Bearing Remaining Useful Life Prediction based on TSDAE and Pathformer-TCLSTM

Guanghua Fu, Yujie Yang, Yonghui Liu, Xuegen Wang

Institute of Logistics Science and Engineering, Shanghai Maritime University, Shanghai 201306, China

ghfu@shmtu.edu.cn, 202330510056@stu.shmtu.edu.cn, 202330510022@stu.shmtu.edu.cn, 1910566541@qq.com

ABSTRACT

The remaining useful life (RUL) prediction of bearings is crucial for the stable operation and effective maintenance. Conventional RUL prediction approaches extract the restricted features that would affect the prediction results, and the computational efficacy is often influenced by the redundancy of the features and domain knowledge. To address these problems, this paper proposes a RUL prediction approach mainly based on temporal sparse denoising autoencoders (TSDAE) for feature selecting, and Pathformer-temporal convolutional long short-term memory (Pathformer-TCLSTM) for predicting. Firstly, the original signal is denoised via wavelet thresholding. Subsequently, the denoised signal is decomposed using empirical mode decomposition (EMD) to extract the features of the time-domain, frequency-domain, and time-frequency domain to resolve the problem of restricted features. Moreover, the TSDAE feature selection technique is implemented to eliminate redundant features and address the limitation of domain knowledge utilized in traditional feature selection. Finally, the Pathformer-TCLSTM model is adopted for RUL prediction, which captures the multi-scale global information, local information, and long-range dependency. The validation on the PHM2012 and XJTU-SY bearing datasets shows that the proposed model has satisfactory predictive performance.

1. INTRODUCTION

The security of mechanical equipment is essential for the operation of industrial production. Rolling bearings are critical to the mechanical system and are utilized extensively in a variety of rotating equipment (Carlos & Gil, 2022; Zhu et al., 2023). The bearings of mechanical equipment frequently operate in harsh working environments, and their operational states are affected by various factors and are prone to deterioration and failure. Therefore, accurately predicting the RUL can ensure the stable operation of mechanical systems (Dong et al., 2023). Additionally, RUL prediction of bearings can mitigate safety risks (Yang et al., 2025; Yu et al., 2021).

Approaches to RUL prediction can be categorized into two groups: physical model-based approaches and data-driven approaches (Medjaher et al., 2012; Zhu et al., 2020). Approaches in the first group encompass the Lundberg-Palmgren model (Paris & Erdogan, 1963) and the Ioannides-Harris model (Ioannides & Harris, 1985). Their application in real engineering systems is a great challenge, as it depends on the complexity of real-world scenarios and extensive expert knowledge.

In contrast, approaches in the second group do not consider the physical mechanisms, wear patterns, or progression of failure. Thus, the characteristics mentioned above have led to the development of data-driven techniques in mechanical RUL prediction. For example, Li et al. (2022) proposed an approach based on multi-support vector regression fusion and adaptive weight update for RUL prediction. Similarly, Alfarizi et al. (2022) proposed a RUL prediction approach that employed an optimized random forest model. However, these approaches heavily depend on manual feature selection, which might limit the accuracy of RUL prediction.

Data-driven approaches based on deep learning have proven to be more effective in feature extraction and nonlinear computation (Barraza-Barraza et al., 2020; Najdi et al., 2025; wahhab Lourari et al., 2024; Wang et al., 2020). Yao et al. (2021) proposed a RUL prediction approach for rolling bearings, which is based on an improved one-dimensional (1-D) convolutional neural networks (CNNs) and simple recurrent units. Guo et al. (2017) adopted a bearing RUL prediction technique based on recurrent neural networks. Zhang et al. (2020) proposed a long short-term memory (LSTM) network based on attention for rotatory machine

https://doi.org/10.36001/IJPHM.2025.v16i2.4297

remaining useful life prediction; this approach overcame the limitations of traditional machine learning algorithms in dealing with complex nonlinear signals. Cao et al. (2021) proposed a novel temporal convolutional network with a residual self-attention mechanism for remaining useful life prediction of rolling bearings, this approach can learn both time-frequency and temporal information of signals. Gao et al. (2025) proposed a multiscale spatiotemporal attention network for remaining useful life prediction of mechanical systems. Ye et al. (2024) proposed an adaptive multi-adaptive graph neural networks with temporal convolutional networks for bearings remaining useful life prediction. However, the gradient may vanish during the backpropagation process due to the long-range dependency problem throughout the training of RNNs; The temporal convolutional network (TCN) lacks the capacity to extract the global information embedded within time series; And the LSTM has limited extraction of local information within time series. Moreover, the extraction of single-domain features could not provide comprehensive information for RUL prediction, which may influence its accuracy.

The Transformer is a popular data-driven deep learning approach, which has substantially contributed to the prediction of time series in recent years, and the model has been employed for RUL prediction. Ding & Jia (2022) proposed a convolutional Transformer for bearing RUL prediction. Peng et al. (2023) proposed a local enhancement Transformer RUL prediction approach based on the temporal convolutional attention mechanism. Chen et al. (2024) proposed multi-scale transformers with adaptive pathways for time series forecasting, it integrated both temporal resolution and temporal distance for multi-scale modeling. However, the Transformer model may struggle with capturing local information, which restricts their RUL prediction accuracy; The Pathformer lacks the ability to extract the long-range dependency and local information within time series, which may influence the accuracy of time series forecasting.

Feature selection aims to identify the optimal subsets that are suitable for model training. This approach can reduce information loss and minimize the decline in learning performance, thereby enhancing the efficiency of RUL prediction (Li, 2017). Feature selection approaches could be categorized into three types (filter approach, wrapper approach, and embedded approach) (Atashgahi et al., 2022). Recently, many deep learning-based models have been developed to select features to improve the RUL prediction performance. Y. Wang et al. (2022) proposed a remaining useful life prediction of rolling bearings based on Pearson correlation-KPCA multi-feature fusion, it utilized the Pearson correlation analysis to select the features for RUL prediction. Atashgahi et al. (2022) proposed a sparse denoising autoencoder for feature selection, the sparsity utilized in sparse denoising autoencoder can reduce the complexity of the original data. However, techniques for computing the feature-target correlation might incur high computational complexity and inefficiency with the growth of feature dimensions; Additionally, the feature selection approaches mentioned above may rely on domain knowledge, which may influence the accuracy of RUL prediction; The sparse denoising autoencoder-based feature selection technique has neglected the long-range dependency of time series.

Although the techniques mentioned above provide a potential solution to bearing RUL prediction, the available features for RUL prediction in existing research are restricted, which in turn affects the results of RUL prediction (Motahari-Nezhad & Jafari, 2021). Additionally, the predictive efficacy could be affected by the presence of redundancies within the feature sets (Li et al., 2019). Traditional feature selection works may rely on domain knowledge and have higher computational complexity. The transformer model has the advantage of capturing global information, but it disregards local details (Vaswani et al., 2017). Furthermore, the inefficiency of 1-D CNNs to predict time series and the problem of gradient vanishing in RNNs could not be disregarded (Ren, Sun, Wang, et al., 2018). Moreover, the TCN cannot extract the global information within time series; The LSTM has limited extraction of local information within time series; The Pathformer cannot extract the long-range dependency and local information within time series. Finally, the sparse denoising autoencoder-based feature selection approach is unable to extract the long-range dependency within time series.

To address the problems mentioned above, this paper proposed a bearing RUL prediction approach based on TSDAE feature selection and Pathformer-TCLSTM. In response to the problem of restricted features, this paper applied EMD to the original dataset. Subsequently, the features of the time-domain, frequency-domain, and time-frequency domain were extracted from EMD. To address feature redundancy, domain knowledge utilized in traditional feature selection, and computational complexity, the TSDAE feature selection strategy is adopted to select the extracted features and obtain the optimal feature subsets. To tackle the lack of local information and long-range dependency in Pathformer, the inefficiency of TCN in extracting the global information, and the limitation of LSTM in extracting the local information within time series, a parallel Pathformer-TCLSTM prediction model was implemented. This parallel model aims to capture multi-scale global information, local information, and long-range dependency in time series. Finally, the proposed model was validated with the PHM2012 bearing degradation dataset and XJTU-SY bearing degradation dataset. The contribution of this paper can be summarized as follows:

1) The time-domain, frequency-domain, and time-frequency domain features are extracted from the EMD as input

features, which can supply more features for model training to improve the accuracy of bearing RUL prediction.

2) Temporal sparse denoising autoencoders (TSDAE) strategy is adopted for feature selection, which aims to select the optimal subsets and enhance computational efficiency. Previous feature selection works have often relied on domain knowledge and had higher computational complexity. Our module can autonomously learn intricate linear relationships between input features and targets, which can avoid parameter adjustment.

3) The Pathformer-temporal convolutional long short-term memory (Pathformer-TCLSTM) parallel network is adopted to extract the complementary features and improve the accuracy of RUL prediction, where Pathformer extracts the multi-scale global features, temporal convolutional extracts the local features, and long short-term memory extracts the long-range dependency within time series.

This paper can be divided into five sections: section 2 presents the theoretical framework of the proposed model, section 3 introduces the concept and procedural steps of the methodology, section 4 validates the efficacy and superiority of the proposed model, and section 5 draws conclusions of this paper.

2. THE PROPOSED METHOD FOR RUL PREDICTION

2.1. Pathformer

Pathformer is a multi-scale time series prediction model that incorporates a Transformer architecture (Chen et al., 2024). This model aims to improve the ability to model comprehensive multi-scale time series by incorporating temporal resolution and temporal distance from various perspectives. The principle necessitates the collaborative operation of a multi-scale router and an aggregator to adaptively extract dynamic characteristics from input time series, thereby accomplishing adaptive multi-scale modelling. The structure of the Pathformer and the multi-scale router are illustrated in Figure 1. First of all, the multi-scale router adaptively selects specific sizes of patch from time series, and allocates the specific weight to each patch. Then, the multi-scale division of different multi-scale Transformer blocks divides these patches to obtain new patches with different time resolutions. Meanwhile, the multi-scale Transformer block is performed over these new patches to extract features. After that, the multi-scale aggregator is utilized to integrate information from different paths. There are three cascading adaptive multi-scale blocks utilized in Pathformer to extract features. Finally, the predicted results are obtained from a predictor.

