scientific reports

Check for updates

OPEN

A novel transformer-based DL model enhanced by position-sensitive attention and gated hierarchical LSTM for aero-engine RUL prediction

Xinping Chen

Accurate prediction of remaining useful life (RUL) for aircraft engines is essential for proactive maintenance and safety assurance. However, existing methods such as physics-based models, classical recurrent neural networks, and convolutional neural networks face limitations in capturing long-term dependencies and modeling complex degradation patterns. In this study, we propose a novel deep-learning model based on the Transformer architecture to address these limitations. Specifically, to address the issue of insensitivity to local context in the attention mechanism employed by the Transformer encoder, we introduce a position-sensitive self-attention (PSA) unit to enhance the model's ability to incorporate local context by attending to the positional relationships of the input data at each time step. Additionally, a gated hierarchical long short-term memory network (GHLSTM) is designed to perform regression prediction at different time scales on the latent features, thereby improving the accuracy of RUL estimation for mechanical equipment. Experiments on the C-MAPSS dataset demonstrate that the proposed model outperforms existing methods in RUL prediction, showcasing its effectiveness in modeling complex degradation patterns and long-term dependencies.

Keywords RUL prediction, Transformer encoder, Aero-engine, Attention, LSTM

Accurate prediction of remaining useful life (RUL) is crucial for proactive maintenance, reducing casualties and economic losses. RUL prediction methods are classified into physics-based, artificial intelligence-based, and hybrid models[1–7]. Physics-based methods use specific models based on failure mechanisms to explain degradation patterns and integrate real-time monitoring data for RUL assessment. However, they face limitations in complex mechanical systems[8–11]. Artificial intelligence methods learn degradation patterns from observational data without expert knowledge. They excel in predicting complex systems where physical or statistical models are inadequate and have gained attention with advancements in technology[12]. Hybrid methods integrate the advantages of different approaches but may be limited in complex rotating machinery[13].

With the accumulation of valuable data and the rapid advancement of computing power, deep learning (DL) has become a hot topic and has been successfully applied in various engineering fields. DL + PHM has gained popularity in both academia and industry. For instance, in the early days, some methods employed classical RNN models for regression tasks on time series data. However, RNN models face challenges such as the gradient vanishing or exploding problem[14], limiting their performance in long sequence prediction tasks. As a solution, RNN variants like LSTM[15,16] and GRU[17] emerged, which use nonlinear gating mechanisms to control the flow of information and alleviate these limitations to some extent. The research on using gated networks for RUL prediction has been growing rapidly. Zhang et al.[18] proposed an LSTM-Fusion network structure for estimating the RUL of aircraft engines. This network integrates observation sequences of different lengths to extract hidden information effectively. Miao et al.[19] introduced a novel dual-task stacked LSTM method that simultaneously evaluates the degradation stages and predicts the RUL of aircraft engines. Liu et al.[20] presented a multi-level prediction approach for aircraft engine health using LSTM and statistical process analysis for bearing fault prediction. Zhang et al.[21] proposed a dual-task network structure based on bidirectional GRU and a mixture of

College of Artificial Intelligence and Big Data, Chongqing College of Electronic Engineering, Chongqing 401331, China. email: 202321001@cqcet.edu.cn

multiple gating expert units. This structure enables simultaneous evaluation of aircraft engine health status and prediction of the RUL. Ma et al.[22] introduced a new prediction model based on deep wavelet sequence gated recurrent units for RUL prediction of rotating machinery. The proposed wavelet sequence gated recurrent units generate wavelet sequences of different scales through a wavelet layer. Xiao et al.[23] enhanced the robustness of the BLSTM model for RUL prediction by adding Gaussian white noise to the health indicators based on principal component analysis. Song et al.[24] constructed aircraft engine health indicators using variational autoencoders and employed the BLSTM model for RUL prediction.

In addition to enhancing the model's temporal data processing capability using recurrent approaches, another alternative is the use of convolutional neural networks (CNNs), which employ shared receptive fields to improve spatial feature extraction[25]. CNN-based models have also been successfully applied in RUL prediction and have shown competitive performance. Zhu et al.[26] proposed a multi-scale CNN for predicting the RUL of bearings. Compared to traditional CNNs, this network maintains synchronization of global and local information. Li et al.[27] introduced a new approach based on deep CNNs for RUL prediction using raw data. Yang et al.[28] employed a dual CNN model for RUL prediction. In this model, the first CNN model identifies early fault points, while the second CNN model predicts the RUL. Jiang et al.[29] transformed time series data into multi-channel data and used CNN to construct health indicators, leading to improved accuracy in residual life prediction.

