/10.1109/CCDC.2017.7978700> (2017).
- Liu, J. *et al.* Fault prediction of bearings based on LSTM and statistical process analysis. *Reliab. Eng. Syst. Saf.* **214**(4), 107646. <https://doi.org/10.1016/j.res.2021.107646> (2021).
- Zainuddin, Z., Akhir, E. A. & Hasan, M. H. Predicting machine failure using recurrent neural network-gated recurrent unit (RNN-GRU) through time series data. *Bull. Electric. Eng. Inform.* **10**(2), 870–878. <https://doi.org/10.11591/eei.v10i2.2036> (2021).
- Levent, E. Bearing fault detection by one-dimensional convolutional neural networks. *Math. Probl. Eng.* **2017**, 1–9. <https://doi.org/10.1155/2017/8617315> (2017).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. Attention is all you need. in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. arXiv e-prints: [arXiv:1706.03762](https://arxiv.org/abs/1706.03762). (Curran Associates Inc., 2017).
- Ding, Y. *et al.* A novel time-frequency transformer and its application in fault diagnosis of rolling bearings. *Mech. Syst. Signal Process.* **168**, 108616. <https://doi.org/10.1016/j.ymssp.2021.108616> (2021).
- Alexakos, C. T., Karnavas, Y. L., Drakaki, M. & Tzifafettas, I. A. A combined short time Fourier transform and image classification transformer model for rolling element bearings fault diagnosis in electric motors. *Mach. Learn. Knowl. Extract.* **3**(1), 228–242. <https://doi.org/10.3390/make3010011> (2021).
- He, Z., Shao, H., Ding, Z., Jiang, H. & Cheng, J. Modified deep autoencoder driven by multisource parameters for fault transfer prognosis of aeroengine. *IEEE Trans. Industr. Electron.* **69**(1), 845–855. <https://doi.org/10.1109/TIE.2021.3050382> (2021).

20. Ding, H. *et al.* A remaining useful life prediction method for bearing based on deep neural networks. *Measurement* **172**, 108878. <https://doi.org/10.1016/j.measurement.2020.108878> (2020).
21. Liao, L. Discovering prognostic features using genetic programming in remaining useful life prediction. *IEEE Trans. Industr. Electron.* **61**(5), 2464–2472. <https://doi.org/10.1109/TIE.2013.2270212> (2013).
22. Han, M. *et al.* Nonuniform state space reconstruction for multivariate chaotic time series. *IEEE Trans. Cybern.* **49**(5), 1885–1895. <https://doi.org/10.1109/TCYB.2018.2816657> (2019).
23. Takens, F. *Detecting Strange Attractors in Turbulence*. <https://doi.org/10.1007/BFb0091924> (Springer, 1981).
24. Cao, L. Practical method for determining the minimum embedding dimension of a scalar time series. *Phys. D Nonlinear Phenomena* **110**(1–2), 43–50. [https://doi.org/10.1016/S0167-2789\(97\)00118-8](https://doi.org/10.1016/S0167-2789(97)00118-8) (1997).
25. Cao, Y. *et al.* A novel temporal convolutional network with residual self-attention mechanism for remaining useful life prediction of rolling bearings. *Reliab. Eng. Syst. Saf.* **215**, 107813. <https://doi.org/10.1016/j.ress.2021.107813> (2021).
26. Lee, J., Qiu, H., Yu, G., & Lin, J. *Rexnord Technical Services, NSF I/UCR Center for Intelligent Maintenance Systems, Milwaukee*. <http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/>.
27. Gal, Y., & Ghahramani, Z. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. *arXiv e-prints: arXiv:1506.02142* (2015).
28. Shamsi, A., Asgharnezhad, H., Abdar, M., *et al.* *Improving MC-Dropout Uncertainty Estimates with Calibration Error-based Optimization*. *arXiv e-prints: arXiv:2110.03260* (2021).
29. Li, Q. & Yang, Y. Transformer model for remaining useful life prediction of aeroengine. *J. Phys. Conf. Ser.* **2171**(1), 012072 (2022).
30. Gjylapi, D. & Proko, E. Recurrent neural networks in time series prediction. *New Res. Adv. Comput. Sci. Inf. Technol. NRACSIT-2017* **5**(10), 2458–9403 (2017).
31. Wang, X., Guo, J., Wang, J., Liu, C. & Du, C. Prediction of bearing remaining useful life based on LSTM network. *Mech. Eng. Mater.* **100**, 81–90. https://doi.org/10.1007/978-3-030-68303-0_7 (2021).
32. Wu, J. *et al.* Ensemble recurrent neural network-based residual useful life prognostics of aircraft engines. *SDHM Struct. Durab. Health Monit.* **13**(3), 317–329. <https://doi.org/10.32604/sdhm.2019.05571> (2019).
33. Qin, Y., Cai, N., Gao, C., *et al.* *Remaining Useful Life Prediction Using Temporal Deep Degradation Network for Complex Machinery with Attention-Based Feature Extraction*. *arXiv eprint arXiv:2202.10916* (2022).
34. Abdelli, K., Griefser, H., & Pachnicke, S. A hybrid CNN-LSTM approach for laser remaining useful life prediction. in *2021 Opto-Electronics and Communications Conference (OECC)*, 2021. 1–3. *arXiv eprint arXiv:2203.12415* (2021).
35. Pal, M. & Parija, S. Prediction of heart diseases using random forest. *J. Phys. Conf. Ser.* **1817**(1), 012009 (2021).
36. Huang, H. Z. *et al.* Support vector machine based estimation of remaining useful life: Current research status and future trends. *J. Mech. Sci. Technol.* **29**(1), 151–163. <https://doi.org/10.1007/s12206-014-1222-z> (2015).
37. Acharya, M. S., Armaan, A. & Antony, A. S. A comparison of regression models for prediction of graduate admissions. *Int. Conf. Comput. Intell. Data Sci. (ICCIDS)* **2019**, 1–5. <https://doi.org/10.1109/ICCIDS.2019.8862140> (2019).
38. Zhang, X. & Chen, X. Research on breach prediction for big data through hybrid ensemble learning and logistic regression. *J. Phys. Conf. Ser.* **1982**(1), 012049. <https://doi.org/10.1088/1742-6596/1982/1/012049> (2021).
39. Lin, G., Lin, A. & Gu, D. Using support vector regression and K-nearest neighbors for short-term traffic flow prediction based on maximal information coefficient. *Inf. Sci.* **608**, 517–531. <https://doi.org/10.1016/j.ins.2022.06.090> (2022).
40. Tian, H. A., Jp, A. & Acct, B. Remaining useful life prediction of bearing based on stacked autoencoder and recurrent neural network. *J. Manuf. Syst.* **61**, 576–591. <https://doi.org/10.1016/j.jmsy.2021.10.011> (2021).
41. Zhih, Z., Jingj, Z. & Ziy, W. Rolling bearing fault diagnosis based on BiLSTM network. *Vib. Shock* **40**(01), 95–101. <https://doi.org/10.13465/j.cnki.jvs.2021.01.013> (2021).

Author contributions

All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by H.Z., S.Z., L.Q., Y.Z., Y.W., Z.W., and G.Y. The first draft of the manuscript was written by H.Z. and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This study were funded by National Natural Science Foundation of China (Grant No. 51875516) and Zhejiang Provincial Natural Science Foundation of China (Grant No. LQ22E050005).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022