Spatial-Attention-based Convolutional Transformer for Bearing Remaining Useful Life Prediction

Chong Chen[a], Tao Wang[a, b*], Ying Liu[c], Lianglun Cheng[a] and Jian Qin[d]

[a] *Guangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology, Guangzhou 510006, China*

[b] *School of Automation, Guangdong University of Technology, Guangzhou 510006, China.*

[c] *Department of Mechanical Engineering, School of Engineering, Cardiff University, Cardiff CF24 3AA, UK*

[d] *Welding Engineering and Laser Processing Centre, School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, UK*

[*] *Corresponding author. E-mail: wangtao_cps@gdut.edu.cn*

Abstract

The remaining useful life (RUL) prediction is of significance to the health management of bearing. Recently, deep learning has been widely investigated in bearing RUL prediction due to its great success in sequence learning. However, the improvement of the prediction accuracy of the existing deep learning algorithms heavily relies on feature engineering such as handcrafted features generation and time-frequency transformation, which increase the complexity and difficulty of the actual deployment. In this paper, a novel spatial attention-based convolutional Transformer (SAConvFormer) is proposed to establish an accurate bearing RUL prediction model based on the raw vibration data without priori knowledge or feature engineering. In this algorithm, firstly, a convolutional neural network (CNN) enhanced by spatial attention mechanism is proposed to squeeze the feature maps and extract the local and global features from raw bearing vibration data effectively. Then, the extracted senior features are fed into a Transformer network to further explore the sequential patterns relevant to the bearing RUL. An experimental study using the XJTU-SY rolling bearings dataset revealed the merits of the proposed deep learning algorithm in terms of RMSE and MAE in comparison with other state-of-the-art algorithms.

Keywords: Remaining useful life; Prognostic and health management; Deep learning; Transformer network; CNN.

1. Introduction

With the rapid development of industrial Internet-of-things (IIoT) and artificial intelligence (AI), asset prognostic and health management (PHM) has developed rapidly in recent years [1, 2]. The advance of IIoT allows a large number of condition monitoring data to be collected, which can be analysed via the emerging AI techniques. With the advance of the above techniques, condition-based maintenance has been developed rapidly, which has been widely studied in the modern industrial system in recent years [3]. Bearing is an important component in a large number of rotating machinery in the industry. The failure of the bearing may result in the machinery shutdown, breakdown of the production line and casualty [4]. The RUL of bearing can indicate the health status of bearing. With an accurate prediction of bearing RUL, appropriate maintenance can be scheduled before catastrophic failure occurs, and therefore the maintenance cost can be lower and the productivity of the machine can be improved [3].

Data-driven approaches are effective to model the RUL from measured data using machine learning techniques [5]. With the development of deep learning, various algorithms such as CNN [6], long-short term memory networks (LSTM) network [7], and autoencoder [8] have shown merits in automatic feature extraction. Recently, the studies of combining feature engineering or statistical approach with deep learning algorithms have gained increasing attention in bearing RUL prediction. For example, extracting time or frequency domain features using wavelet transform or fast Fourier transform can boost the RUL prediction accuracy using deep learning, while the priori knowledge of wavelet transforms or fast Fourier transform is required [9]. It increases the complexity of the modelling process and raises the bar for actual deployment. Hence, an end-to-end modelling algorithm is necessary to be investigated to get accurate RUL prediction without priori knowledge.

It is well known that attention plays an important role in human perception [10]. Deep learning is a type of algorithm inspired by human perception. However, most deep learning algorithms

do not have an effective mechanism to locate the key position of the input data [11]. Attention mechanism was proposed to identify the importance of each feature in the input data [12]. The variants of the attention mechanism such as self-attention [13], spatial attention [14], and channel attention [15] have been widely investigated in deep learning algorithms to promote algorithm performance. Spatial attention [14] can be introduced into CNN to enhance the global patterns learning ability, which is essential in the bearing RUL modelling. Multi-head self-attention mechanism is the core part of the Transformer network, which is effective in learning the feature from different aspects [13]. CNN can aggregate the local information at lower layers, while it is short of global patterns learning. In contrast, the Transformer network is able to access any part of the historical data regardless of distance, which enables the network to capture the long-term dependency. However, the Transformer network is not sensitive to the local contexts [16, 17], which contains important patterns highly relevant to the asset's health status. An accurate RUL prediction needs the contribution of both global and local patterns mining from the vibration data [9]. Meanwhile, the high sampling frequency of vibration data poses a challenge in computational cost to the Transformer network. Feeding a long sequence into a Transformer network can lead to a high computational cost and low modelling performance. Hence, it is worthwhile to explore a computationally efficient and end-to-end Transformer network which is able to extract both global and local patterns from the vibration data so to achieve accurate RUL prediction.

In this study, an end-to-end RUL modelling algorithm called SAConvFormer, which can achieve accurate RUL prediction based on the raw vibration data without extra feature engineering was proposed. In the lower level of SAConvFormer, spatial attention is adopted to enhance the global patterns extraction of CNN. With the global and local features captured by spatial attention enhanced CNN, the subsequent Transformer block is able to model RUL accurately. The main contribution of this study is three-fold: (1) Different from the existing studies that heavily rely on extra feature engineering, this study proposed an end-to-end algorithm to directly model the RUL based on the raw vibration data; (2) Since directly feeding

the long sequence of raw vibration data into a Transformer network leads to a huge computational cost and the challenge of modelling training, spatial attention mechanism was introduced to CNN so to capture both local and global patterns and reduce the feature size; (3) A Transformer block was adopted to further processed the extracted features from the spatial attention based CNN (SA-CNN) so to obtain an accurate prediction of bearing RUL. The overall structure of the paper is organised as follows: Section 2 reviews the related works on RUL modelling and the applications of attention mechanism in PHM. Section 3 details the methodology of this paper. Section 4 introduces the experimental setup and the experimental results are demonstrated in Section 5. Finally, Section 6 discusses, and Section 7 concludes.

2. Literature Review

2.1. The studies of RUL prediction

Recently, deep learning has been widely investigated in PHM. The studies of deep learning in RUL modelling mainly consisted of two types which are direct RUL modelling and health-index (HI) based RUL modelling. Direct RUL modelling is that mapping the relationship between input data and RUL using deep learning directly. Wang et al. [18] proposed a deep separable convolutional network for machinery residual connection RUL prediction. A separable convolutional layer joint with residual connection was designed as a process block. By stacking multiple process blocks, a deep neural network was obtained to learn the hidden representations in the raw sensor data. CNN as a prevailing deep learning algorithm has been widely studied in industrial applications [19-22]. Li et al. [21] proposed a WaveletKernelNet architecture, where a continuous wavelet convolutional layer was designed as the first hidden layer in CNN. In the continuous wavelet convolutional layer, the convolutional operation and continuous wavelet transform were combined to extract the useful features in the sensor signal. Another study that focuses on time-frequency features is that Li et al. proposed [22] a multi-scales CNN for the RUL modelling. In this study, the raw signal data was first transformed into

image-style input, which was then processed using multiscale CNN.

HI-based RUL modelling aims to obtain the health degradation curve of an asset in the first place and then estimate the RUL according to the health status. The HI-based RUL modelling approach has the restriction that the relation between the multi-sensor data and the asset degradation needs to follow a specific distribution such as linear or exponential [5]. In order to overcome this limitation, Wang et al. [23] proposed an indirect gradient descent algorithm to train a deep neural network and a long-short term memory network to fuse multi-sensor signals to obtain the asset HI. With the pre-determined failure threshold and the estimated HI, the RUL can be predicted. Zhao et al. [24] proposed a normalised CNN algorithm to identify the degradation point in the bearing life cycle. The proposed CNN model is able to extract the salient features from the raw signal data of bearing. Liu et al. [25] combined the advantage of the LSTM network with statistical process analysis to predict the degradation of bearings. First, the time-domain features were extracted and segmented using statistical process analysis. The root-means-square value of vibration was adopted as the degradation degree. Subsequently, the data was fed into the LSTM network to get the degradation prediction. She et al. [26] reported a HI construction method based on a sparse auto-encoder with regularization model for rolling bearings. The proposed model extracts the original features for the construction of the HI. With minimal quantization error, features with the highest trendability are selected for constructing an HI by sorting them based on their trendability.

Besides modelling RUL using pure deep learning algorithms, combining deep learning with other statistical approaches or machine learning approaches has been considered by different researchers. In order to take advantage of both the data-driven approach and the statistical approach, Wang et al. [27] proposed a hybrid approach for bearing RUL prediction modelling. In this approach, relevance vector machine regressions with different kernel parameters were firstly adopted to represent the degradation of bearing vibration data. Then an exponential

model coupled with the Fr´echet distance was utilised to predict the bearing RUL. Li et al. [6] proposed an end-to-end RUL prediction algorithm that combined LSTM and encoding-decoding framework for the data processing model. Then the processed data is subsequently fed into a temporal convolutional network based on the CNN. Tang et al. [28] proposed a new RVM model, called the weight-tracking relevance vector machine. To prevent overfitting, an adaptive sequential optimal feature selection method is proposed within the proposed model.

Furthermore, an adaptive modification method is proposed to improve the RUL prediction accuracy. Huang et al. [9] proposed a deep convolutional neural network-bootstrap integrated method for bearing RUL prediction. In this approach, the time-frequency features were obtained via continuous wavelet transform. A multi-modal network was used to process the time-frequency features and handcrafted features. Then a bootstrap scheme was adopted to get the RUL prediction with prediction interval. Gao et al. [29] proposed an enhanced LSTM network and combined it with ensemble empirical mode decomposition (EEMD) energy moment entropy. In this approach, the energy moment entropy of the intrinsic mode functions were extracted as the input of the enhanced LSTM network. Zou et al. [30] proposed a multi-domain adversarial network-based approach to reduce the discrepancy between different working conditions in RUL prediction transfer learning. The data samples are first selected via particle swarm optimisation algorithm, before the Hilbert envelope spectrum and degradation signal spectrum were extracted. The extracted features were sent into the proposed network to reduce the discrepancy in feature distribution between target and source domains. Besides the literature above, researchers also have proposed various types of deep learning algorithms in component or machinery RUL prediction, such as Bi-LSTM network [31], dilated-CNN [32], deep attention residual neural network [33] and deep adversarial neural networks [34].