The Transformer model[30–32], as one of the most popular deep learning architectures in recent years, has been introduced for sequence data modeling. It efficiently handles long sequences of parallel data and can be applied to time series data of varying lengths. It has achieved remarkable success in various industrial applications, including natural language processing[33], machine vision[34], medical diagnosis[35], and more. In recent years, it has also been gradually applied in the field of RUL prediction. Zhang et al.[36] introduced a novel Transformer-based bidirectional self-attention deep model for RUL prediction. This method is a fully self-attention-based encoder-decoder structure without any RNN/CNN modules. Su et al.[37] proposed an adaptive Transformer that combines attention mechanisms and recurrent structures for predicting the RUL of rolling bearings. It directly models the relationship between shallow features and RUL, mitigating the vanishing gradient problem and better representing complex time degradation patterns. Based on the proposed shared temporal attention layer, Chadha et al.[38] developed two Transformer models specifically designed for handling multivariate time series data and applied them to predict the RUL of aircraft engines. Chang et al.[39] proposed a novel Transformer model for RUL prediction based on a sparse multi-head self-attention mechanism and knowledge distillation technique. It effectively reduces the computational burden of the model and improves domain adaptation capability for raw signal data of rolling bearings. Ren et al.[40] introduced a T2 tensor-assisted multiscale Transformer model to accurately predict the RUL of industrial components. Ding et al.[41] presented a new convolutional Transformer model capable of extracting degradation-related information from both local and global original signals.

In this study, we propose a DL model based on a Transformer-based auto-encoder for the task of RUL prediction. Unlike RNN and CNN models, the Transformer architecture allows for the processing of a sequence of data in a single pass by leveraging attention mechanisms, enabling access to any part of the historical data without being limited by distance. This makes it potentially more powerful in capturing long-term dependencies. However, the adopted dot-product self-attention in Transformers results in the extracted high-level features being insensitive to their local context at each time step[34], which requires the model to invest more effort in estimating the corresponding RUL. Therefore, we introduce position-aware self-attention units (PSA) to enhance the model's ability to focus on the positional relationships of the input data at each time step and improve the incorporation of local context. Additionally, to leverage the improved features extracted by the encoder, we design a gated hierarchical long short-term memory network (GHLSTM) for regression predictions at different time scales, further enhancing the accuracy of RUL prediction for mechanical equipment. The main contributions in the article are as follows.

(1) The traditional attention mechanism used in the Transformer encoder is insensitive to the local context, which is essential for predicting remaining useful life. The proposed position-aware self-attention (PSA) mechanism captures the positional relationships of input data, enabling the model to incorporate local context and generate more effective hidden features. This leads to improved accuracy in predicting remaining useful life.
(2) For enhancing the ability to model long-term dependencies and improve performance in handling large-scale sequential data, the gated hierarchical long short-term memory (GHLSTM) network is proposed, which learns features at different time scales, enables regression predictions at multiple scales, and provides comprehensive feature learning. This results in improved accuracy in predicting RUL.
(3) Experimental results on a widely used aerospace dataset demonstrate the superiority of our proposed method over other existing methods based on quantitative evaluation metrics.

The outline of the article is as follows. Section "Introduction" provides an introduction to the research topic. Section "Theoretical basis" presents the theoretical basis. Section "Proposed methodology" gives a detailed deduction of the proposed DL model. Section "Experimental analysis" is the content of experiments and relevant analysis. And finally, a conclusion is given in Section "Conclusion".

Theoretical basis

The Transformer was first introduced in 2017 for NLP tasks[42]. It is a sequence-to-sequence model that essentially functions as an auto-encoder, composed of a sophisticated encoder module and a decoder module. The encoder module maps the input sequence to a high-dimensional hidden vector, which is then fed into the decoder to generate the output sequence. Unlike recurrent networks with their sequential data input nature, the Transformer

is capable of capturing long-term dependencies by utilizing self-attention mechanisms based on dot products. Transformer-based models have achieved remarkable performance in various time series tasks, including natural language processing, computer vision, and PHM.

The proposed model primarily focuses on the improved structure of the Transformer encoder module. Therefore, in this section, we provide a detailed explanation of the main components and architecture of the Transformer encoder module. The Transformer encoder structure, as shown in Fig. 1, mainly consists of multi-head attention, feed-forward networks, and position encoding.