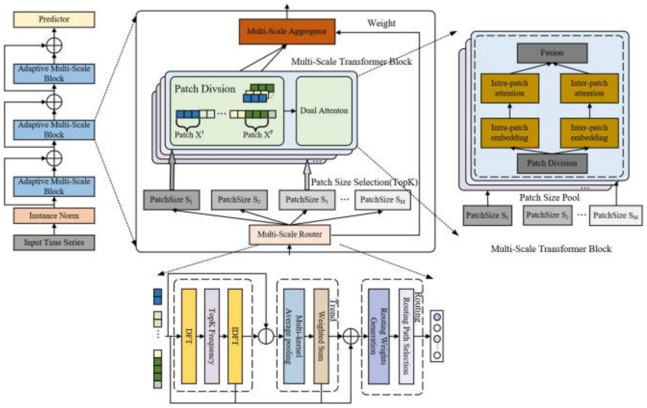


Figure 1. The structure of Pathformer.

2.2. LSTM

Long short-term memory (LSTM) is a particular type of recurrent neural networks. It is notably well-suited for a variety of tasks, including language modelling, and is capable of effectively capturing long-range dependency (Ma & Mao, 2020). Furthermore, LSTM can solve the problem of gradient expansion in recurrent neural networks. The LSTM network diagram is shown in Figure 2.

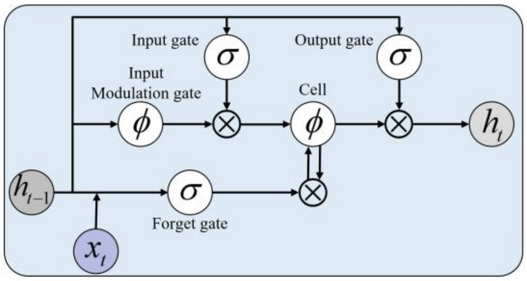


Figure 2. The structure of LSTM.

The LSTM primarily comprises an input gate, a forget gate, an output gate, and a cell state (Saufi & Hassan, 2021). The forget gate determines whether to retain or discard information. Furthermore, the Sigmoid activation function is used to convey the current input information and the previous hidden state. The retention of information is determined by output values that are near 1, while the loss of information is indicated by output values that are near 0. The input gate determines the extent to which new information is incorporated into the current cell state. The output gate regulates the amount of cell state information transferred to the hidden state. The cell state serves as a long-term memory repository, selectively retaining and propagating critical historical information through sequential time steps.

2.3. TCN

A temporal convolutional network (TCN) is a time series prediction model that typically captures latent temporal dependencies within the input time series (Bai et al., 2018). In contrast to RNNs, TCN models utilize causal convolution rather than recurrent connections, which allows for parallel data processing (Wang et al., 2023). Its architecture is shown in Figure 3.

The input layer and output layer in 1-D fully convolutional networks are similar to 1-D convolution, while each hidden layer has the same time length as the input layer.

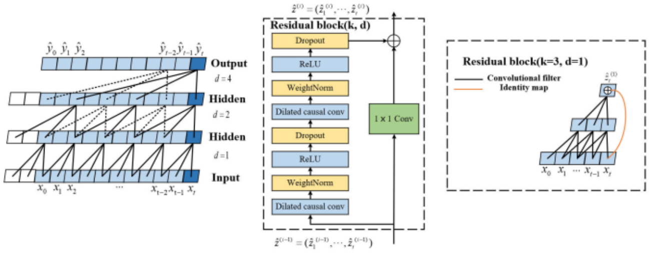
The causal convolution is utilized in TCN, where the output at time t convolves with elements at time t. However, an immense amount of computational complexity may be caused by the linear relationship between the acceptive field of output information and filter size, network depth. In order to capture the long-range historical relationships, it is imperative to augment the number of network layers. Consequently, the dilated convolution is employed to ensure that the receptive field of output information is proportional to the number of layers (Qiu et al., 2023).

For a 1-D input sequence $x \in \mathbb{R}$ and a filter $f: \{0, \dots, k-1\} \rightarrow \mathbb{R}$, the dilated convolution operation on s can be defined as follows (Yang et al., 2020).

$$F(s) = (x *_d f)(s) = \sum_{i=1}^{k-1} f(i) \cdot x_{s-d \cdot i} \quad (1)$$

where d represents the dilation factor, k is the size of the filter, $s - d \cdot i$ denotes the direction in the past. As the number of layers increases, the degree of swelling also expands. The output at the top level in TCN can effectively broaden the input acceptance range of the convolution network, encompassing a wider spectrum of inputs. The significance of field perception in time series modelling is derived from its inherent limitation in capturing periodic characteristics within a specific layer. Figure 3 (a) depicts an extended causal convolution with an expansion factor $d = 1{,}24$, and filter size $k = 3$.

Residual connection is employed in TCN to address the problem of gradient vanishing. The structure of the residual block is shown in Figure 3 (b). The residual block of TCN utilizes the ReLU as the activation function, and it employs two extended causal convolutions and two nonlinear activation layers. Additionally, the weight normalization and dropout layer are implemented to expedite the model training process, achieving satisfactory generalization performance. Figure 3 (c) illustrates an instance of a residual connection within a TCN. The black line represents the filter function within the residual, while the orange line depicts the identity mapping.


$\hat{y}_0$ $\hat{y}_1$ $\hat{y}_2$ $\hat{y}_{t-2}$ $\hat{y}_{t-1}$ $\hat{y}_t$
Output
$d=4$
Hidden
$d=2$
Hidden
$d=1$
Input
x_0 x_1 x_2 … x_{t-2} x_{t-1} x_t
$\hat{z}^{(i)} = (\hat{z}_1^{(i)}, \cdots, \hat{z}_t^{(i)})$
Residual block(k, d)
Dropout
ReLU
WeightNorm
Dilated causal conv
Dropout
ReLU
WeightNorm
Dilated causal conv
1 x 1 Conv
$\hat{z}^{(i-1)} = (\hat{z}_1^{(i-1)}, \cdots, \hat{z}_t^{(i-1)})$
Residual block(k=3, d=1)
Convolutional filter
Identity map
$\hat{z}_t^{(i)}$
x_0 x_1 … x_{t-1} x_t

(a) (b) (c)

Figure 3. The structure of TCN.

2.4. EMD

(Empirical mode decomposition) EMD is an adaptive analytical approach that could accommodate non-linear and non-stationary signals. Compared with other methods (e.g., ensemble empirical mode decomposition and variational mode decomposition), EMD demonstrates a faster decomposition speed. EMD involves the decomposition of a non-stationary time series into a set of intrinsic mode function (IMF) and residual (Guo et al., 2021). Additionally, the original signal can be retrieved by aggregating the residual values and IMF. The various frequency components extracted via EMD represent distinct information within the signal.

1) Calculate the average of the upper envelope and the lower envelope.

$$m(t) = \frac{E_{max}(t) + E_{min}(t)}{2} \quad (2)$$

where $E_{max}(t)$ denotes the upper envelope of the original input data $x(t)$, $E_{min}(t)$ represents the lower envelope.

2) Calculate the difference $h(t)$ between the original input data $x(t)$ and the average envelope $m(t)$.

$$h(t) = x(t) - m(t) \quad (3)$$

3) Determine whether $h(t)$ satisfies the constraint conditions of IMF.

If $h(t)$ is not an IMF, $x(t) = h(t)$. Repeat 1) and 2) until the condition is satisfied.

If $h(t)$ is an IMF, $h(t)$ will be the first component of IMF:

$$c_1(t) = h(t) \quad (4)$$

4) Obtain the residual component $r_1(t)$ of $x(t)$.

$$r_1(t) = x(t) - c_1(t) \quad (5)$$

5) The revised sequence of input is represented by $r_1(t)$, then repeat the previously mentioned steps. Until emergence of the second IMF $c_2(t)$, iterate the previously mentioned process n times. This process can be depicted as:

$$x(t) = \sum_i^n c_\mathrm{i}(t) + r_\mathrm{n}(t) \quad (6)$$

where $c_{\mathrm{i}}(t)$ represents the i-th IMF, $r_{\mathrm{n}}(t)$ denotes the residual. The updated representation of (13) is formulated as follows.

$$x(t) = \sum_{i}^{n} IMF_{\mathrm{i}}(t) + r_{\mathrm{n}}(t) \qquad (7)$$

2.5. Sparse auto-encoder

The sparse auto-encoder (SAE) is the variant of (Auto-encoder) AE, which adopts the sparsity constraint to reduce the complexity (Sun et al., 2019). The AE is an unsupervised neural network, and it consists of an encoder and a decoder (Yang et al., 2016).

The main process of SDAE is as follows:

1) Add noise to the original input data to generate corrupted input data. The purpose of adding noise to the raw input data is to enhance model robustness.

2) The encoder extracts the information from corrupted input data to reduce the dimensional space.

3) The decoder processes the hidden layer information to obtain the output data. The purpose of this stage is to reconstruct clean data from the learned feature representations, thereby effectively eliminating noise.

4) Calculate the error between input data and output data to optimize model parameters.

5) Incorporating sparse penalties into the reconstruction error minimizes neuronal activity. Introducing a penalty term serves to reduce overfitting while improving model generalization.

2.6. RUL prediction based on TSDAE and Pathformer-TCLSTM

2.6.1. RUL prediction framework

To deal with the problem of restricted features in RUL prediction of bearings, this paper applied EMD to extract features. In response to address feature redundancy, domain knowledge utilized in traditional feature selection, and computational complexity, this paper adopted the TSDAE approach to obtain the optimal feature subsets. To tackle the lack of local information and long-range dependency in Pathformer, the inefficiency of TCN in extracting the global information, the limitation of LSTM in extracting the local information within time series, the parallel Pathformer-TCLSTM prediction model was employed. The proposed RUL prediction framework is illustrated in Figure 4. It can be divided into three parts: signal denoising, feature extraction and selection, and RUL prediction. In the first part, the original signal is denoised through wavelet threshold denoising. In the second part, the signal is decomposed using EMD. Additionally, the time-domain, frequency-domain, and time-frequency domain features are extracted from EMD. The TSDAE feature selection technique is subsequently implemented to evaluate the importance of the extracted features and to determine the appropriate subsets of features for each operating condition. In the last part, the optimal feature subsets will be selected, followed by data partitioning. Moreover, the sliding window method is utilized to extract fixed-length segments of input features, which are subsequently input into the model for training. Subsequently, the feature will be fed into the Pathformer-TCLSTM network for prediction.