2.2. The studies of attention mechanism in PHM

Recently, the attention mechanism has been widely investigated in deep learning since it was

proposed in 2014 [12]. RNN has shown advantages in processing sequential data. The main focus of the attention mechanism is the combination with RNN, which has shown merits in PHM. Liu et al. [3] proposed a new feature-based attention mechanism that can adaptively weigh the features of the input data. With the deployment of the proposed attention mechanism, the Bi-directional gated recurrent unit (GRU) network can achieve better performance in turbofan engine RUL modelling. Duan et al. [35] designed a Bi-GRU autoencoder with an attention mechanism and skipped connection to locate the important features and reduce the decoding burden. In this study, the Bi-GRU autoencoder was designed to learn the sequential features within the multi-sensor data in an unsupervised manner. Chen et al. [36]embedded the attention mechanism into the LSTM network, which is helpful to learn the importance of features and time steps. Besides the studies of applied attention mechanism in RNN, CNN is another structure that attention mechanism can work with. Wang et al. [37] proposed a temporal CNN with soft thresholding and attention mechanism for machinery RUL prediction. In the algorithm, CNN with dilation operation and causal padding were adopted to capture the sequential features. Meanwhile, soft thresholding was proposed to modify the activation function so as to obtain more useful features. Huang et al. [38] proposed a frequency Hoyer attention-based CNN. In this network, adaptive weighting of the feature map is calculated using the frequency Hoyer attention, which can extract the hidden representation of vibration data from different perspectives.

With the advance of the multi-head self-attention mechanism, a Transformer neural network was proposed in 2017 [13]. The structure of the Transformer gets rid of RNN and CNN and extracts sequential patterns based on the attention mechanism. This is an effective way to realise parallel computing and long-term sequential feature extraction. In PHM, Liu et al. [39] proposed a multi-modal Transformer neural network for tool wear estimation. In this approach, three statistical features of tool-wear signals, which are maximum, mean, and variation, were first extracted. Then the three types of features were then fed into three sub-networks individually. The subnetwork was a modification of a standard Transformer block, which uses

the LSTM layer to replace the fully connected layer. The output of the sub-networks was then concatenated and sent into a fully connected layer for tool wear estimation. Mo et al. [40] modified the standard Transformer for modelling the RUL of a turbofan engine. By introducing a gated convolution unit, the proposed algorithm is able to capture the local context of the data. In the proposed architecture, a gated convolutional unit layer was deployed as the first hidden layer, followed by a linear layer and position encoding operation. Then the data was fed into a Transformer block to further extract the senior features. Finally, the output of Transformer block was sent into a linear layer to yield the predicted RUL. Chen et al. [41] proposed a Transformer-based framework for vibration signal classification. In this framework, the time-frequency spectrum features were extracted from raw signal data using discrete Fourier transform and short-term Fourier transform. Then the extracted features were further processed via Mel filter bank and discrete cosine transform. Finally, the processed features were fed into a modified Transformer network which a Bi-LSTM layer was introduced in the hidden layers.

It can be seen from the state-of-the-art that most studies of deep learning in PHM strongly rely on feature engineering or statistical process to improve the RUL prediction accuracy, which imposes a challenge in deploying these approaches in the real world. Another issue is that the research on Transformer in PHM is limited, and most of them require extra feature engineering. Different from existing studies, this paper proposed a new deep learning algorithm called SAConvFormer, which takes advantage of both SA-CNN and Transformer encoder. The proposed algorithm can be deployed to build an end-to-end bearing RUL prediction model without priori knowledge.

3. Methodology

The general flowchart of the experiment is illustrated in Figure 1. Firstly, the data is collected from the experiment platform. Secondly, the first prediction time (FPT) of bearing is

determined. Thirdly, the data is pre-processed via sliding window processing and training/ testing data split. Then the training data is fed into the proposed SAConvFormer algorithm to train a bearing RUL prediction model. Finally, the testing set is sent into the trained model to obtain the RUL prediction. The details of each stage are explained in the following sub-sections.

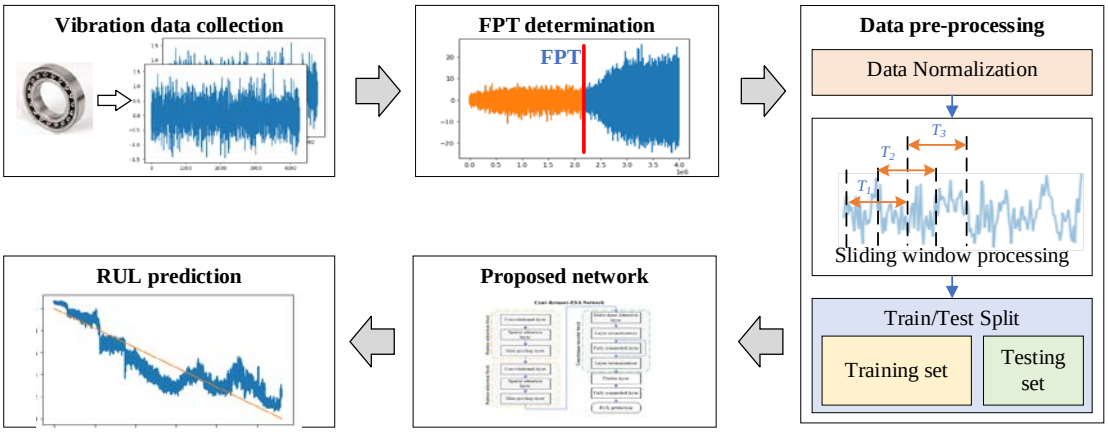


Figure 1. The flow chart of the methodology

3.1. FPT determination

The bearing degradation can be roughly classified as healthy stage and unhealthy (degradation) stage [1]. The boundary between the healthy stage and the unhealthy stage is FPT. The health of bearing before FPT decreases slowly, which is challenging to capture its degradation pattern via the analysis of vibration signal. In contrast, the bearing degrades rapidly after FPT, where the change of vibration signal is significant. Hence, the vibration data before FPT is truncated since its degradation pattern is feeble, and FPT is deemed as the degradation start point in the bearing life cycle. With an accurate determination of FPT, the modelling performance of bearing RUL can be leveraged.

Among different approaches to determining the FPT of bearing, the kurtosis-based fault detection method is one of the most effective and convenient approaches that has been widely adopted [5]. The kurtosis-based approach is able to capture the feeble degradation patterns in the early stage. Moreover, it is also advantageous in the robustness of the random noise. The

detail of this approach is shown as follows:

(1) Assuming the early stage of bearing running period is in the healthy stage. Calculating the mean value μ and standard deviation σ of the kurtosis of the vibration data in this period.

(2) Computing the threshold of the anomaly detection point. When the measured kurtosis is out of the interval $[\mu - 3\sigma, \mu + 3\sigma]$, it is considered an anomaly point.

(3) Defining the tolerance of anomaly points number. Taking three as an instance. It means that if there are three contiguous measured signals out of the pre-set interval, it is considered as the degradation start point, which is the FPT.

3.2. Data pre-processing

After the FPT is determined, the data in the healthy stage is truncated. Subsequently, the data need to be pre-processed. Firstly, the lifetime and RUL of the bearings vary from each other, while the degradation data of each bearing is recorded from 100% healthy stage to complete failure. Directly allocating the RUL as the data label cannot represent a bearing's actual healthy stage. Hence, the RUL of each bearing is scaled to the range from 0 to 1, which is more representative of the actual degradation. Furthermore, the stoppage of each bearing can be different, which is determined according to the maximum vibration amplitude of each bearing. Therefore, the vibration data in each bearing need to be normalised to the range from 0 to 1as well.

Since the vibration data is collected continuously at a high sampling rate, treating each observation point as a data instance is not reasonable because the sequential dependency between regional observation points is ignored. One of the widely used approaches to get the data instance is sliding window processing. Given a window length and strides, the time window slides from the start point of the signal to the end of the signal, and each sliding generates a data instance. By adopting the sliding window strategy, each data instance contains multiple data points, which contains more information. The parameters of length and strides

need to be determined in the actual case [42]. After the sliding window processing, the data is then split into a training set and a testing set for modelling training and testing. The details of the data spilt are detailed in Section 4.2.

3.3. Proposed network

In this sub-section, the details of the proposed SAConvFormer algorithm are elaborated. The proposed algorithm consists of a feature extraction block for sequential features extraction and a Transformer block for sequential patterns mining. The overall architecture of SAConvFormer is shown in Figure 2.