Multi-head self-attention

The multi-head self-attention mechanism is a variant of the attention mechanism widely used in natural language processing and machine translation tasks. It is an extension of the self-attention mechanism designed to enhance the modeling capacity of the model for different semantic information. The self-attention mechanism allows the model to interact and exchange information between different positions in the input sequence, while the multi-head self-attention mechanism further expands this interactive capability. It achieves this by applying the attention mechanism to different projections in multiple subspaces, creating multiple attention heads. Each attention head has its own set of parameters and can learn different attention weights to capture the associations between different semantic information.

The calculation process of the multi-head self-attention mechanism is as follows. We project the input sequence, i.e. $\mathbf{f} = \{f_i\}_{i=1}^{t}$ with f_i *w.r.t* x_i and $f_i \in R^d$, into multiple subspaces through linear transformations. For each attention head, we use different parameter matrices to perform the projection, obtaining representations for each sub-space. We denote the parallel attention calculations as H, which represents the multi-head attention mechanism:

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(\{head_j\}_{j=1}^{H})\mathbf{W}^A, \tag{1}$$

$$head_j = Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V})_j = soft\max(\frac{\mathbf{Q}_j\mathbf{K}_j^T}{\sqrt{d_k}})\mathbf{V}_j, \tag{2}$$

$$\begin{aligned} \mathbf{K}_j &= \mathbf{f}\mathbf{W}_j^k \\ \mathbf{V}_j &= \mathbf{f}\mathbf{W}_j^v \\ \mathbf{Q}_j &= \mathbf{f}\mathbf{W}_j^q \end{aligned} \tag{3}$$

where $\mathbf{W}^A \in \mathbf{R}^{H_{d_k} \times d}$ and $d_k = d/H$;$\mathbf{k}_j$,$\mathbf{V}_j$ and $\mathbf{Q}_j$ are the key, value and query vectors; $head_j$ is the jth attention head; $\mathbf{W}_j^k$, $\mathbf{W}_j^v$, $\mathbf{W}_j^q \in \mathbf{R}^{d \times d_k}$ are the trainable matrixes.

Feedforward neural network and position encoding

The feed-forward NN is composed of two full connection (FC) layers with ReLU activation function, whose formula is as follows,

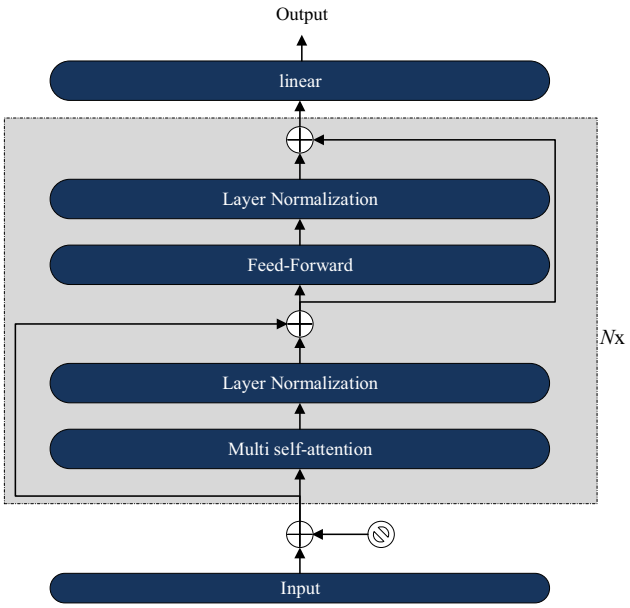

Figure 1. The diagram of the Transformer encoder module.

$$F(\mathbf{x}) = \mathbf{W}_2 \cdot \mathrm{Re}LU(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2, \tag{4}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the weights and bias of the following connected FC layers,$\mathbf{x}$ is the input of the forward neural network.

The formula of position encoding is demonstrated as follows,

$$\begin{aligned} p_i^{(2s)} &= \sin(i/10000^{2s/d}) \\ p_i^{(2s+1)} &= \cos(i/10000^{2s/d}) \end{aligned} \tag{5}$$

By the above design, for given input with any length l, p_i and p_{i+l} has a linear relationship, which helps the regression model learn the sequence relationship effectively. Thus the final input of the transformer encoder module is $\mathbf{X} = \mathbf{x} + \mathbf{p}$.