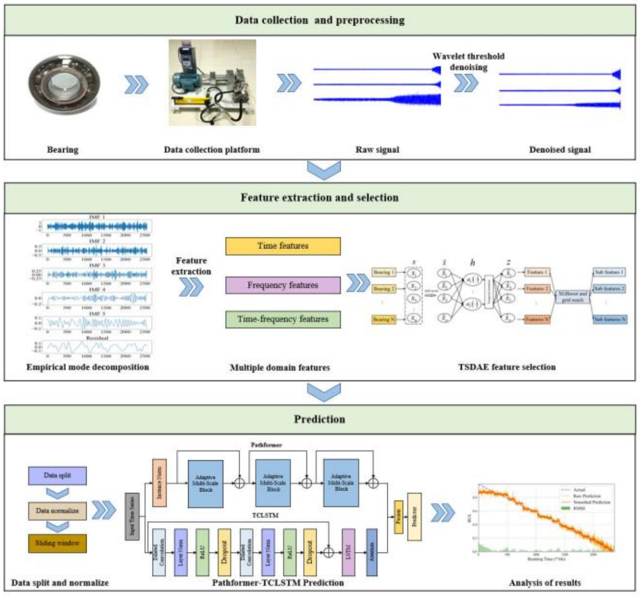


Figure 4. The framework of RUL prediction.

2.6.2. Signal denoising

The wavelet threshold denoising technique demonstrates characteristics of adaptability that can be applied in a wide range of fields. Firstly, the signal is decomposed into various frequency components using wavelet transform, followed by the application of a threshold. Then, the wavelet coefficients below the threshold are deemed as noise and assigned a value of zero, while these exceeding the threshold are preserved. Finally, the denoised signal is obtained by processing the wavelet coefficients and reconstructing them using the wavelet inverse transform (Aminou et al., 2023).

$$S(t) = Z(t) + \sigma n N(t) \qquad (8)$$

where $S(t)$ represents the noise signals, $Z(t)$ denotes the original signal, σn is the noise factor, $N(t)$ denotes the standard normal distribution of noise characterized by a mean of 0 and variance of 1. The main stages of wavelet threshold denoising are depicted as follows：

i) Sample the signal $S(t)$ at intervals of $S(i)$;

ii) Conduct a discrete wavelet transformation on $S(i)$;

iii) Set the threshold for wavelet coefficients $\lambda = \sqrt{2\sigma_n^2 log_n}$, n represents the length of $S(i)$;

iv) Compute the inverse discrete wavelet transform.

A set of trials confirmed in this article suggest that the soft threshold approach should be used and the wavelet threshold should set at 0.05.

2.6.3. Feature extraction

Following the EMD decomposition of the samples, three types of feature extraction are implemented for IMF: time-domain feature extraction, frequency-domain feature extraction, and time-frequency domain feature extraction.

i) Time-domain feature extraction. Time-domain features can capture the fundamental signal shape, serving as a cornerstone for signal analysis. Moreover, twenty-one time-domain features (Mean, variance, median, energy, root mean square, peak factor, kurtosis, inverse hyperbolic sine, arc tangent standard deviation, skewness, coefficient of skewness, gap factor, kurtosis factor, standard deviation, maximum value, minimum value, crest,peak-to-peak, inverse hyperbolic sine standard deviation, pulse factor, and absolute mean amplitude) utilized in this study.

ii) Frequency-domain feature extraction. Frequency-domain features reveal the frequency components of the signal, which are essential for the analysis of periodic, vibrational, and other related issues. This paper implemented the Fourier transform on the original signal (Ren, Sun, Cui, et al., 2018; Zhao et al., 2021), thereby transforming it from the time-domain to the frequency-domain. Subsequently, relevant features are computed within the frequency-domain. Moreover, eight frequency-domain features (Mean, kurtosis, ratio, coefficient of variation, standard deviation, center of gravity, skewness, root mean square in frequency-domain) are utilized in this paper.

iii) Time-frequency domain feature extraction. Time-frequency domain features can capture both temporal and spectral characteristics of a signal, providing significant utility in the analysis of non-stationary, transient, or multi-component signals. Discrete wavelet transform is employed to convert the time-domain signal into the time-frequency domain representation. Subsequently, the corresponding time-frequency characteristics are obtained by computing the coefficients of each wavelet. Moreover, eight time-frequency domain features (Energy, standard deviation, mean, maximum value, minimum value, variance, skewness, and median in the time-frequency domain) are utilized in this paper.

The dimension of time-domain, frequency-domain, and time-frequency domain features extracted from the EMD is 185. Thus, extracting features from various frequencies of the EMD facilitates the acquisition of comprehensive feature sets.

2.6.4. TSDAE feature selection

A feature selection approach of sparse denoising autoencoder is employed to measure the importance of features through the strength of neurons in a sparse neural network (Atashgahi et al., 2022). To address the deficiency of long-range dependency in sparse denoising autoencoder, this paper adopted TSDAE feature selection. TSDAE utilized the sparsity to selectively activate neurons in the hidden layer, thereby accelerating model training and enhancing its generalization capability. TSDAE can learn the relationship between input features and targets via model training, thereby addressing the reliance of traditional methods on domain-specific knowledge.

This structure employs a temporal attention mechanism after the hidden layer to obtain long-range dependency relationships between time series. Furthermore, the ExtraTreesClassifier of the original evaluation technique in the sparse denoising autoencoder above is substituted by (Extreme Gradient Boosting) XGBoost for the regression prediction. Ultimately, the grid search technique is implemented to select the optimal subsets.

At first, the raw features and target under a single bearing are input into the TSDAE module to learn the correlation between features and target. After that, the TSDAE outputs the importance ranking of features under a single bearing according to the strength of neurons in a sparse neural network. Moreover, the grid search and XGBoost are utilized to determine which combinations of output importance ranking features under single bearing are optimal. Next, the optimal features of all bearings under a single condition will be selected. Ultimately, we calculate the frequency of each raw feature among all the optimal features selected from the grid search and XGBoost, and each raw feature with high frequency is considered to be the best training feature. TSDAE can reduce the computational complexity and improve the efficiency of machine learning algorithms in RUL prediction. The structure of the TSDAE features selection is depicted in Figure 5. The implementation stages of TSDAE feature selection are as follows.

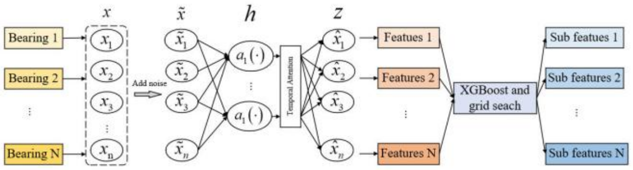


Figure 5. The structure of the TSDAE feature selection.

1) Initialize sparse connections between consecutive layers of neural networks to prevent overfitting.

$$P\left(W_{ij}^{l}\right) = \frac{\epsilon(n^{l-1}+n^{l})}{n^{l-1}\times n^{l}} \tag{9}$$

where ϵ is the parameter of sparse layer, n^l represents the number of neural in l-th layer, W_{ij}^l denotes the connection between neural i in $l-1$-th layer and neural j in l-th layer. And W_{ij}^l stores in sparse weight matrix $\boldsymbol{W^l}$**.**

2) Add noise to the raw input data to enhance model robustness.

$$\tilde{x} = x + nfN(\mu, \sigma^2) \tag{10}$$

where x is input data, nf represents the noise factor, $N(\mu, \sigma^2)$ denotes the Gaussian noise.

3) Reconstruct the output to obtain the preliminary results of model training. The reconstruction of output z is utilized by the information of hidden layer.

$$h = a(W^1\tilde{x} + b^1) \tag{11}$$

$$h' = TempAtten(h) \tag{12}$$

$$z = a(W^2h' + b^2) \tag{13}$$

where W^1 is the sparse weight of hidden layer, W^2 is the sparse weight of the output layer, b^1 and b^2 are bias, a represents activation function, h denotes the output of hidden layer, h' is output of temporal attention, z represents the output of the decoder.

4) Calculate loss function to adjust the parameters of model training. The mean squared error is utilized as a loss function, which is used to determine the discrepancy between the original features and the reconstructed output.

$$L_{MSE} = \| z - x \|_2^2 \tag{14}$$

5) Calculate the neural strength of input data. The relative strength of the input neurons observed from the model training determines the significance of features. The intensity of each neuron is subsequently approximated by calculating the total absolute weights of its outgoing connections.

$$s_i = \sum_{j=1}^{n^1} |W_{ij}^1| \tag{15}$$

where n^1 denotes the number of neurons in the first hidden layer, W_{ij}^1 represents the connection weights between input neural i and hidden neural j. s_i is the strength of input feature.

6) Feature selection under the single bearing. In this stage, we utilize XGBoost to assess the predictive performance of various feature sets, and employ grid search to identify the feature combination with the lowest RMSE.

The strength of all input neural under single bearing constitute a set $S_i = \{s_1, s_2, \cdots, s_n | s_i > s_j, j > i\}$. The feature set $F_i = \{f_1, f_2, \cdots, f_n | f_i > f_j, j > i\}$ corresponds to the input set S_i, the label set corresponding to F_i is $Y_i = \{y_1, y_2, \cdots, y_n\}$. F_{ik} is defined as the set composed of the top k features in F_i. Input F_{ik} into XGBoost to evaluate:

$$Y'_{ipre} = XGBoost(F_{ik}) \tag{16}$$

$$RMSE_{ik} = \sqrt{\frac{1}{n}\sum_{i=12}^{n}\left(Y_i - Y'_{ipre}\right)^2} \tag{17}$$

$$\mathrm{r} = \min\{RMSE_{i1}, RMSE_{i2}, \cdots, RMSE_{ik}\} \tag{18}$$

where Y'_{ipre} denotes the prediction value of XGBoost, $RMSE_{ik}$ represents the root mean square error, min () is operation of computing the minimum value, r represents the minimum value. One of the feature sets corresponding to r is defined as F'_{ik}.

7) Feature selection under single condition. Under each condition, we calculate the frequency of each original feature within the optimal combinations. The higher the frequency, the more important the feature is considered.