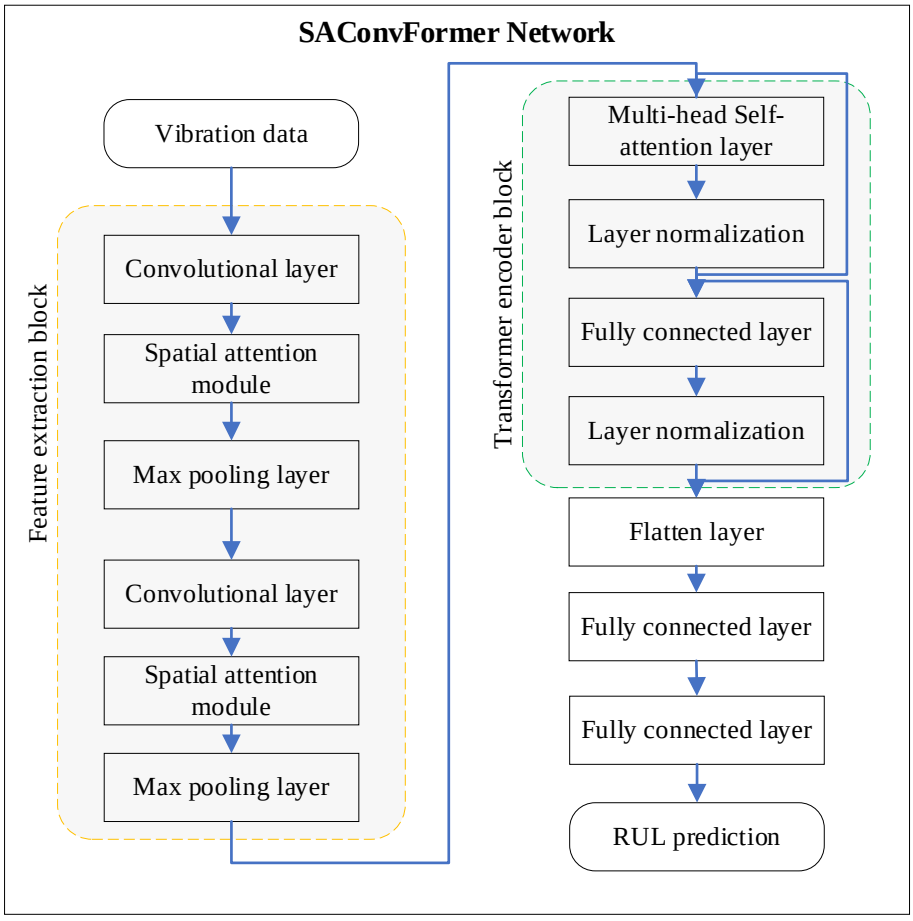


Figure 2. The architecture of the proposed algorithm

3.3.1. *Feature extraction block*

The Feature extraction block consists of convolutional layers, spatial attention modules and max-pooling layers. Besides reducing the input dimension and extracting the important features, the first convolutional layer in the feature extraction block is used to fuse the input of multiple sensor data. With the application of the spatial attention module, the spatial patterns between different convolutional features can be located, which is beneficial to the feature extraction. Furthermore, max-pooling layers are used to reduce the dimension of extracted features.

Convolutional layer: The multi-vibration signals are prepared in a two-dimensional format. Specifically, the horizontal vibration and vertical vibration signals of a bearing are stacked as the input data of the convolutional layer, which is able to extract the features relevant to the bearing RUL. A convolutional layer consists of multiple learnable kernels to generate new feature maps. The convolutional operation can be expressed as:

$$x_j^l = \sigma(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l) \quad (1)$$

where * represents convolutional operation, and x_i is ith input feature map. k is one of the convolutional kernels, which can be deemed as a filter. b is an additive bias. M_j is a feature map of the convolutional layer. l is the index of the convolutional layers. $\sigma()$ is the activation function.

Pooling layer: The pooling layer is normally located after the convolutional layer, which is adopted to subsample the feature map and locate the most significant features. There are various types of pooling strategies, such as maximum polling and average pooling. The max-pooling operation takes the maximum value in the selected region, while average-pooling takes the average value instead. With the adoption of the pooling layer, the computational load can be lowered without sacrificing the important information in the feature maps.

Spatial attention module: Spatial attention is an effective and efficient attention module that

can be helpful in the feature extraction of CNN [14]. The spatial attention module aims to locate the position of the informative part of features via mining the inter-spatial relationship between features, which enhances the global feature extraction of CNN. The flow chart of getting spatial attention is shown in Figure 3. To obtain spatial attention, average-pooling and max-pooling along the channel axis are adopted to get a new feature descriptor by concatenating the output of both pooling operations. Then the new feature descriptor is fed into a convolutional layer to yield a spatial feature map, which the important features are assigned with higher weights and the unimportant features are assigned with lower weights. Specifically, the channel information from the last convolutional layer is aggregated by a feature map generated by two pooling operations and a convolutional operation. The spatial attention feature map can be computed as:

$$M_s(F) = \sigma(f^{n*n}(AvgPool(F); MaxPool(F))) \quad (2)$$

where σ denotes the sigmoid activation function and $f()$ is the convolution operation with the kernel size is $n * n$.

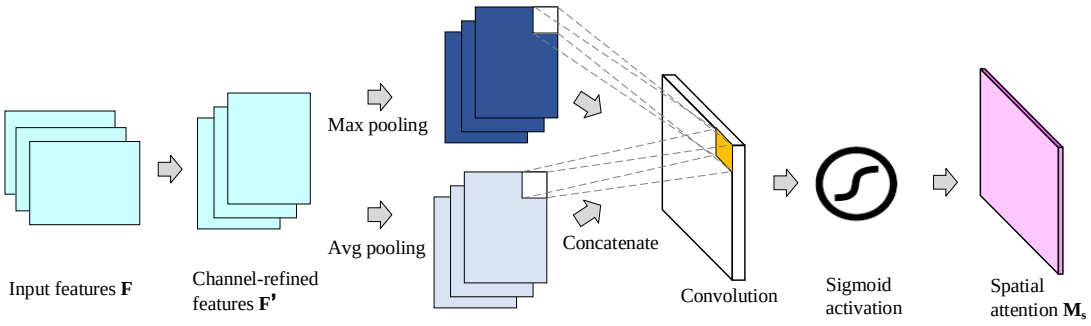


Figure 3. The flow chart of getting spatial attention

3.3.2. *Transformer network*

With the advance of the multi-head self-attention mechanism, the Transformer network is able to capture the long- and short-term dependency within data. A standard Transformer network consists of a position encoding layer, a multi-head self-attention layer, two layer-normalisation layers, and a fully connected layer. The skip connection is adopted in the Transformer network

to overcome the gradient vanish problem [43]. The ability of the Transformer network to learn the sequential pattern within data is deficient since the self-attention mechanism is not sensitive to the positional information within the data. In order to address this issue, the position encoding layer is utilised to provide the sequential information so that the sequential pattern within the data can be captured. Because of that, the structure is advantageous in parallel computing and long-term sequential pattern mining [13]. In this study, the input of the Transformer network is the output of the feature extraction block, in which the sequential pattern is captured via spatial attention and convolution layer. Hence, positional encoding is not a necessary component in this study. The number of training parameters in Transformer network is relevant to its input, which is the output of the SA-CNN. If the input of the Transformer network is too large, it may cause a huge computation cost and difficulty for the Transformer network to learn the sequential patterns. Meanwhile, if the input of the Transformer network is too small, it may cause information loss and result in unsatisfactory performance. Hence, the structure of the SA-CNN needs to be well designed in order to send the appropriate size of input data to the Transformer network.

The multi-head self-attention layer is the key component in the Transformer network. The extracted features from the feature extraction block are firstly sent to the multi-head self-attention layer, which adopts multiple self-attention modules to learn the important features from different perspectives. In order to get the attention output, the relationship between a query and a set of key-value pairs to output needs to be determined. The attention output is a weighted sum of the values, which can indicate the location of the important features. The weights are obtained via the computation of a compatibility function of the query with the corresponding key. The standard self-attention layer obtains the attention score via computing the query vector q, key vector k and value vector v. When a new input i is fed into a self-attention attention layer, the attention score is calculated as:

$$Score = softmax(q_i * k_i) * v_i \quad (3)$$

Unlike the standard self-attention layer, multi-head self-attention deploys three matrices Q, K and V to replace the vectors q, k and v. The rich and complex information within the matrix is beneficial to comprehensively determine the feature importance from different aspects. Given an input data $X = [x_1, x_2, \dots, x_n]$. A linear transformation is first deployed to the input data to yield the matrices Q, K and V, which can be expressed as:

$$Q = XW^q \tag{4}$$

$$K = XW^k \tag{5}$$

$$V = XW^v \tag{6}$$

, where W^q, W^k and W^v are trainable projection matrices.

Then the derived matrices Q, K and V are used as the input of scaled dot-product attention. The attention score of a head is then computed as:

$$Head\_Score_1 = \frac{softmax(Q \times K^T \times V)}{\sqrt{d}} \tag{7}$$

, where d is a scalable factor.

In order to obtain the attention score from different perspectives, multiple heads are used to generate different attention scores. Then the attention scores obtained from different heads are concatenated, which can be formulated as:

$$MultiHead(Q, K, V) = Concat(Head_{Score_1}, Head_{Score_2}, \dots, Head_{Score_m})W^o \tag{8}$$

, where m is the number of heads, and W^o is a trainable weighted matrix.

The output of the multi-head self-attention is then concatenated with its own input and then processed by a layer normalisation layer to avoid overfitting [44], which can be expressed as follows:

$$y_{norm} = LayerNorm[X + MultiHead(X)] \tag{9}$$

, where X is the input to the multi-head self-attention layer.

Subsequently, the extracted features y_{norm} are sent into two feed-forward layers with skip-connection. Finally, another layer-normalisation operation is adopted to generate the output of the Transformer network. The operation can be expressed as:

$$y_{feedforward} = LayerNorm[y_{norm} + (\max(0, y_{norm}W_1 + b_1)W_2 + b_2)] \quad (10)$$

, where W_1 and W_2 are the weight of the feedforward layers, and b_1 and b_2 are the bias of the feedforward layers.

With the process of the Transformer block, the long-term dependency within data can be further exposed. In the next stage, the output of the Transformer block is flattened and fed into two fully connected layers to yield an RUL prediction.

4. Experimental Setup

4.1. Data Collection and Pre-processing

The XJTU-SY rolling bearings datasets [27] were adopted in this experiment. The XJTU-SY dataset consists of 15 rolling element bearings' run-to-failure sub-datasets. The datasets were collected via an accelerated degradation test. The testing platform is shown in Figure 4. In the accelerated degradation test, there are three operational conditions which are 2100rpm (35Hz) and 12kN, 2250rpm (37.5Hz) and 11kN, and 2400rpm (40Hz) and 10kN. In order to collect the vibration signals, two accelerometers were installed in the horizontal and vertical direction on the housing of the rolling element bearing. The sampling frequency of the accelerometers is 25.6kHz, with the sampling duration set as 1.28s. The sampling interval was set as 1min. The experiment stopped when one of the accelerometers' maximum amplitudes exceeded $10 * A_h$, where A_h is the maximum amplitudes in the normal operation stage. In the accelerated degradation test, four common failure modes of bearings were recorded, which are inner race, wear, cage fracture, outer race wear, and outer race fractures.