Proposed methodology

The proposed enhanced Transformer model

The proposed enhanced Transformer model consists of three parts: the feature extraction module, the encoding module, and the regression module, as shown in Fig. 2. The feature extraction module consists of a simple fully connected (FC) layer and position encoding, which performs a simple non-linear dimensionality reduction on the multidimensional raw data and incorporates positional information. The encoding module further compresses and extracts valuable latent features from the extracted features. Compared to the encoding module of traditional Transformers, the proposed model mainly adopts Position-Sensitive Attention (PSA) to replace the self-attention component, enabling the encoding module to capture more contextual information. The PSA unit is integrated to address the insensitivity to local context in the Transformer encoder, thus enhancing the model's ability to incorporate positional relationships and local context at each time step. PSA collectively contributes to the generation of latent features with higher efficacy, which benefits the remaining useful life prediction in the regression module. The regression module utilizes the proposed GHLSTM with multiple hidden features at different time scales for regression prediction. Compared to ordinary linear regression or recursive network-based regression, it can more effectively learn from hidden features, thereby improving the accuracy of RUL predictions.

Supposed that for each sample i, the predicted RUL is $\overline{Rul}_i$ and the true RUL is Rul_i. Mean square error (MSE) is adopted as a loss function to tune the learnable parameters θ of the proposed enhanced Transformer model during the training stage by the optimization Adam, whose formula is given below,

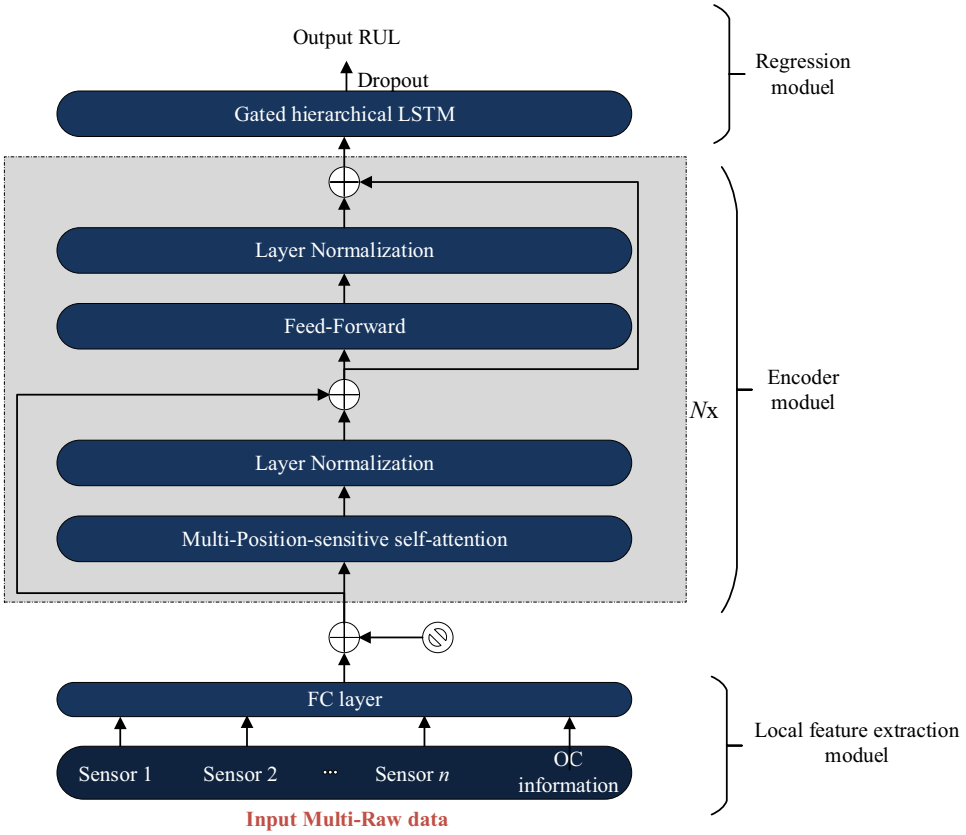

Figure 2. The proposed enhanced Transformer model.

$$L(MSE,\theta)=\frac{1}{2}\sum_{i=1}^{N}\left(\overline{Rul}_i-Rul_i\right)^2. \tag{6}$$

Table 1 shows the hyper-parameters of the proposed DL method. The optimized hyper-parameters of the model are obtained by the grid search. The pseudocode of the proposed prediction method has been summarized in Table 2.