The set of all feature set F'_{ik} under single condition constitute a set $Q = \sum_{k=1}^{n} F'_{ik}$, each element is q_i. The set of all extracted features is $D = \{d_1, d, \cdots, d_n\}$. Count the frequency of d_i in Q.

$$I_{d_i}(q_i) = \begin{cases} 1, & d_i = q_i \\ 0, & d_i \neq q_i \end{cases} \tag{19}$$

$$count = \sum_{q_i \in Q} I_{d_i}(q_i) \tag{20}$$

where $I_{d_i}(q_i)$ denotes indicator function. If $d_i = q_i$, $I_{d_i}(q_i) = 1$, if $d_i \neq q_i$, $I_{d_i}(q_i) = 0$. The count is the result of frequency of d_i appears in Q. The set of all d_i corresponding to the $count$ is denoted as $D' = \{d_1, d_2, \cdots, d_k\}$. The number of elements in set M can be defined as K:

$$K = card(D') \tag{21}$$

where $card()$ denotes the operation of counting the number of elements in D'.

2.6.5. Pathformer-TCLSTM

To tackle the lack of local information and long-range dependency in Pathformer, the inefficiency of TCN in extracting the global information, and the limitation of LSTM in extracting the local information within time series, a parallel Pathformer-TCLSTM network is proposed. The Pathformer can extract the multi-scale global information, which can solve the deficiency of TCN and LSTM in global information extraction; The TCN can extract the local information, which can address the limitation of Pathformer and LSTM in local information extraction; The LSTM can extract the long-range dependency information, which can address the deficiency of Pathformer in long-range dependency information extraction. The Pathformer-TCLSTM model can be divided into two modules, the first module Pathformer aims at extracting the multi-scale global information, the second module TCLSTM aims at extracting the local information and long-range dependency. The features extracted from the TCN are then passed through the

LSTM, which is capable of capturing long-range dependencies within the local information.

The construction of Pathformer-TCLSTM is depicted in Figure 6, and the parameter of network is shown in Table 1.

(i) The first module is Pathformer. It aims to extract global multiscale information from time series data. Firstly, the input sequence is divided into several scales, and the scale-specific information is extracted by adaptive multi-scale blocks. Subsequently, global information is obtained from three cascaded adaptive multi-scale blocks. The advantage of this module lies in obtaining multi-scale global details.

(ii) The second module is a cascade of TCN and LSTM. It aims to capture local information and long-range dependency. First of all, the input sequence is fed into dilated causal convolution; Then, the normalized results are passed through a layer of ReLU activation function. Additionally, to reduce the parameters of the model, we set the parameter Dropout to 0.2. Then, the information is input into a dilated causal convolution, followed by layer normalization, and subsequently passed through a ReLU activation function, with the Dropout layer also being applied. At this stage, the original data is incrementally integrated with the extracted information via residual connections on an element-by-element basis. Subsequently, the information extracted from TCN is fed into an LSTM to capture long-range dependency. Finally, the attention mechanism is adopted to focus on the most important information in extracted features.

In the last stage, the information extracted from Pathformer is integrated with the information obtained from TCLSTM along the first dimension, then the result of RUL prediction is obtained from a linear layer.

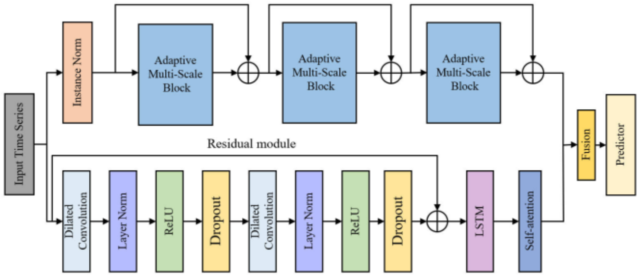


Figure 6. The framework of Pathformer-TCLSTM.

Table 1. Parameters in Pathformer-TCLSTM.

Layer	Output size	Activation function	Dropout	LN
Input	(b,91,96)	—	—	N
Pathformer block	(b,91,1536)	—	—	N
Linear	(b,1,96)	—	—	N
TCN	(b,16,96)	ReLU	0.2	Y
LSTM	(b,16,1)	—	—	N
Self-attention	(b,16,1)	—	—	N
Fusion	(b,107)	—	—	N
Linear	(b,1)	—	—	N

3. THE EXPERIMENTAL VALIDATION

To validate the generalizability and efficacy of the proposed model for RUL prediction, two experiments were conducted in the PHM2012 bearing dataset and the XJTU-SY bearing dataset via Pytorch 1.11 and run on ubuntu20.04 with an Intel (R) Xeon (R) Platinum 8457C CPU, 100GB RAM, and L20 48GB GPU.

3.1. Case study 1: IEEE-PHM-2012-Challenge

3.1.1. Dataset description

The degradation vibration data utilized in this investigation is collected from the experimental platform of the FEMMO-ST institute (Nectoux et al., 2012). This platform is illustrated in Figure 7. Horizontal and vertical acceleration sensors accumulate data throughout the experiment. Moreover, this data is collected every 10 seconds with a sampling frequency of 25.6 kHz, which results in the accumulation of 2,560 data points per sample. Furthermore, the dataset comprises the conditions of 17 bearings under three distinct scenarios (1800 rpm and 4000 N, 1650 rpm and 4200 N, and 1500 rpm and 5000 N). The effectiveness of the proposed model is verified in the horizontal direction under condition 1 and condition 2. The data of condition 1 and condition 2 are utilized in this paper. The PHM2012 degradation data for the bearings under condition 1 and condition 2 are summarized in Table 2.

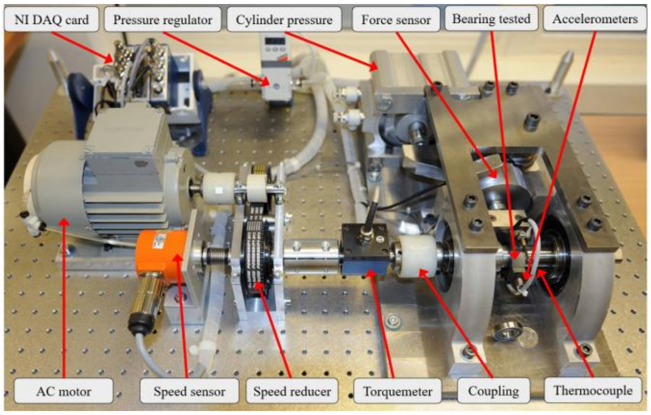


Figure 7. The FEMTO rolling bearing experimental platform.

Table 2. The information of PHM2012 degradation data.

Operating conditions	Radial force (N)	Speed (rpm)	Bearings

1	4000	1800	Learning: 1_1, 1_2 Testing: 1_3, 1_4, 1_5, 1_6, 1_7
2	4200	1650	Learning: 2_1, 2_2 Testing: 2_3, 2_4, 2_5, 2_6, 2_7

3.1.2. Normalization

The degradation data of PHM2012 bearings demonstrates the variability that occurs under various operating conditions. Additionally, we normalized the input data to reduce the difficulty of model training. The process of data normalization is depicted in Eq. (22).

$$\overline{X_t} = \frac{X_t - X_{min}}{X_{max} - X_{min}} \tag{22}$$

where X_t is the data at time t, X_{min} denotes the minimum value in X_t, X_{max} represents the maximum value in X_t.

The RUL values of bearings under various operating conditions are also normalized to simplify the model and mitigate overfitting whose process is depicted in Eq. (23) and Eq. (24).

$$Current\_RUL_t = Total\_Life - t \tag{23}$$

$$Norm\_RUL_t = \frac{Current\_RUL_t}{Total\_Life} \tag{24}$$

where $Total\_Life$ represents the total RUL of single bearing, $Current\_RUL_t$ denotes the bearing RUL at time t, $Norm\_RUL_t$ is the RUL of normalization, its value ranges from 0 to 1.

3.1.3. Evaluation metrics

The root mean square error (RMSE) and mean absolute error (MAE) are employed as metrics to evaluate the performance of RUL prediction. The RMSE is particularly sensitive to larger errors, whereas the MAE assigns equal weight to each error (Wang & Lu, 2018).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{25}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{26}$$

where Y_i represents the actual RUL, $\hat{Y}_i$ is predicted RUL.

3.1.4. Sliding window

This paper utilizes a sliding window approach to process the data. It employs diverse window sizes to extract a segment of features preceding the current moment. The different window sizes represent the information acquired over distinct future time intervals. Moreover, it can capture the changing trends of RUL within the same window size, thereby enhancing the accuracy of RUL prediction (H. Wang et al., 2022). The structure of sliding window is illustrated in Figure 8.

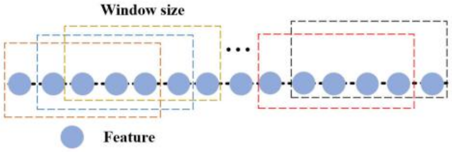


Figure 8. The structure of sliding window.

3.1.5. Results

During model training, we set the learning rate to 0.001; The patience is set to 5. The learning rate decay strategy is adopted, and the minimum learning rate is set to 0.00001. The Adam optimizer is utilized to train the model. L1 regularization was used in the experiment to prevent model overfitting. Epoch is set to 50, batch is set to 128. The window size is set to 10.

To eliminate irrelevant features and improve the accuracy of RUL prediction, TSDAE is employed to select the features extracted from the PHM2012 dataset. The raw features and targets under single bearing are input into the TSDAE to obtain the importance ranking of features, the XGBoost and grid search are utilized to determine which sub feature dimensions are optimal. The optimal sub features results under different bearings on PHM2012 dataset are presented in Figure 9, which (a) represents the condition 1 and (b) represents the condition 2.

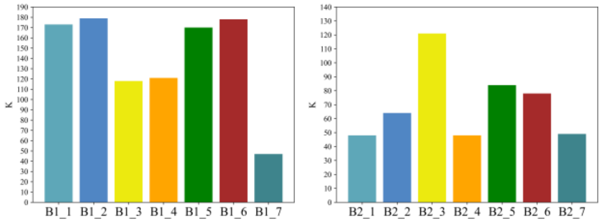
Figure 9 reveals that the feature sets exhibit variability across different bearings, it also denotes that XGBoost and grid search can determine which sub feature dimensions are optimal under different bearings. Subsequently, we calculate the frequency of each raw feature among all the sub optimal features, and each raw feature with high frequency is considered to be the best feature for model training. We integrate the highly frequent features to form the optimal feature set under a single operating condition.

To identify the most influential features, we calculated the proportions of time-domain, frequency-domain, and time-frequency domain features among the top 30 features.