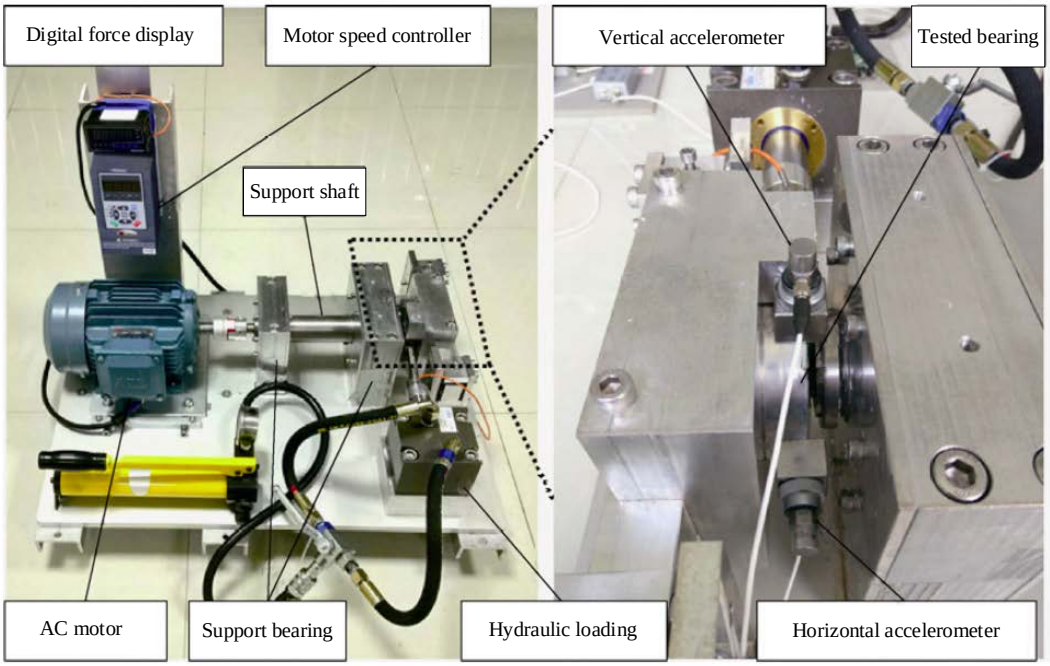


Figure 4. The tested platform of the rolling element bearing

The details of the 15 run-to-failure bearings are listed in Table 1. After the mechanism analysis of all the bearing data, it is worthwhile to mention the characteristic of Bearing1_4 and Bearing3_2. Bearing 1_4 experienced a sudden failure, and the degradation pattern was only partly recorded. Bearing3_2 suffered from various types of failures, and there is a great variation in the vibration signal. Therefore, the data of both bearings were not considered in this study. There are 13 bearings datasets in three different operational conditions that were used in the modelling stage. The FTP of each bearing is calculated according to the kurtosis-based approach. The tolerance of anomaly points number was set as three in this study. It can be seen the FTPs and the actual degradation time were listed in the fourth and fifth columns of Table 1.

Table1. The details of the XJTU-SY dataset

Operational condition	Bearing index	Bearing lifetime (minutes)	FTP (minutes)	RUL after FTP (minutes)	Fault element
35Hz/ 12kN	Bearing1_1	123	76	47	Outer race
	Bearing1_2	161	44	117	Outer race
	Bearing1_3	158	60	98	Outer race
	Bearing1_4	122	-	-	Cage
	Bearing1_5	52	39	13	Inner race and outer race
37.5Hz/ 11kN	Bearing2_1	491	455	36	Inner race
	Bearing2_2	161	48	113	Outer race
	Bearing2_3	533	327	206	Cage
	Bearing2_4	42	32	10	Outer race
	Bearing2_5	339	141	198	Outer race
40Hz/ 10kN	Bearing3_1	2538	2344	194	Outer race
	Bearing3_2	2496	-	-	Inner race, ball, cage and outer race
	Bearing3_3	371	340	31	Inner race
	Bearing3_4	1515	1418	97	Inner race
	Bearing3_5	114	9	105	Outer race

4.2. Experimental and Model Setup

After the determination of FPT, the data was further pre-processed via data normalisation and sliding window processing. The window length was set as 1024 data points, and the strides were set as 256 data points. The data label is RUL, which was normalised to the scale from 0 to 1. The data was then split into the training set and testing set. The modelling was implemented based on different operational conditions. For the dataset in each operational condition, the leave-one-out cross-validation strategy was adopted to yield more comprehensive results. Taking the modelling of Bearing1_3 as an example, the datasets of Bearing1_1, Bearing1_2, and Bearing1_5 were used as training set, and the dataset of Bearing1_3 was adopted as testing set. Subsequently, the data instances in the training set were reshuffled to increase the data integrity. In the model training stage, the modelling was repeated five times and the mean value and standard deviation of the results were marked for further analysis.

The setting of the network parameters can significantly affect the algorithm performance. The parameters of the proposed network are shown in Figure 5. Besides the network structure, the

training parameters also need to be determined. Firstly, Adam was adopted as the optimiser. The learning rate, batch size and training epoch were set as 0.001, 500 and 100, respectively. Early stopping strategy was adopted to get the best results during the training process. Secondly, $l2$ regularisers [45] were added to the model to avoid overfitting. Meanwhile, the He normal initialiser was selected to initialise the parameters of the network, which is helpful to the convergence of the training process. Finally, the mean square error was selected as the loss function since it is suitable for the regression mission.

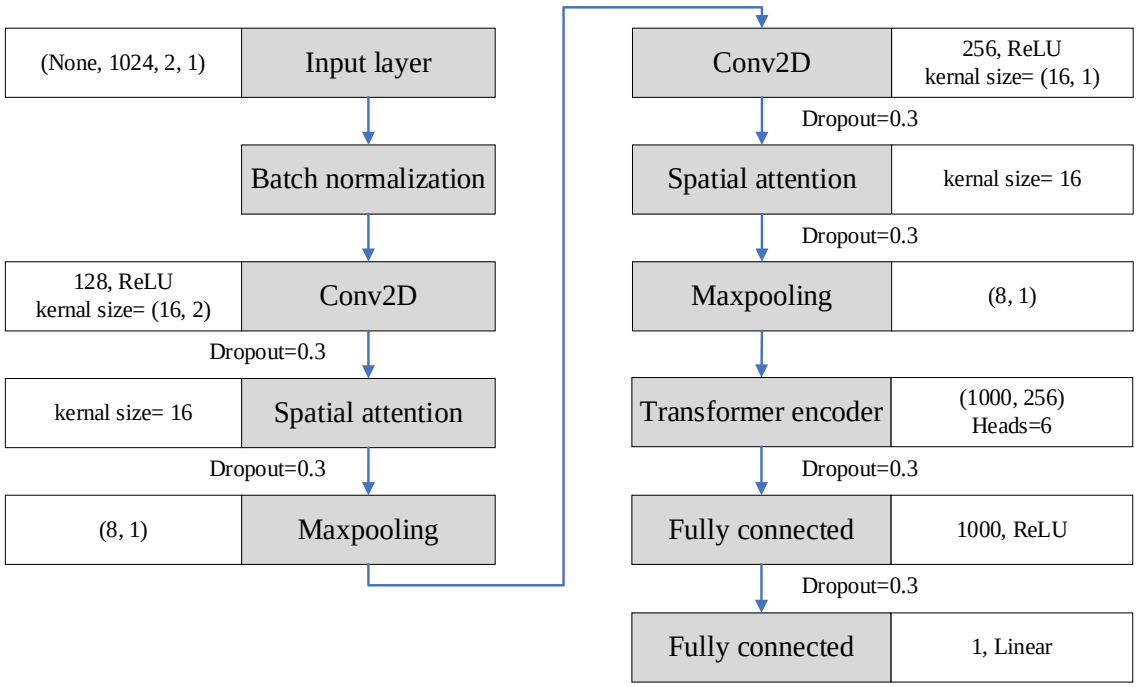
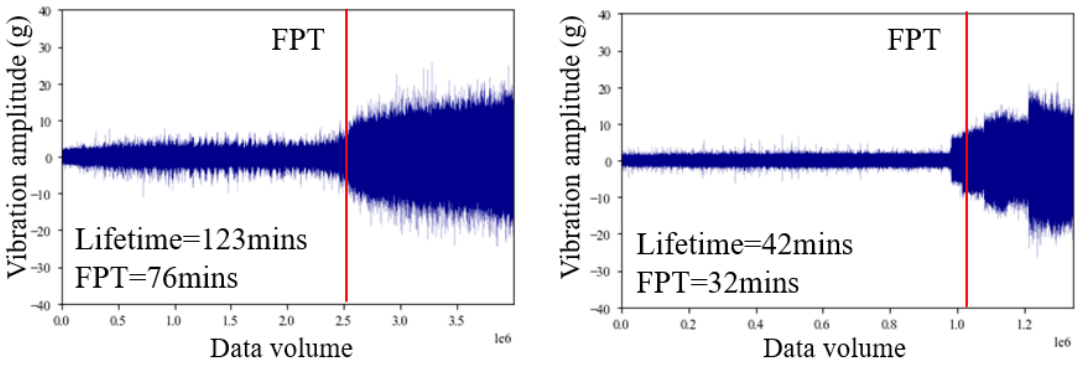


Figure 5. The parameters setting of SAConvFormer

In the experiment, two scenarios were designed to demonstrate the performance of the proposed algorithm. In scenario 1, an ablation experiment was set up. As mentioned in Section 3.3.2, the convolutional kernel size and pooling size are two key parameters in SA-CNN, which affect the input size of the Transformer network. In this scenario, two convolutional layers and two max-pooling layers were adopted in SA-CNN. Different convolutional kernel sizes and pooling sizes were adopted to yield RUL predictions. Bearing 1_1 was adopted to reveal the impact of kernel size and pooling size on the performance of SAConvFormer. After that, all the 13 bearing

datasets were used to evaluate the performance of the benchmarking algorithms. The difference between the lifetime and degradation trends of these three datasets are obvious, which can be used to evaluate the performance of the algorithm in different operational conditions. The Horizontal vibration signals of the two datasets are illustrated in Figure 6.