Under condition 1, time-domain features constituted 66.67% of the top features, including crest, absolute mean amplitude, inverse hyperbolic sine, arc tangent standard deviation, inverse hyperbolic sine standard deviation, maximum value, mean, median, minimum value, root mean square, gap factor, standard deviation, and variance. Frequency-domain features accounted for 13.33%, specifically, spectral centroid, coefficient of variation, kurtosis, and root mean square in frequency-domain. Time-frequency domain features represented 20%, comprising: mean, minimum, and standard deviation in time-frequency domain.

Under condition 2, time-domain features constituted 60% of the top features, including coefficient of skewness, pulse factor, inverse hyperbolic sine standard deviation, kurtosis factor, maximum value, mean, median, minimum value, root mean square, gap factor, skewness, and variance. Frequency-domain features accounted for 23.33%, specifically,center of gravity, coefficient of variation, kurtosis, root mean square, and ratio in frequency-domain. Time-frequency domain features represented 16.67%, comprising maximum value, median, minimum value, and variance in time-frequency domain.

In the PHM dataset, the optimal feature set primarily consists of time-domain features, while frequency-domain and time-frequency features constitute a significantly smaller portion.



(a) (b)

Figure 9. The optimal sub features result under different bearings on PHM2012 dataset.

The split of the training set and the testing set on PHM2012 dataset is shown in Table 3.

Table 3. The split of the training set and the testing set on PHM2012 dataset.

Operating conditions	Bearings
1	Training: 1_1, 1_2 Testing: 1_3, 1_4, 1_5, 1_6
2	Training: 2_1, 2_2, 2_3 Testing: 2_4, 2_5, 2_6

In each condition, we select three optimal feature combinations and input them into the model for verification. The feature selection results on PHM2012 are presented in Table 4.

According to the results in Table 4, when the dimension is 91 under condition 1, the model has the lowest RMSE and MAE. Moreover, the performance of dimension 185, which represents the model without TSDAE, has higher RMSE and MAE. When the dimension is 96 under condition 2, the model has the lowest RMSE and MAE. The results of the model without TSDAE are less satisfactory than the proposed model. It also denotes that the TSDAE feature selection proposed in this paper has better feature selection effect. Ultimately, the optimal dimension under condition 1 is 91, and the optimal dimension under condition 2 is 96.

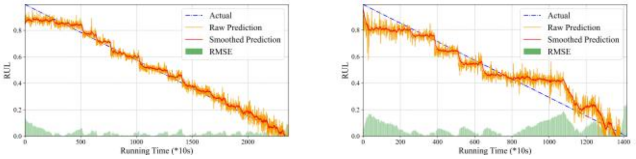
Table 4. The feature selection results on the PHM2012 dataset.

Operating conditions	Dimension	RMSE	MAE
1	142	0.0719	0.05574
	91	**0.0419**	**0.0313**
	29	0.1002	0.0736
	185	0.0884	0.0718
2	142	0.1268	0.1050
	96	**0.0966**	**0.0783**
	64	0.1231	0.1026
	185	0.1113	0.0937

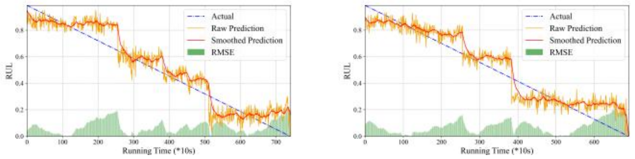
The results of bearing RUL prediction on PHM2012 dataset are presented in Table 5 and depicted in Figure 10. Figure 10 reveals that the RMSE and MAE under each bearing mainly are concentrated around 0.

Table 5. The results of bearing RUL prediction on PHM2012 dataset.

Bearings	RMSE	MAE
Bearing1_3	0.0419	0.0313
Bearing1_4	0.0861	0.0643
Bearing1_5	0.1520	0.1213
Bearing1_6	0.1371	0.1095
Bearing2_4	0.0966	0.0783
Bearing2_5	0.2149	0.1892
Bearing2_6	0.0985	0.0809



(a) (b)



(c) (d)

Figure 10. The partial visualization results of RUL prediction on PHM2012 dataset, (a) bearing 1_3, (b) bearing 1_4, (c) bearing 2_4, (d) bearing 2_6.

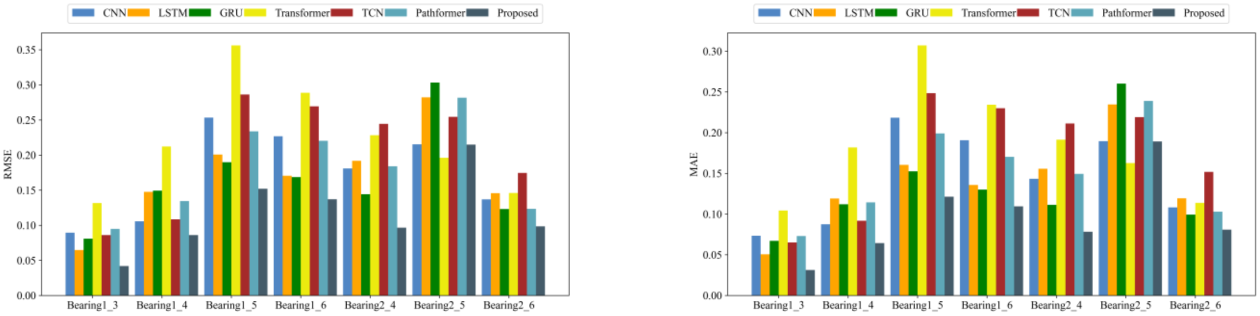
3.1.6. Comparative analysis

To demonstrate the performance of the proposed model in the PHM2012 dataset, it is benchmarked against six other models: CNNs, LSTM, GRU, Transformer, TCN, and Pathformer. The CNNs model incorporates two convolutional layers with three kernels and two max pooling layers with two kernels. The LSTM model includes two LSTM layers with 32 hidden units and 64 hidden units. The GRU model includes two GRU layer with 32 hidden units and 64 hidden units. The Transformer model contains two encoder layers with 128 hidden dimensions and an average pooling layer. The structure of the Pathformer and TCN model is adopted from this article. To guarantee the reliability of validation, each algorithm is executed for five times, with the average value serving as the ultimate predictive result. The comparative results of different algorithms on PHM2012 dataset are presented in Table 6, and the bar chart of comparative results of different models on PHM2012 dataset are illustrated in Figure 11.

Table 6 reveals that the RMSE and MAE of the proposed model reach 0.0419 and 0.0313 under Bearing1_3, outperforming the comparison models. Moreover, the average RMSE and MAE of the proposed model under condition 1 and condition 2 are lower than the comparison model, it denotes that the proposed model has stable performance under various conditions. In comparison with single model (LSTM, TCN, and Pathformer), the proposed model has better RUL prediction results under various conditions, it also denotes that the proposed parallel Pathformer-TCLSTM model can address the limitation of the Pathformer, TCN, and LSTM models. Except for the Bearing1_1, Bearing1_5, Bearing1_6, and Bearing2_5, the RMSE and MAE of the comparison models are all above 0.10, it also denotes that the RUL prediction results of these models should be improved. Compared with the Transformer and Pathformer, the proposed model has better RUL prediction results under various bearings. Additionally, the features extracted from the time-domain, frequency-domain, and time-frequency domain, which are obtained from EMD, can provide effective training set for model training under various conditions. Ultimately, the proposed model can achieve satisfactory RUL prediction results on PHM2012 dataset.

Table 6. The comparative results of different algorithm on PHM2012 dataset.

Test	CNN		LSTM		GRU		Transformer		TCN		Pathformer		Proposed	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
B1_3	0.0895	0.0734	0.0647	0.0506	0.0810	0.0671	0.1317	0.1044	0.0861	0.0651	0.0949	0.0730	**0.0419**	**0.0313**
B1_4	0.1056	0.0875	0.1476	0.1191	0.1494	0.1120	0.2122	0.1818	0.1085	0.0918	0.1345	0.1144	**0.0861**	**0.0643**
B1_5	0.2534	0.2183	0.2008	0.1604	0.1898	0.1525	0.3562	0.3071	0.2863	0.2485	0.2338	0.1990	**0.1520**	**0.1213**
B1_6	0.2268	0.1906	0.1705	0.1357	0.1686	0.1300	0.2887	0.2343	0.2695	0.2300	0.2203	0.1703	**0.1371**	**0.1095**
Average	0.1688	0.14245	0.1459	0.1164	0.1472	0.1154	0.2472	0.2069	0.1876	0.1588	0.1708	0.1391	**0.1042**	**0.0816**
B2_4	0.1809	0.1433	0.1918	0.1556	0.1442	0.1113	0.2283	0.1913	0.2445	0.2113	0.1841	0.1494	**0.0966**	**0.0783**
B2_5	0.2153	0.1895	0.2822	0.2346	0.3033	0.2602	**0.1962**	**0.1626**	0.2544	0.2190	0.2817	0.2389	0.2149	0.1892
B2_6	0.1370	0.1082	0.1457	0.1194	0.1232	0.0993	0.1460	0.1138	0.1746	0.1518	0.1235	0.1030	**0.0985**	**0.0809**
Average	0.1777	0.1470	0.2065	0.1698	0.1902	0.1569	0.1901	0.1559	0.2245	0.1940	0.1964	0.1637	**0.1366**	**0.1161**



(a) (b)

Figure 11. The bar chart of comparative results of different models on PHM2012 dataset, (a) RMSE, (b) MAE.

3.2. Case study 2: XJTU-SY Bearing Dataset

3.2.1. Dataset description

The experimental platform of XJTU-SY (Wang et al., 2018) is depicted in Figure 12. It collects data from 15 bearings across three operational conditions, capturing bearing data from normal operation to failure. Vibration signals are acquired from both horizontal and vertical acceleration sensors. Furthermore, the signal is sampled at a frequency of 25.6 kHz, with a sampling interval of one minute. Additionally, each signal sample comprises 32,768 data points, and the sampling interval is set at 1.28 seconds. The information from the degradation data of XJTU-SY is summarized in Table 7.