(a) Bearing 1_1 (b) Bearing 2_4

Figure 6. Demonstration of the horizontal vibration signals of (a) Bearing 1_1 and (b) Bearing 2_4

Three benchmarking algorithms, which are standard CNN, spatial attention-based CNN, and CNN +Transformer network, were adopted to evaluate the impact of different components on the algorithm performance. The standard CNN was designed by replacing the Transformer network with a fully connected layer and removing the spatial attention layers in SAConvFormer. Based on standard CNN, adding spatial attention layers can obtain spatial attention-based CNN. Meanwhile, removing the spatial attention layers in SAConvFormer can yield a CNN +Transformer encoder. It should be noted that the vanilla Transformers network was not adopted in the ablation experiment. The parallel computing mechanism of the Transformers network leads to a huge computational load in the modelling process based on the extremely long sequential input data. As a result, it takes a long time to update the model parameters and the training loss is hard to converge.

In scenario 2, the results of the proposed algorithm were compared with the results released from the state-of-the-art studies and prevailing algorithms investigated in RUL modelling. The

benchmarking algorithms and approaches for bearing RUL modelling include:

(1) BiLSTM network [31]: a neural network that consists of three bi-directional LSTM layers, and a fully connected layer. The number of nodes in the hidden layers was set as 1000.

(2) Dilated CNN [32]: a neural network that consists of four dilated convolutional layers and a fully connected layer. The number of nodes in the dilated convolutional layers and fully connected layer were set as 128 and 1000.

(3) Multiscale CNN [46]: using time-frequency representation of the signals as input. It consists of three convolution, three max-pooling layers, and two fully connected layers. A skip connection is applied in the last convolution layer.

(4) DCNN-bootstrap [9]: using time-frequency representation of the signals as input. It consists of four convolution layers and two fully connected layers. This network is embedded in a bootstrap-based framework that can quantify the RUL prediction interval.

The algorithms and approaches prevailing in PHM were adopted in this scenario to indicate the performance of the proposed algorithm. All tests were conducted on an Intel i9-10920X 3.50Ghz CPU with Nvidia GeForce RTX 3090 graphics card.

4.3. Evaluation Metrics

In order to evaluate the performance of the proposed algorithm with other benchmarking algorithms, mean-absolute-error (MAE) and root-mean-square-error (RMSE) were selected as the evaluation metrics. The MAE is the mean value of the absolute value of the deviation from the arithmetic mean of all individual observations. It can avoid the error counteraction issue of different observations. The MAE can be expressed as:

$$MAE = \frac{\sum_{i=1}^{n} |x_i - x_p|}{n} \tag{11}$$

, where n is the number of observations, x_i is the actual value and x_p is the predicted value.

RMSE can measure the difference between the prediction values and the actual values. It also can reflect the divergence of the prediction error. The expression of RMSE is:

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-x_p)^2}{n}} \tag{12}$$

5. Experimental Results

5.1. Scenario 1: Ablation experiment of the proposed algorithm

The setting of SA-CNN is highly relevant to the algorithm performance of SAConvFormer and the model parameters. The impact of different convolutional kernel sizes and pooling sizes on the algorithm performance and the network parameters of SAConvFormer are shown in table 2. From the results, what evident is that the adoption of a small pooling size causes the large training parameters of the network and the higher RMSE and MAE. When the pooling size was set at 4, the network parameters reached over 15M, which requires a large computational cost. In contrast, the large pooling size, which is 16, can reduce the model parameters greatly, while the RMSE and MAE achieved by SAConvFormer are over 0.152 and 0.122, respectively. Meanwhile, in the comparison of convolutional kernel size, it can be seen that when the kernel size was set at 32, the algorithm performance of SAConvFormer was promoted. The best RMSE and MAE obtained in this experiment are 0.125 and 0.104, respectively, when the kernel size was set at 32 and the pooling size was set at 8. Meanwhile, in this setting, the number of network parameters is 5.9M, which is under the average of all the experiments. Hence, 32 and 8 were adopted as the setting of convolutional kernel size and pooling size in the following experiments.

Table 2. The impact of different SA-CNN settings on the performance of SAConvFormer

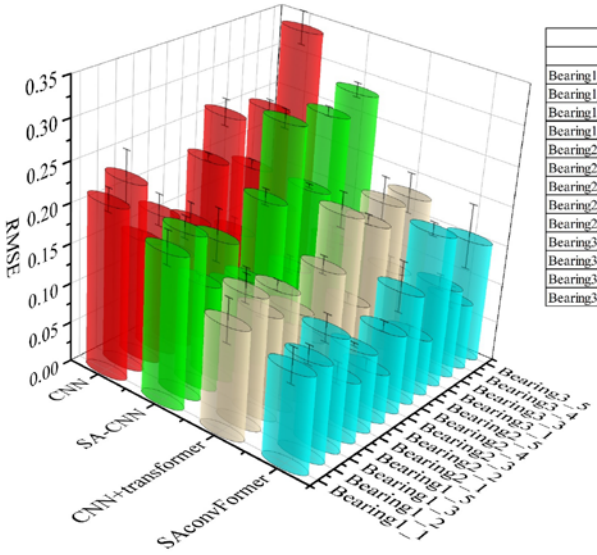
Conv_kernal_size	Pooling_size	# Parameters	Output_size of SA-CNN	RMSE	MAE
8	4	17.9M	(61,256)	0.159±0.0095	0.142±0.0091
8	8	6.1M	(120, 256)	0.138±0.0096	0.115±0.0087
8	16	3.1M	(3, 256)	0.182±0.0089	0.161±0.0086
16	4	17.7M	(59, 256)	0.152±0.0170	0.128±0.0156
16	8	5.9M	(13, 256)	0.142±0.0083	0.123±0.0086
16	16	3,3M	(3, 256)	0.173±0.0174	0.146±0.0168
32	4	16.9M	(54, 256)	0.136±0.0053	0.121±0.0048

32	8	5.9M	(11, 256)	**0.125**±0.0088	**0.104**±0.0079
32	16	3.4M	(1, 256)	0.152±0.0126	0.122±0.0134
64	4	15.4M	(44, 256)	0.182±0.0096	0.162±0.0099
64	8	6,.0M	(7, 256)	0.131±0.0193	0.109±0.0178

In this scenario, four algorithms which are CNN, SA-CNN, CNN+ Transformer and SAConvFormer were adopted to reveal the impact of the spatial attention layer and Transformer on the algorithm performance in terms of RMSE and MAE. The experiments used all the 13 bearing datasets to evaluate the performance of the four algorithms. The modelling results are shown in Figure 7 and Figure 8. It can be seen from Figure 7, that with the enhancement of spatial attention, the RMSE of RUL prediction in most bearing datasets using SA-CNN are promoted. Only Bearing 2_4, Bearing 3_1, and Bearing 3_3 shows worse RMSE with the introduction of SA-CNN. With the introduction of Transformer network to CNN, the RMSE of all the experiments were decreased. Among all the bearing datasets, the experiment implemented based on Bearing 3_5 dataset achieved the greatest promotion, which is 0.158, while the decrease of RMSE of Bearing 3_5 is the smallest, which is mere 0.006. With the enhancement of spatial attention and the Transformer network, SAConvFormer achieved the lowest RMSE in most of the experiments (11 out of 13). The RMSEs achieved by SAConvFormer are higher than CNN +Transformer scheme in the experiments of Bearing 2_4 and Bearing3_1. CNN is the baseline model in this scenario. In comparison with CNN, SAConvFormer lowered the algorithm performance in terms of RMSE by around 0.08 in all the experiments. The greatest progress of the experiments in terms of RMSE is Bearing 3_3 dataset experiment, which progress is 54.7%.

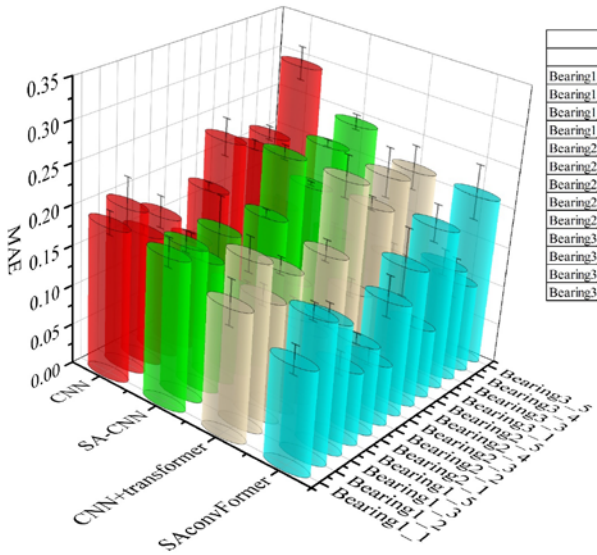
Figure 8 illustrates the comparison of algorithm performance in terms of MAE. Similar to the results of RMSE, spatial attention decreased the MAE of most experiments, except Bearing1_3, Bearing1_5, Bearing2_4 and Bearing3_3. The introduction of Transformer greatly reduced the algorithm performance in terms of MAE in nine bearing datasets experiments, while the MAEs of Bearing 1_2, Bearing 2_4, and Bearing 3_3 increased. Meanwhile, similar to the results of

RMSE, SAConvFormer achieved the lowest RMSE in most of the experiments (10 out of 13). In general, the proposed algorithm achieved the average progress, which is around 0.038 in all the experiments in terms of MAE.