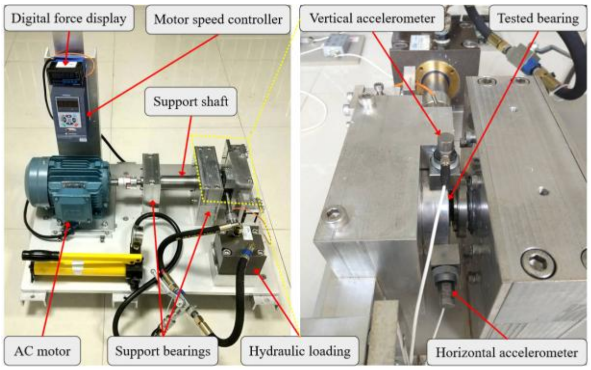


Figure 12. The experimental platform of XJTU-SY.

Table 7. The information of XJTU-SY degradation data.

Operating conditions	Radial force (N)	Speed (rpm)	Bearings
1	12000	2100	Learning: 1_1, 1_2, 1_3, 1_4 Testing:1_5
2	11000	2250	Learning: 2_1, 2_2, 2_3, 2_4 Testing: 2_5
3	10000	2400	Learning: 3_1, 3_2, 3_3, 3_4 Testing: 3_5

3.2.2. Results

The TSDAE approach is utilized to select the features extracted from XJTU-SY bearing degradation, and the optimal sub features results under different bearings on XJTU-SY dataset are illustrated in Figure 13, where (a) represents the condition 1, (b) represents the condition 2, (c) represents the condition 3.

Figure 13 reveals that the feature sets exhibit variability across different bearings, it also denotes that XGBoost and grid search has the ability to determine which sub feature dimensions are optimal under different bearings. Furthermore, the frequency of each raw feature among all the sub optimal features will be calculated. In each condition, we select three optimal feature combinations and input them into the model for verification.

Moreover, we calculated the proportions of time-domain, frequency-domain, and time-frequency domain features among the top 30 features.

Under condition 1, time-domain features constituted 63.33% of the top features, including: kurtosis, absolute mean amplitude, deviation, skewness, energy, inverse hyperbolic sine standard deviation, kurtosis factor, maximum value, median, minimum value, gap factor, skewness, kurtosis factor, and standard deviation.

Frequency-domain features accounted for 23.33%, center of gravity, skurtosis, ratio, root mean square and standard deviation in frequency-domain.

Time-frequency domain features represented 13.33%, comprising: energy, median, minimum value and skewness in time-frequency domain.

Under condition 2, time-domain features constituted 90% of the top features, including: crest, kurtosis, inverse hyperbolic sine, arc tangent standard deviation, coefficient of skewness, energy, pulse factor, inverse hyperbolic sine standard deviation, mean, median, root mean square, skewness, kurtosis factor, standard deviation, and variance.

Frequency-domain features accounted for 6%, specifically: Mean, and standard deviation in frequency-domain.

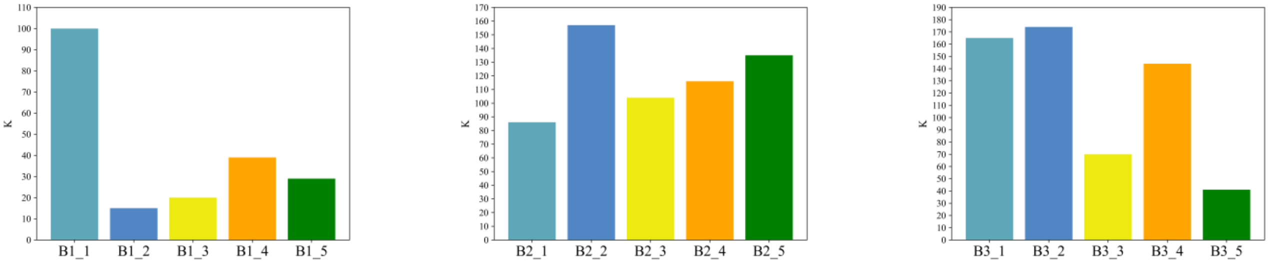
Time-frequency domain features represented 4%, comprising: energy and median in time-frequency domain.

Under condition 3, time-domain features constituted 93.33% of the top features, including: Crest, kurtosis, inverse hyperbolic sine, energy, kurtosis factor, maximum value, mean, minimum value, peak-to-peak, root mean square, skewness, coefficient of skewness, and standard deviation, variance.

Frequency-domain features accounted for 6.67%, specifically: ratio and standard deviation in frequency-domain. Time-frequency domain features represented 4%, comprising: energy and median in time-frequency domain

In the XJTU-SY dataset, time-domain features predominate, with frequency-domain features being secondary.

In each condition, four bearings are designated for the training set, while the remaining one is reserved for the testing set.



(a) (b) (c)

Figure 13. The optimal sub features result under different bearings on XJTU-SY dataset.

The feature selection results on XJTU-SY dataset are presented in Table 8. As seen from Table 8, when the dimension is 137, 135, and 72 under condition 1, condition 2, and condition 3, the model has the lowest RMSE and MAE. Moreover, the performance of dimension 185, which represents the model without TSDAE, has higher RMSE and MAE under each condition. It also denotes that the TSDAE feature selection proposed in this paper has better feature selection effect

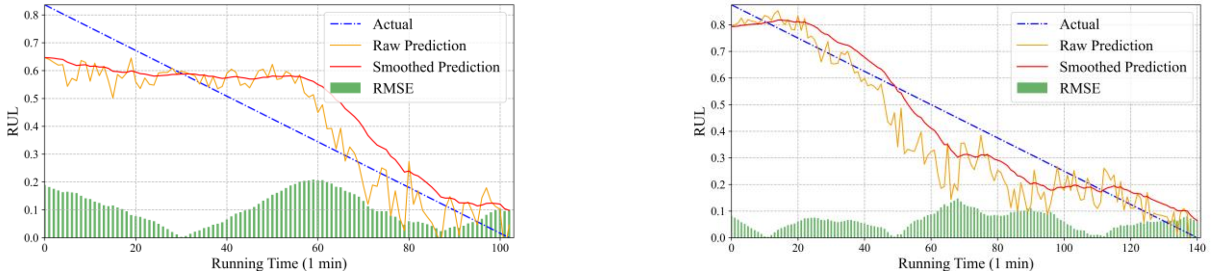
Table 8. The feature selection results of XJTU-SY dataset.

Operating conditions	**Dimension**	**RMSE**	**MAE**
1	**137**	**0.0957**	**0.0763**
	53	0.1280	0.1082
	12	0.1420	0.0900
	185	0.1287	0.1059
2	171	0.1178	0.0960
	135	**0.0947**	**0.0760**
	83	0.1226	0.0953
	185	0.1295	0.1002
3			
	178	0.1697	0.1436
	145	0.1540	0.1257
	72	**0.1060**	**0.0820**
	185	0.2360	0.1980

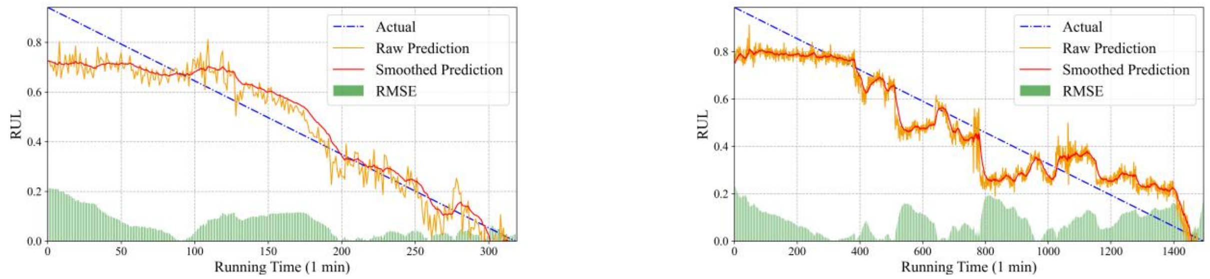
The results of bearing RUL prediction on XJTU-SY dataset are presented in Table 9. Table 9 reveals that the RMSE and MAE under Bearing1_1 are 0.0920, 0.0773. Remarkably, the RSME and MAE under the bearing1_3 and Bearing2_2 are less than 0.1. Moreover, the RMSE and MAE under Bearing1_2, Bearing1_4, Bearing2_3, Bearing2_5, Bearing3_4, Beaing3_5 are less than 0.15. The proposed model in this paper achieves satisfactory RUL prediction results under different conditions on PHM2012 dataset. Furthermore, the partial visualization results of RUL prediction on XJTU-SY dataset are illustrated in Figure 14. In Figure 14, the predicted results are smoothed by the exponential smoothing, and the RMSE of each data point is calculated.

Table 9. The results of bearing RUL prediction on XJTU-SY dataset.

Bearings	**RMSE**	**MAE**
Bearing1_1	0.0920	0.0773
Bearing1_2	0.1063	0.0884
Bearing1_3	0.0957	0.0763
Bearing1_4	0.1200	0.0937
Bearing2_1	0.2276	0.1993
Bearing2_2	0.0929	0.0687
Bearing2_3	0.1068	0.0870
Bearing2_5	0.0947	0.0760
Bearing3_1	0.2217	0.1819
Bearing3_2	0.1532	0.1526
Bearing3_3	0.2021	0.2054
Bearing3_4	0.1060	0.0820
Bearing3_5	0.1335	0.1138



(a) (b)



(c)

(d)

Figure 14. The partial visualization results of RUL prediction on XJTU-SY dataset, (a) bearing 1_1, (b) bearing 1_2, (c) bearing 2_5, (d) bearing 3_4.

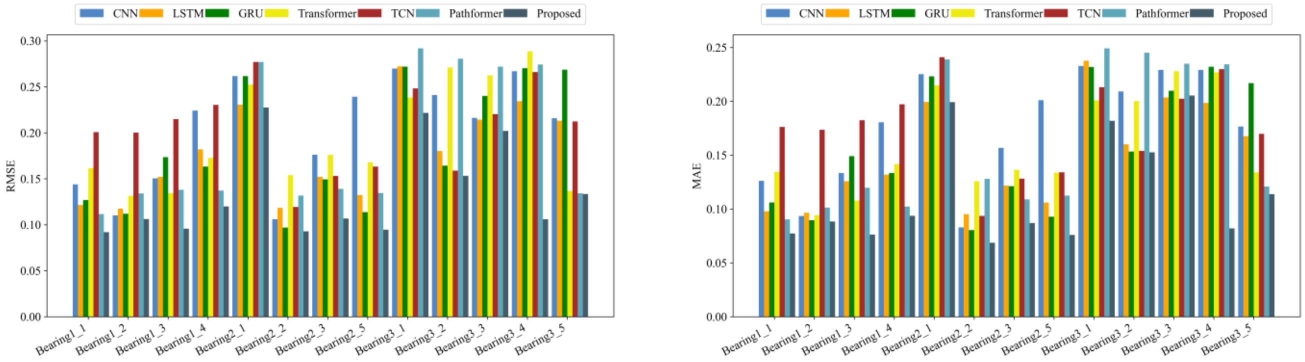
3.2.3. Comparative analysis

To demonstrate the performance of the proposed model in the XJTU-SY dataset, it is benchmarked against six other models: CNNs, LSTM, GRU, Transformer, TCN, and Pathformer. To guarantee the reliability of validation, each algorithm is executed for five times, with the average value serving as the ultimate predictive result. The comparative results of different algorithms on XJTU-SY dataset are presented in Table 10, and the bar chart of comparative results of different models on XJTU-SY dataset are illustrated in Figure 15.

Table 10 shows that for Bearing1_1, the proposed model has RMSE and MAE of 0.0920 and 0.0773, and for Bearing2_5, the RMSE and MAE are 0.0947 and 0.0760, and for Bearing3_4, the RMSE and MAE are 0.1060 and 0.0820, outperforming the comparison models. Additionally, the average RMSE and MAE of the proposed model under condition 1, condition 2, and condition 3 are lower than the comparison models, it denotes that the proposed model has stable performance under various conditions. Compared with the single model (LSTM, TCN, and Pathformer), the proposed parallel Pathformer-TCLSTM model has lower RMSE and MAE under different bearings, it also denotes that the proposed model can address the limitations of the single model (LSTM, TCN, and Pathformer). Compared with the Transformer and Pathformer, the proposed model has better RUL prediction under various bearings. The performance of the proposed model under various bearings indicates that the features extracted from time-domain, frequency-domain, and time-frequency domain, which are obtained from EMD, are effective for the model training. Consequently, the proposed model can achieve satisfactory RUL prediction results on XJTU-SY dataset.

Table 10. The comparative results of different algorithm on XJTU-SY dataset.

Test	**CNN**		**LSTM**		**GRU**		**Transformer**		**TCN**		**Pathformer**		**Proposed**	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
B1_1	0.1440	0.1262	0.1215	0.0979	0.1269	0.1061	0.1616	0.1345	0.2008	0.1762	0.1117	0.0904	**0.0920**	**0.0773**
B1_2	0.1102	0.0935	0.1175	0.0966	0.1120	0.0896	0.1315	0.0943	0.2002	0.1736	0.1341	0.1014	**0.1063**	**0.0884**
B1_3	0.1505	0.1335	0.1521	0.1258	0.1736	0.1491	0.1344	0.1079	0.2149	0.1825	0.1381	0.1198	**0.0957**	**0.0763**
B1_4	0.2242	0.1806	0.1820	0.1320	0.1633	0.1335	0.1728	0.1417	0.2304	0.1973	0.1372	0.1023	**0.1200**	**0.0937**
Average	0.1572	0.1334	0.1432	0.1130	0.1439	0.1195	0.1500	0.1196	0.2115	0.1824	0.1302	0.1034	**0.1035**	**0.0839**
B2_1	0.2617	0.2252	0.2307	0.1995	0.2617	0.2232	0.2523	0.2147	0.2771	0.2410	0.2771	0.2389	**0.2276**	**0.1993**
B2_2	0.1060	0.0830	0.1184	0.0952	0.0970	0.0804	0.1540	0.1258	0.1194	0.0936	0.1318	0.1280	**0.0929**	**0.0687**
B2_3	0.1762	0.1568	0.1521	0.1218	0.1494	0.1212	0.1760	0.1364	0.1531	0.1282	0.1391	0.1090	**0.1068**	**0.0870**
B2_5	0.2393	0.2011	0.1323	0.1060	0.1138	0.0929	0.1679	0.1338	0.1633	0.1341	0.1346	0.1125	**0.0947**	**0.0760**
Average	0.1958	0.1665	0.1583	0.1306	0.1554	0.1294	0.1875	0.1526	0.1782	0.1492	0.1706	0.1471	**0.1305**	**0.1077**
B3_1	0.2699	0.2329	0.2723	0.2377	0.2720	0.2319	0.2384	0.2009	0.2483	0.2131	0.2919	0.2492	**0.2217**	**0.1819**
B3_2	0.2411	0.2092	0.1802	0.1601	0.1643	0.1534	0.2711	0.2002	0.1589	0.1540	0.2806	0.2452	**0.1532**	**0.1526**
B3_3	0.2163	0.2292	0.2144	0.2036	0.2401	0.2098	0.2626	0.2280	0.2204	0.2024	0.2720	0.2349	**0.2021**	**0.2054**
B3_4	0.2670	0.2292	0.2344	0.1985	0.2703	0.2321	0.2886	0.2270	0.2661	0.2299	0.2743	0.2344	**0.1060**	**0.0820**
B3_5	0.2159	0.1766	0.2132	0.1676	0.2687	0.2169	0.1368	0.1340	0.2124	0.1698	0.1342	0.1209	**0.1335**	**0.1138**
Average	0.2420	0.2154	0.2229	0.1935	0.2430	0.2088	0.2395	0.1980	0.2212	0.1938	0.2506	0.2169	**0.1633**	**0.1471**



(a) (b)

Figure 15. The bar chart of comparative results of different models on XJTU-SY dataset, (a) RMSE, (b) MAE.

4. CONCLUSION & FUTURE WORK

This paper introduces an approach for RUL prediction of bearings, combining TSDAE feature selection and Pathformer-TCLSTM. (i) To address the problem of restricted features in RUL prediction, the features of the time-domain, frequency-domain, and time-frequency domain are extracted from EMD; (ii) To deal with the redundancy in extracted features and the limitation of domain knowledge utilized in traditional feature selection, TSDAE feature selection technique is adopted to select the optimal features under various conditions for model training; (iii) To tackle the lack of local information and long-range dependency in Pathformer, the inefficiency of TCN in extracting the global information, the limitation of LSTM in extracting the local information within time series, a parallel Pathformer-TCLSTM prediction model is implemented.

In the proposed RUL prediction model based on TSDAE feature selection and Pathformer-TCLSTM, (i) Features extracted from time-domain, frequency-domain, and time-frequency domain, which are obtained from EMD, can provide more beneficial features for model training and achieve lower RMSE and MAE under various bearings in experiment; (ii) TSADE feature selection adopted in this paper can select the optimal features under various conditions, these optimal features can assist the model to achieve better RUL prediction results under various bearings in experiment. Moreover, TSDAE can learn the relationship between input features and targets via model training, thereby avoiding parameter adjustment and enhancing computational efficiency for RUL prediction; (iii) The proposed parallel Pathformer-TCLSTM model can extracts multi-scale global features while also incorporating local features and long-range dependencies to improve RUL prediction. Compared with the single model in experiment, the proposed model can extract complementary features and achieve better RUL prediction results on PHM2012 dataset and XJTU-SY dataset.

Although the proposed model in this paper can extract the complementary features to improve the accuracy of bearing remaining useful life prediction, the Pathformer adopted in this article is complex. Therefore, future enhancements to the proposed model will focus on adopting a lightweight Pathformer architecture to reduce model complexity and improve efficiency. In addition, the performance of the proposed model under various bearings is fluctuating. Consequently, adopting the transfer learning approach to predict the RUL under various operating conditions of bearings to improve generalization performance will be the main direction in the feature.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 62572293).

REFERENCES

Alfarizi, M. G., Tajiani, B., Vatn, J., & Yin, S. (2022). Optimized random forest model for remaining useful life prediction of experimental bearings. IEEE Transactions on Industrial Informatics, 19(6), 7771-7779. https://doi.org/10.1109/tii.2022.3206339

Aminou, H., Youssoufa, M., Abba, A. A. A., & Gbadoubissa, Z. E. J. (2023). Review of wavelet denoising algorithms. Multimedia Tools and Applications, 82(27), 41539-41569. https://doi.org/10.1007/s11042-023-15127-0

Atashgahi, Z., Sokar, G., van der Lee, T., Mocanu, E., Mocanu, D. C., Veldhuis, R., & Pechenizkiy, M. (2022). Quick and robust feature selection: the strength of energy-efficient sparse training for autoencoders. Machine Learning, 111(1), 377-414. https://doi.org/10.1007/s10994-021-06063-x

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Barraza-Barraza, D., Tercero-Gómez, V. G., Cordero-Franco, A. E., & Beruvides, M. T. (2020). Remaining useful life estimation based on detection of explosive changes: Analysis of bearing vibration. International Journal of Prognostics and Health Management, 11(1). https://doi.org/10.36001/IJPHM.2020.V11I1.2609

Cao, Y., Ding, Y., Jia, M., & Tian, R. (2021). A novel temporal convolutional network with residual self-attention mechanism for remaining useful life prediction of rolling bearings. Reliability Engineering and System Safety, 215, 107813. https://doi.org/10.1016/J.RESS.2021.107813

Carlos, F., & Gil, G. (2022). Remaining Useful Life prediction and challenges: A literature review on the use of Machine Learning Methods. Journal of Manufacturing Systems, 63, 550-562. https://doi.org/10.1016/J.JMSY.2022.05.010

Chen, P., Zhang, Y., Cheng, Y., Shu, Y., Wang, Y., Wen, Q., Yang, B., & Guo, C. (2024). Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. arXiv preprint arXiv:2402.05956.