	CNN		SA-CNN		CNN+transformer		SAconvFormer	
	mean	std	mean	std	mean	std	mean	std
Bearing1_1	0.214	0.015	0.189	0.021	0.136	0.027	**0.120**	0.022
Bearing1_2	0.233	0.032	0.204	0.013	0.164	0.017	**0.121**	0.012
Bearing1_3	0.152	0.005	0.132	0.005	**0.130**	0.011	0.141	0.017
Bearing1_5	0.175	0.020	0.175	0.033	0.147	0.007	**0.075**	0.014
Bearing2_1	0.117	0.005	0.106	0.006	0.111	0.022	**0.099**	0.002
Bearing2_2	0.146	0.020	0.105	0.010	0.065	0.008	**0.062**	0.008
Bearing2_3	0.216	0.005	0.195	0.010	0.142	0.011	**0.099**	0.007
Bearing2_4	0.264	0.017	0.283	0.012	0.201	0.022	**0.138**	0.032
Bearing2_5	0.105	0.014	0.089	0.011	0.078	0.007	**0.071**	0.007
Bearing3_1	0.182	0.006	0.183	0.003	**0.170**	0.007	0.190	0.008
Bearing3_3	0.249	0.008	0.269	0.006	0.188	0.027	**0.113**	0.010
Bearing3_4	0.112	0.010	0.097	0.014	0.089	0.014	**0.084**	0.014
Bearing3_5	0.335	0.021	0.281	0.009	0.177	0.026	**0.152**	0.041

Figure 7. The comparison of algorithm performance in terms of RMSE



	CNN		SA-CNN		CNN+transformer		SAconvFormer	
	mean	std	mean	std	mean	std	mean	std
Bearing1_1	0.189	0.015	0.186	0.018	0.160	0.028	**0.129**	0.018
Bearing1_2	0.204	0.032	0.189	0.013	0.206	0.021	**0.174**	0.011
Bearing1_3	0.162	0.005	0.163	0.006	**0.148**	0.013	0.159	0.014
Bearing1_5	0.175	0.020	0.186	0.039	0.173	0.011	**0.090**	0.014
Bearing2_1	0.109	0.003	0.108	0.006	**0.106**	0.018	0.109	0.006
Bearing2_2	0.125	0.005	0.084	0.016	0.075	0.010	**0.074**	0.011
Bearing2_3	0.181	0.005	0.177	0.005	0.162	0.014	**0.139**	0.026
Bearing2_4	0.235	0.024	0.244	0.009	0.244	0.026	**0.171**	0.042
Bearing2_5	0.078	0.008	0.074	0.009	0.095	0.012	**0.086**	0.007
Bearing3_1	0.210	0.006	**0.188**	0.002	0.193	0.008	0.199	0.023
Bearing3_3	0.217	0.009	0.232	0.004	0.223	0.031	**0.141**	0.016
Bearing3_4	0.115	0.009	0.107	0.011	0.108	0.015	**0.111**	0.025
Bearing3_5	0.293	0.021	0.243	0.010	0.213	0.028	**0.211**	0.034

Figure 8. The comparison of algorithm performance in terms of MAE

Figures 7 and 8 demonstrated that the experiment of Bearing 2_2 using SAConvFormer achieved the best algorithm performance in terms of RMSE and MAE, which are 0.065 and

0.051, respectively. The prediction results of the Bearing 2_2 experiment using SAConvFormer was shown in Figure 9. The diagonal line in the figures is the actual RUL of the bearing 1_1 after FPT. The RMSE and MAE in Figure 8 are calculated based on the prediction value and actual value. Kalman filtering [47], as an effective denoising tool, was used to smooth the prediction value to make a clear comparison between the prediction and actual values. It can be seen that the prediction is close to the actual RUL in most stages, while the prediction error in the range from 2000 to 3000 is slightly enlarged.

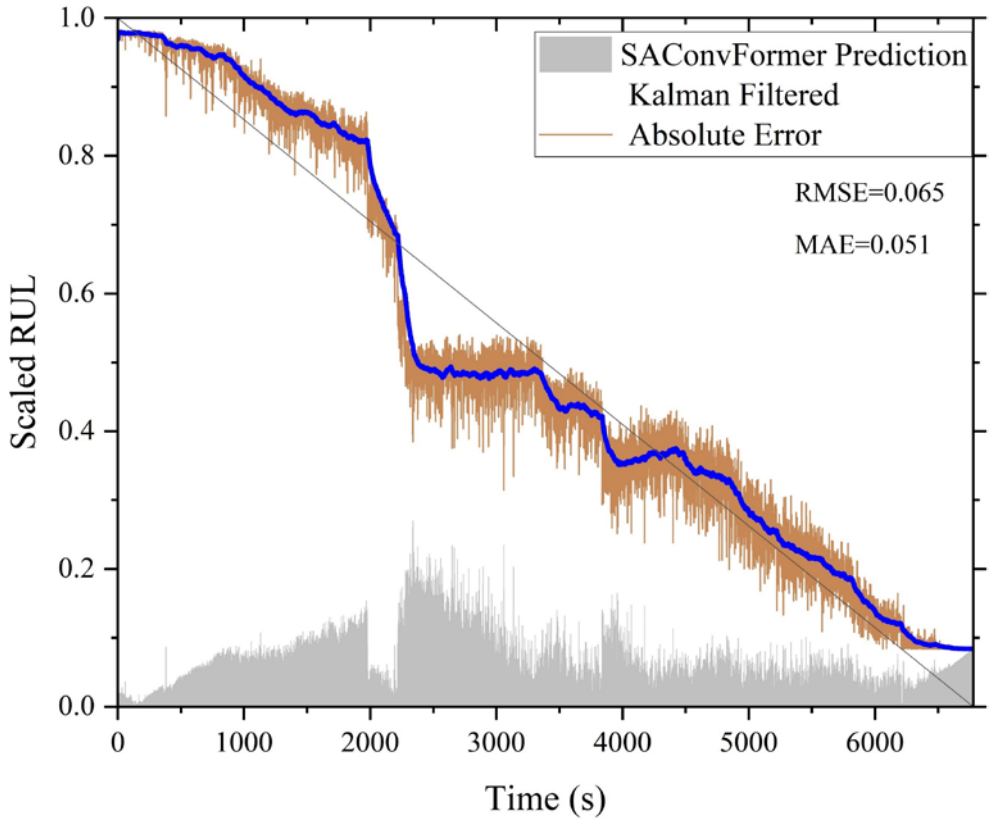


Figure 9. The prediction results for Bearing 2_2

In order to further compare the performance of the algorithms adopted in this scenario, the four algorithms' prediction values of Bearing 1_1 were plotted in Figure 10. It is obvious that the prediction value of all the algorithms experienced a sharp decline in the vicinity of 250s, which can be caused by the sudden acceleration of the bearing failure. The prediction of CNN has the largest error in all the cases. With the help of spatial attention, the prediction error of SA-CNN

is smaller than CNN. SA-CNN showed better performance in the period from 750s to 1500s. In the middle and the later periods of the bearing lifecycle, the prediction of CNN and SA-CNN are similar, where the degradation trend is not significant. With the introduction of Transformer to CNN, the prediction error in the first half is decreased, while the prediction error in the second half is still considerable. When the spatial attention and Transformer were jointly adopted, it can be seen that the prediction error of SAConvFormer was further reduced in all the stages in comparison with the algorithms above.

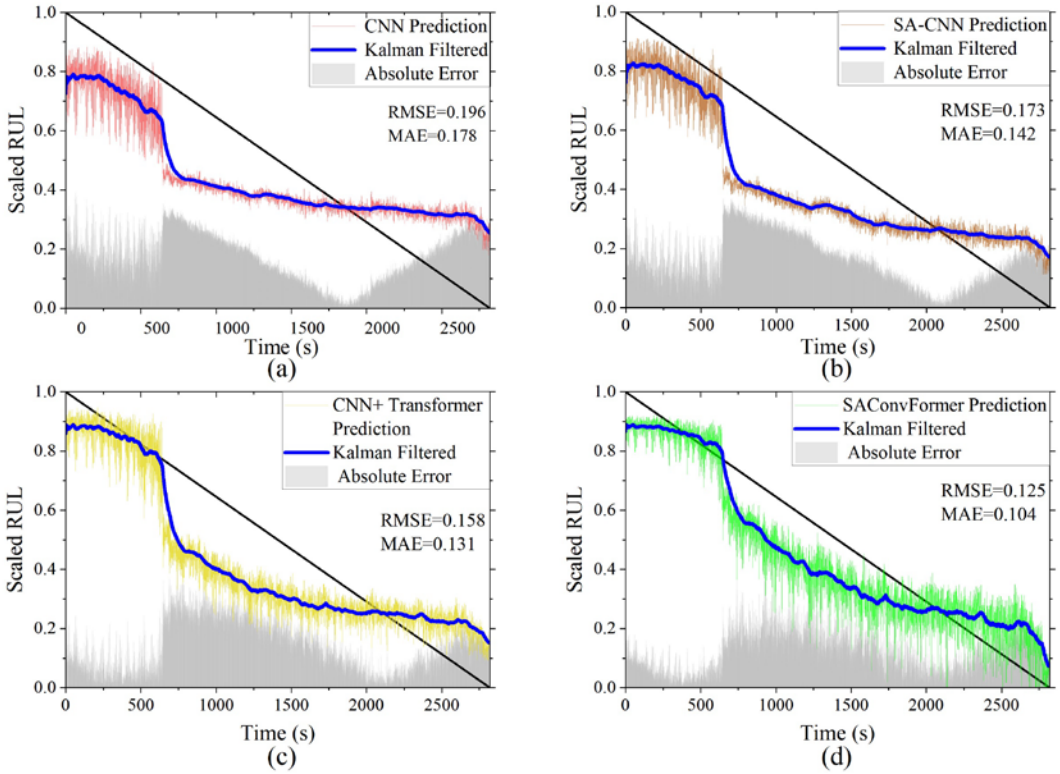


Figure 10. The prediction results for bearing 1_1: (a) CNN; (b) SA-CNN; (C) CNN+Transformer; (d) SAConvFormer

5.2. Scenario 2: Comparison with the state-of-the-art algorithms and methods

In this scenario, two prevailing algorithms, which are BiLSTM [31] network and Dilated CNN [32] were chosen as the benchmark algorithms. Besides, two prevailing bearing RUL modelling methods based on MSCNN [46] and DCNN-MLP [9] were also adopted to evaluate the

effectiveness of the proposed algorithm.

The comparison of algorithm performance in terms of RMSE and MAE for all the bearing datasets is illustrated in Figure 11. In the experiments of 35Hz/12kN operational condition (Bearing 1_1 to Bearing 1_5), it is obvious that the BiLSTM network gets the worst algorithm performance in all the datasets. The algorithm performance in terms of RMSE and MAE of BiLSTM in Bearing 1_5 is dramatically high, which are 0.510 and 0.428. However, in most of the results of 37.5Hz/11kN operational condition (Bearing 2_1 to Bearing 2_5), the algorithm performance of BiLSTM surpasses that of Dilated CNN and MSCNN. In the results of the 40Hz/10kN operational condition (Bearing 3_1 to Bearing 3_5), BiLSTM shows better performance in RMSE and worse performance in MAE. Meanwhile, dilated CNN shows stable algorithm performance in the experiments of 35Hz/12kN operational condition, while its algorithm performance in terms of RMSE and MAE are higher than that of MSCNN, DCNN-MLP and SAConvFormer.

In comparison with the MSCNN and DCNN-MLP based methods that used time-frequency images and handcrafted features as input, the proposed SAConvFormer can achieve better performance in terms of RMSE and MAE in all the bearing datasets. DCNN-MLP achieves better performance in terms of RMSE and MAE in comparison with the rest algorithms. SAConvFormer showed great merits in the results of the 35Hz/12kN operational condition. The difference between RMSE in SAConvFormer and DCNN-MLP for Bearing 1_1, Bearing 1_2, and Bearing 1_4 is 0.077, 0.067 and 0.094, respectively. In the results of the 37.5Hz/11kN operational condition, DCNN-MLP and SAConvFormer tend to yield better results, where the RMSE and MAE are situated around 0.100. the merits of SAconvFormer are not obvious. The largest margin between the RMSE and MAE of SAConvFormer and DCNN-MLP in this group are merely 0.047 and 0.040. For the datasets such as Bearing 3_1 and Bearing 3_5, the MAE of all the benchmarking algorithms and methods are out of the range of 0.200. However, SAConvFormer can achieve the MAE of 0.190 and 0.151.

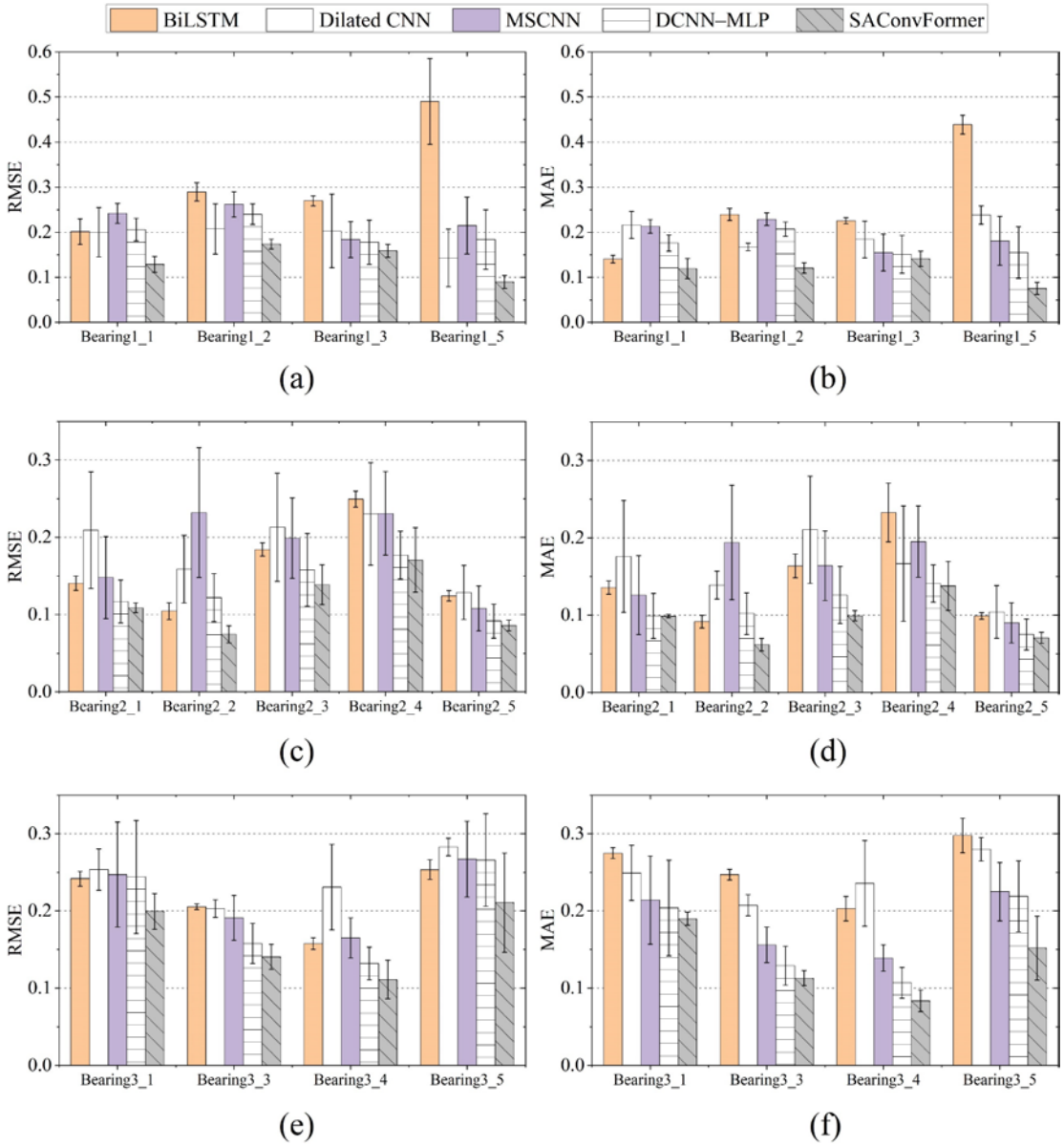


Figure 11. The comparison of algorithm performance in terms of RMSE and MAE for all the bearing datasets

6. Discussion

The combination of CNN and Transformer takes advantage of both algorithms. CNN is effective in feature extraction and dimension reduction, while the extracted features are only processed by fully connected layers, which are short of sequence learning capability. Meanwhile, it is challenging to directly apply the Transformer network for the modelling based on bearing vibration data because the length of the input sequence is extremely long, which

leads to the high computation cost of the algorithm. The high level and low dimension features extracted from CNN are processed in the Transformer, which can greatly leverage the algorithm performance. Furthermore, the introduction of spatial attention is beneficial for the global feature extraction of CNN. The standard CNN does not investigate the relationship between different feature maps, which causes the information loss of the global patterns. This issue can further impede the sequence learning of the Transformer network since the sequence information was not extracted in CNN. Hence, the proposed SAConvFormer can achieve satisfactory performance in the bearing RUL prediction modelling. In scenario 1, the optimal setting of SA-CNN was determined. The output of SA-CNN has a great impact on the algorithm performance of SAConvFormer. Under the context that the structure of the Transformer block is fixed, with smaller convolutional kernel size and pooling size, the output size of SA-CNN tends to be larger, which causes high computation cost and inaccurate RUL prediction. When the convolutional kernel size and pooling size increase, the output size of SA-CNN tends to be smaller, which impedes the Transformer block to learn the hidden patterns relevant to RUL. The results of the ablation experiment reveal the impact of each component in SAConvFormer on performance improvement. It can be seen that the introduction of Transformer can greatly promote the algorithm performance in the experiments of Bearing 1_1 and Bearing 2_4, while its algorithm performance in Bearing 3_1 was worse. The introduction of spatial attention can lower the RMSE and MAE in all three experiments, while its impact is not as considerable as Transformer in the experiment of Bearing 1_1 and Bearing 2_1. The introduction of spatial attention can slightly lower the RMSE and MAE in the experiment of Bearing 3_1. With the combination of Transformer and spatial attention, SAConvFormer achieve the reduction of RMSE and MAE by 23.7% and 16.9% in the experiment of Bearing 3_1, which indicated the effectiveness of the proposed algorithm. Besides, the results of Bearing 2_1 to Bearing 2_5 are generally better than that of the experiments. The Bearing 2_2 achieved the best algorithm performance in terms of RMSE and MAE, which are 0.065 and 0.051. Two reasons might cause the good performance of this experiment. Firstly, the bearing degradation patterns in 37.5Hz/11Kn working conditions can be obvious, where the SAConvFormer is able to identify.

The second reason can be there are five bearing datasets available in the group of 37.5Hz/11Kn working conditions, while only four bearing datasets are available in the group of 35Hz/12kN and 40Hz/10kN. With more training data, the performance of SAConvFormer can be leveraged. With more available datasets in the future, the impact of data size on the algorithm performance will be investigated in future works.

In scenario 2, the results of the benchmarking experiments demonstrated that the proposed algorithm shows merits in all the bearing experiments. The largest margin between the proposed algorithm and the second-best algorithm in terms of RMSE and Mae are 0.077 and 0.079, while the smallest margin is 0.005 and 0.004. The advantage of SAConvFormer is considerable in the experiment of Bearing 3_5. Bearing 3_5 has an extremely short lifecycle which is 114mins. The characteristic of this dataset poses a challenge for the algorithms to learn its degradation patterns. The benchmarking algorithms can only get RMSE and MAE about 0.25 and 0.22, while the proposed algorithm can achieve RMSE and MAE about 0.210 and 0.152.

Another merit of the proposed algorithm is its applicability. In the benchmarking experiments, BiLSTM and dilated CNN used the same input data of the proposed algorithm, while MSCNN and DCNN-MLP utilised the time-frequency image as input data. The time-frequency transformation can expose the frequency-domain features that are relevant to the bearing failure. In the existing studies, using time-frequency transformation and other signal processing approaches to generate new features can promote the algorithm performance, while it requires extra domain knowledge and procedures. In this study, the proposed SAConvFormer only use the raw vibration data as input to construct an end-to-end RUL prediction model without priori knowledge, which is easy to deploy in the actual industrial scenario.