Ding, Y., & Jia, M. (2022). Convolutional transformer: An enhanced attention mechanism architecture for remaining useful life estimation of bearings. IEEE Transactions on Instrumentation and Measurement, 71, 1-10. https://doi.org/10.1109/tim.2022.3181933

Dong, S., Xiao, J., Hu, X., Fang, N., Liu, L., & Yao, J. (2023). Deep transfer learning based on Bi-LSTM and attention for remaining useful life prediction of rolling bearing. Reliability Engineering & System Safety, 230, 108914. https://doi.org/10.1016/J.RESS.2022.108914

Gao, Z., Jiang, W., Wu, J., & Dai, T. (2025). Multiscale Spatiotemporal Attention Network for Remaining Useful Life Prediction of Mechanical Systems. IEEE sensors journal, 1-1. https://doi.org/10.1109/JSEN.2024.3523176

Guo, J., Wang, J., Wang, Z., Gong, Y., Qi, J., Wang, G., & Tang, C. (2023). A CNN ‐ BiLSTM ‐ Bootstrap integrated method for remaining useful life prediction of rolling bearings. Quality and Reliability Engineering International, 39(5), 1796-1813. https://doi.org/10.1002/qre.3314

Guo, L., Li, N., Jia, F., Lei, Y., & Lin, J. (2017). A recurrent neural network based health indicator for remaining useful life prediction of bearings. Neurocomputing, 240, 98-109. https://doi.org/10.1016/j.neucom.2017.02.045

Guo, R., Wang, Y., Zhang, H., & Zhang, G. (2021). Remaining useful life prediction for rolling bearings using EMD-RISI-LSTM. IEEE Transactions on Instrumentation and Measurement, 70, 1-12. https://doi.org/10.1109/tim.2021.3051717

Ioannides, E., & Harris, T. A. (1985). A new fatigue life model for rolling bearings. 107, 367-377. https://doi.org/10.1115/1.3261081

Li, J. (2017). Feature selection: A data perspective. Comput. Surveys, 50, 6. https://doi.org/10.1145/3136625

Li, X., Zhang, W., & Ding, Q. (2019). Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. Reliability Engineering & System Safety, 182, 208-218. https://doi.org/10.1016/j.ress.2018.11.011

Li, Y., Huang, X., Zhao, C., & Ding, P. (2022). A novel remaining useful life prediction method based on multi-support vector regression fusion and adaptive weight updating. ISA transactions, 131, 444-459. https://doi.org/10.1016/j.isatra.2022.04.042

Ma, M., & Mao, Z. (2020). Deep-convolution-based LSTM network for remaining useful life prediction. IEEE Transactions on Industrial Informatics, 17(3), 1658-1667. https://doi.org/10.1109/tii.2020.2991796

Medjaher, K., Tobon-Mejia, D. A., & Zerhouni, N. (2012). Remaining useful life estimation of critical components with application to bearings. IEEE Transactions on reliability, 61(2), 292-302. https://doi.org/10.1109/tr.2012.2194175

Motahari-Nezhad, M., & Jafari, S. M. (2021). Bearing remaining useful life prediction under starved lubricating condition using time domain acoustic emission signal processing. Expert Systems With Applications, 168, 114391. https://doi.org/10.1016/j.eswa.2020.114391

Najdi, B., Benbrahim, M., & Kabbaj, M. N. (2025). Adaptive Res-LSTM Attention-based Remaining Useful Lifetime Prognosis of Rolling Bearings. International Journal of Prognostics and Health Management, 16(1). https://doi.org/10.36001/ijphm.2025.v16i1.4171

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). PRONOSTIA: An experimental platform for bearings accelerated degradation tests. IEEE International Conference on Prognostics and Health Management, PHM'12.,

Paris, P., & Erdogan, F. (1963). A critical analysis of crack propagation laws. 85(4), 528-533. https://doi.org/10.1115/1.3656900

Peng, H., Jiang, B., Mao, Z., & Liu, S. (2023). Local enhancing transformer with temporal convolutional attention mechanism for bearings remaining useful life prediction. IEEE Transactions on Instrumentation and Measurement, 72, 1-12. https://doi.org/10.1109/tim.2023.3291787

Qiu, H., Niu, Y., Shang, J., Gao, L., & Xu, D. (2023). A piecewise method for bearing remaining useful life estimation using temporal convolutional networks. Journal of Manufacturing Systems, 68, 227-241. https://doi.org/10.1016/j.jmsy.2023.04.002

Ren, L., Sun, Y., Cui, J., & Zhang, L. (2018). Bearing remaining useful life prediction based on deep autoencoder and deep neural networks. Journal of Manufacturing Systems, 48, 71-77. https://doi.org/10.1016/j.jmsy.2018.04.008

Saufi, M. S. R. M., & Hassan, K. A. (2021). Remaining useful life prediction using an integrated Laplacian-LSTM network on machinery components. Applied Soft Computing, 112, 107817. https://doi.org/10.1016/J.ASOC.2021.107817

Sun, C., Ma, M., Zhao, Z., Tian, S., Yan, R., & Chen, X. (2019). Deep Transfer Learning Based on Sparse Autoencoder for Remaining Useful Life Prediction of Tool in Manufacturing. IEEE Transactions on Industrial Informatics, 15(4), 2416-2425. https://doi.org/10.1109/TII.2018.2881543

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://doi.org/10.48550/arXiv.1706.03762

wahhab Lourari, A., Benkedjouh, T., El Yousfi, B., & Soualhi, A. (2024). An anfis-based framework for the prediction of bearing's remaining useful life. International Journal of Prognostics and Health Management, 15(1). https://doi.org/10.36001/ijphm.2024.v15i1.3791

Wang, B., Lei, Y., Li, N., & Li, N. (2018). A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. IEEE Transactions on reliability, 69(1), 401-412. https://doi.org/10.1109/tr.2018.2882682

Wang, H., Wang, D., Liu, H., & Tang, G. (2022). A predictive sliding local outlier correction method with adaptive state change rate determining for bearing remaining useful life estimation. Reliability Engineering and System Safety, 225, 108601. https://doi.org/10.1016/J.RESS.2022.108601

Wang, H., Yang, J., Wang, R., & Shi, L. (2023). Remaining useful life prediction of bearings based on convolution attention mechanism and temporal convolution network. IEEE Access, 11, 24407-24419. https://doi.org/10.1109/access.2023.3255891

Wang, W., & Lu, Y. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. Iop Conference, 324, 012049. https://doi.org/10.1088/1757-899X/324/1/012049

Wang, Y., Zhao, J., Yang, C., Xu, D., & Ge, J. (2022). Remaining useful life prediction of rolling bearings based on Pearson correlation-KPCA multi-feature fusion. Measurement, 201, 111572. https://doi.org/10.1016/J.MEASUREMENT.2022.111572

Wang, Y., Zhao, Y., & Addepalli, S. (2020). Remaining useful life prediction using deep learning approaches: A review. Procedia manufacturing, 49, 81-88. https://doi.org/10.1016/j.promfg.2020.06.015

Wen, L., Su, S., Li, X., Ding, W., & Feng, K. (2024). GRU-AE-wiener: A generative adversarial network assisted hybrid gated recurrent unit with Wiener model for bearing remaining useful life estimation. Mechanical systems and signal processing, 220, 111663. https://doi.org/10.1016/j.ymssp.2024.111663

Yang, W., Yao, Q., Ye, K., & Xu, C.-Z. (2020). Empirical mode decomposition and temporal convolutional networks for remaining useful life estimation. International journal of parallel programming, 48(1), 61-79. https://doi.org/10.1007/s10766-019-00650-1

Yang, X., Chen, D., Huang, J., Wu, X., Chen, Z., & Li, Q. (2025). Remaining Useful Life Prediction Under Multiple Operating Conditions Based on a Novel Dual-Layer Temporal Convolutional Network. IEEE sensors journal, 25(1), 1900-1911. https://doi.org/10.1109/JSEN.2024.3494020

Yang, Y., Wu, Q. M. J., & Wang, Y. (2016). Autoencoder With Invertible Functions for Dimension Reduction and Image Reconstruction. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48(7), 1065-1079. https://doi.org/10.1109/TSMC.2016.2637279

Yao, D., Li, B., Liu, H., Yang, J., & Jia, L. (2021). Remaining useful life prediction of roller bearings based on improved 1D-CNN and simple recurrent unit. Measurement, 175, 109166. https://doi.org/10.1016/j.measurement.2021.109166

Ye, Y., Wang, J., Yang, J., Yao, D., & Zhou, T. (2024). Adaptive MAGNN-TCN: An Innovative Approach for Bearings Remaining Useful Life Prediction. IEEE sensors journal, 1-1. https://doi.org/10.1109/JSEN.2024.3506154

Yu, W., Pi, D., Xie, L., & Luo, Y. (2021). Multiscale attentional residual neural network framework for remaining useful life prediction of bearings. Measurement, 177, 109310. https://doi.org/10.1016/J.MEASUREMENT.2021.109310

Zhang, H., Zhang, Q., Shao, S., Niu, T., & Yang, X. (2020). Attention-Based LSTM Network for Rotatory Machine Remaining Useful Life Prediction. IEEE Access, 8, 132188-132199. https://doi.org/10.1109/access.2020.3010066

Zhao, H., Liu, H., Jin, Y., Dang, X., & Deng, W. (2021). Feature extraction for data-driven remaining useful life prediction of rolling bearings. IEEE Transactions on Instrumentation and Measurement, 70, 1-10. https://doi.org/10.1109/tim.2021.3059500

Zhu, G., Zhu, Z., Xiang, L., Hu, A., & Xu, Y. (2023). Prediction of bearing remaining useful life based on DACN-ConvLSTM model. Measurement, 211, 112600. https://doi.org/10.1016/J.MEASUREMENT.2023.112600

Zhu, J., Chen, N., & Shen, C. (2020). A new data-driven transferable remaining useful life prediction approach for bearing under different working conditions. Mechanical systems and signal processing, 139, 106602. https://doi.org/10.1016/j.ymssp.2019.106602

BIOGRAPHIES

Guanghua Fu received the Ph.D. degree in Management Science and Engineering from Jilin University, Changchun, China, in 2019.

He is currently a lecturer in Shanghai Maritime University, Shanghai, China. His current research interests include intelligent information acquisition and control, mechanical reliability analysis, remaining useful life prediction, and prognostics and health management.

Yujie Yang received the B.S. degree in Automation from Anhui University of Science and Technology, Huainan, China, in 2021. She is currently pursuing the M.S. degree in Control Science and Engineering from Shanghai Maritime University, Shanghai, China. Her current research interests include trajectory tracking control for autonomous vehicles.

Yonghui Liu received the B.S. degree in computer science and technology from Jinling Institute of Technology, Nanjing, China, in 2023. He is currently pursuing the M.S degree in control science and engineering from Shanghai Maritime University, Shanghai, China.

His current research interests include prediction and control of mechanical behaviors, prognostic health management of mechanical systems.

Xuegen Wang received the B.S. degree in automation from Huangshan University, Huangshan, China, in 2022. He is currently pursuing the M.S. degree in control science and engineering from Shanghai Maritime University, Shanghai, China.

His current research interests include prognostic and health management of mechanical systems.