Currently, the proposed algorithm can only be applicable in the degradation stage of bearing. Before the deployment of the proposed algorithm, the determination of FPT is still needed. The FPT determination adopted in this study is the Kurtosis-based approach, which is widely used

in the bearing RUL prediction modelling. If the determination of FPT and the degradation modelling can be jointly processed in the same model, the deployment of the RUL prediction model can be further simplified in the real world, which will accelerate the application of the data-driven RUL prediction approach. Hence, a comprehensive algorithm that can jointly determine the FPT and predict the RUL will be investigated in the future. Meanwhile, the distribution of RUL was not considered in this study. The estimation of the RUL uncertainty relies on the deployment of statistical or reliability models, which requires prioi knowledge and extra efforts. In the future, investigating an end-to-end RUL modelling approach that can measure the uncertainty of the predicted results is necessary.

7. Conclusion

Bearing is an essential part of the industry. An accurate RUL prediction can bring tangible benefits to bearing maintenance management. The existing studies of bearing RUL prediction heavily relied on feature engineering to improve the prediction accuracy, which increases the complexity and difficulty in actual deployment. In this study, an end-to-end bearing RUL modelling algorithm called SAConvFormer was proposed to establish an RUL prediction model without priori knowledge. The major findings of this study are: (1) the adoption of CNN and spatial attention mechanism is effective in extracting the local and global sequential features, which are then processed by a Transformer network; (2) the Transformer network is able to learn the global patterns which is relevant the bearing health status; (3) compared to the existing RUL modelling algorithms and methods, the proposed algorithm has demonstrated its merit in the experiments. Based on the promising performance of the SAConvFormer, the bearing maintenance strategy can be optimised to avoid catastrophic breakdown, achieve better job scheduling, and lower maintenance costs. In future works, the joint determination of FPT and RUL and the estimation of the RUL uncertainty will be further investigated.

Acknowledgement

Our work is supported by multiple funds in China, including the Key Program of NSFC-Guangdong Joint Funds (U1801263), the Natural Science Foundation of Guangdong Province (2020A1515010890, 2020B1515120010), the Dedicated Fund for Promoting High-Quality Economic Development in Guangdong Province (Marine Economic Development Project) GDNRC[2021]44, and the Major project of science and technology plan of Foshan City (1920001001367). Our work is also supported by Guangdong Provincial Key Laboratory of Cyber-Physical System (2020B1212060069).

References

[1] Lei Y, Li N, Guo L, Li N, Yan T, Lin J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*. 2018;**104**:799-834.
[2] Zio E. Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*. 2022;**218**:108119.
[3] Liu H, Liu Z, Jia W, Lin X. Remaining useful life prediction using a novel feature-attention-based end-to-end approach. *IEEE Transactions on Industrial Informatics*. 2020;**17**(2):1197-207.
[4] Yan M, Wang X, Wang B, Chang M, Muhammad I. Bearing remaining useful life prediction using support vector machine and hybrid degradation tracking model. *ISA Transactions*. 2020;**98**:471-82.
[5] Li N, Lei Y, Lin J, Ding SX. An improved exponential model for predicting remaining useful life of rolling element bearings. *IEEE Transactions on Industrial Electronics*. 2015;**62**(12):7762-73.
[6] Li J, Chen R, Huang X. A sequence-to-sequence remaining useful life prediction method combining unsupervised LSTM encoding-decoding and temporal convolutional network. *Measurement Science and Technology*. 2022.
[7] Chen C, Liu Y, Sun X, Di Cairano-Gilfedder C, Titmus S. An integrated deep learning-based approach for automobile maintenance prediction with GIS data. *Reliability Engineering & System Safety*. 2021;**216**:107919.
[8] Chen D, Qin Y, Wang Y, Zhou J. Health indicator construction by quadratic function-based deep convolutional auto-encoder and its application into bearing RUL prediction. *ISA transactions*. 2021;**114**:44-56.
[9] Huang C-G, Huang H-Z, Li Y-F, Peng W. A novel deep convolutional neural network-bootstrap integrated method for RUL prediction of rolling bearing. *Journal of Manufacturing Systems*. 2021.
[10] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. *Neurocomputing*. 2021;**452**:48-62.
[11] Humphreys GW, Sui J. Attentional control and the self: the Self-Attention Network (SAN). *Cognitive neuroscience*. 2016;**7**(1-4):5-17.
[12] Mnih V, Heess N, Graves A, editors. Recurrent models of visual attention. Advances in neural information processing systems; 2014.

[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I, editors. Attention is all you need. Advances in neural information processing systems; 2017.
[14] Woo S, Park J, Lee J-Y, Kweon IS, editors. Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV); 2018.
[15] Choi M, Kim H, Han B, Xu N, Lee KM, editors. Channel attention is all you need for video frame interpolation. Proceedings of the AAAI Conference on Artificial Intelligence; 2020.
[16] Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*. 2021;**34**.
[17] Li S, Jin X, Xuan Y, Zhou X, Chen W, Wang Y-X, Yan X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in Neural Information Processing Systems*. 2019;**32**.
[18] Wang B, Lei Y, Li N, Yan T. Deep separable convolutional network for remaining useful life prediction of machinery. *Mechanical Systems and Signal Processing*. 2019;**134**:106330.
[19] Yu Y, Wang C, Gu X, Li J. A novel deep learning-based method for damage identification of smart building structures. *Structural Health Monitoring*. 2019;**18**(1):143-63.
[20] Yu Y, Rashidi M, Samali B, Mohammadi M, Nguyen TN, Zhou X. Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm. *Structural Health Monitoring*. 2022:14759217211053546.
[21] Li T, Zhao Z, Sun C, Cheng L, Chen X, Yan R, Gao RX. WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 2021.
[22] Li X, Zhang W, Ding Q. Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliability engineering & system safety*. 2019;**182**:208-18.
[23] Wang D, Liu K, Zhang X. A Generic Indirect Deep Learning Approach for Multisensor Degradation Modeling. *IEEE Transactions on Automation Science and Engineering*. 2021.
[24] Zhao B, Zhang X, Zhan Z, Wu Q. A robust construction of normalized CNN for online intelligent condition monitoring of rolling bearings considering variable working conditions and sources. *Measurement*. 2021;**174**:108973.
[25] Liu J, Pan C, Lei F, Hu D, Zuo H. Fault prediction of bearings based on LSTM and statistical process analysis. *Reliability Engineering & System Safety*. 2021;**214**:107646.
[26] She D, Jia M, Pecht MG. Sparse auto-encoder with regularization method for health indicator construction and remaining useful life prediction of rolling bearing. *Measurement Science and Technology*. 2020;**31**(10):105005.
[27] Wang B, Lei Y, Li N, Li N. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*. 2018;**69**(1):401-12.
[28] Tang J, Zheng G, He D, Ding X, Huang W, Shao Y, Wang L. Rolling bearing remaining useful life prediction via weight tracking relevance vector machine. *Measurement Science and Technology*. 2020;**32**(2):024006.
[29] Gao Z, Liu Y, Wang Q, Wang J, Luo Y. Ensemble empirical mode decomposition energy moment entropy and enhanced long short-term memory for early fault prediction of bearing. *Measurement*. 2021:110417.
[30] Zou Y, Li Z, Liu Y, Zhao S, Liu Y, Ding G. A Method for Predicting the Remaining Useful Life of Rolling Bearings Under Different Working Conditions Based on Multi-domain Adversarial Networks. *Measurement*. 2021:110393.
[31] Zhang J, Wang P, Yan R, Gao RX. Long short-term memory for machine remaining life prediction. *Journal of Manufacturing Systems*. 2018;**48**:78-86.
[32] Xu X, Wu Q, Li X, Huang B. Dilated Convolution Neural Network for Remaining Useful Life Prediction. *Journal of Computing and Information Science in Engineering*. 2020;**20**(2).
[33] Zeng F, Li Y, Jiang Y, Song G. A deep attention residual neural network-based remaining useful life prediction of machinery. *Measurement*. 2021:109642.
[34] Li X, Zhang W, Ma H, Luo Z, Li X. Data alignments in machinery remaining useful life prediction using deep adversarial neural networks. *Knowledge-Based Systems*. 2020:105843.
[35] Duan Y, Li H, He M, Zhao D. A BiGRU Autoencoder Remaining Useful Life Prediction Scheme With Attention Mechanism and Skip Connection. *IEEE Sensors Journal*.

2021;**21**(9):10905-14.
[36] Chen Y, Peng G, Zhu Z, Li S. A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. *Applied Soft Computing*. 2020;**86**:105919.
[37] Wang Y, Deng L, Zheng L, Gao RX. Temporal convolutional network with soft thresholding and attention mechanism for machinery prognostics. *Journal of Manufacturing Systems*. 2021;**60**:512-26.
[38] Huang X, Zhang P, Shi W, Dong S, Wen G, Lin H, Chen X. Frequency Hoyer attention based convolutional neural network for remaining useful life prediction of machinery. *Measurement Science and Technology*. 2021;**32**(12):125108.
[39] Liu H, Liu Z, Jia W, Lin X, Zhang S. A novel transformer-based neural network model for tool wear estimation. *Measurement Science and Technology*. 2020;**31**(6):065106.
[40] Mo Y, Wu Q, Li X, Huang B. Remaining useful life estimation via transformer encoder enhanced by a gated convolutional unit. *Journal of Intelligent Manufacturing*. 2021:1-10.
[41] Jin C-c, Chen X. An end-to-end framework combining time–frequency expert knowledge and modified transformer networks for vibration signal classification. *Expert Systems with Applications*. 2021;**171**:114570.
[42] Liu Z, Liu H, Jia W, Zhang D, Tan J. A multi-head neural network with unsymmetrical constraints for remaining useful life prediction. *Advanced Engineering Informatics*. 2021;**50**:101396.
[43] He K, Zhang X, Ren S, Sun J, editors. Identity mappings in deep residual networks. European conference on computer vision; 2016: Springer.
[44] Ba JL, Kiros JR, Hinton GE. Layer normalization. *arXiv preprint arXiv:1607.06450*. 2016.
[45] He K, Zhang X, Ren S, Sun J, editors. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision; 2015.
[46] Zhu J, Chen N, Peng W. Estimation of bearing remaining useful life based on multiscale convolutional neural network. *IEEE Transactions on Industrial Electronics*. 2018;**66**(4):3208-16.
[47] Bishop G, Welch G. An introduction to the kalman filter. *Proc of SIGGRAPH, Course*. 2001;**8**(27599-23175):41.