Position-sensitive self-attention (PSA)

To overcome the issue of insensitivity of high-level features to local context in Transformer encoders, we introduce a position-aware self-attention (PSA) unit in our proposed model. This improvement enables the model to focus on the positional relationships of input data at each time step, thereby enhancing its capability to capture local context. Consequently, this approach computes similarity scores between each input element and all other elements, considering both content and positional encodings. Attention weights are then computed based on these scores, and the output is formed by taking the weighted sum of the inputs. By incorporating positional information, the PSA mechanism enhances the model's ability to capture local context, leading to more accurate attention weights and improved feature representations, generating more effective hidden features for accurately predicting the RUL of mechanical equipment. This enhanced sensitivity to local context is crucial for accurately predicting RUL. The deduction of PSA is described as follows.

(1) Construction of the input:

The input of PSA consists of the input sequence $\mathbf{x}=\{x_1\ x_2\ \cdots\ x_{n-1}\ x_n\}$ and the relevant position encoding $\mathbf{p}=\{p_1\ p_2\ \cdots\ p_{n-1}\ p_n\}$, where x_i is the feature representation of the ith element, and p_i is the position encoding of the position i whose formulas are Eq. (5).

(2) The calculation of the similarity score:

Sub layer	Hyperparameter value	Sub layer	Hyperparameter value
Linear	14	Number of encoder module	2
MPSA	4	Learning rate	0.005
Feedforward	128	Output layer	1
GHLSTM	15	Dropout	0.2

Table 1. The hyper-parameters of the proposed enhanced Transformer model.

Algorithm: RUL prediction based on the enhanced Transformer

Local feature extraction module:

Collect the14 sensor measurements and the three operating condition (OC) information of the aircraft engine as the input $\mathbf{x}=\{x_1\ \ x_2\ \ \cdots\ \ x_{n-1}\ \ x_n\}$ into the FC layer;

By Eq.(8-9) to obtain the relevant position encoding $\mathbf{p}=\{p_1\ \ p_2\ \ \cdots\ \ p_{n-1}\ \ p_n\}$ of input $\mathbf{x}$;

Encoder module:

Position-sensitive self-attention(PSA):

Calculate the similarity Score s_{ij} of other elements x_j in the input sequence $\mathbf{x}$ by Eq.(7);

Calculate the attention weights w_{ij} of x_j based on s_{ij} by Eq.(8);

Generate the output $\tilde{x}_i$ of PSA by Eq.(9) based on w_{ij} and x_j.

Layer Normalization:

Linear Normalized the output matrix $\tilde{\mathbf{x}}$ of PSA to get $\tilde{\mathbf{x}}_1$;

Add the normalized matrix and the output of the local feature extraction module $\mathbf{x}_1=\tilde{\mathbf{x}}_1+\mathbf{x}+\mathbf{p}$.

Input into the feed-forward layer by Eq.(4), linear normalized again to get $\hat{\mathbf{x}}_1$, and add with $\mathbf{x}_1$ to obtain the output of the encoder module.

Repeat the encoder module N times to get the final input of GHLSTM $\mathbf{x}^1$**.**

Regression module:

The whole input $\mathbf{x}^1$ and the half of input $\mathbf{x}^2$ are input into the GHLSTM layer to get the output RUL $\overline{Rul}_i$ by Eq.(10-22).

For the *number of epochs* ***do***

Train the enhanced Transformer by Eq.(6);

Optimizer: Adam optimizer.

End

Output: The trained enhanced Transformer

Table 2. The pseudocode of the proposed RUL prediction method.

For each element x_i in the input sequence, calculate the similarity Score s_{ij} of other elements x_j in the input sequence $\mathbf{x}$, meantime considering the influence of position encoding p_{ij}, thus the formula is deduced as follows,

$$\begin{aligned} s_{ij}^1 &= similarity\left(x_i \quad p_{ij}\right) = \left(x_i \cdot p_{ij}\right) / \left(\left\|x_i\right\| * \left\|p_{ij}\right\|\right) \\ s_{ij}^2 &= similarity\left(x_i \quad p_{ij}\right) = \left(x_i \cdot p_{ij}\right) / \left(\left\|x_i\right\| * \left\|p_{ij}\right\|\right) \\ s_{ij} &= similarity\left(s_{ij}^1 \quad s_{ij}^2\right) = \left(s_{ij}^1 \cdot s_{ij}^2\right) / \left(\left\|s_{ij}^1\right\| * \left\|s_{ij}^2\right\|\right) \end{aligned} \tag{7}$$

(3) The calculation of attention weights:
For each element x_i, calculating the attention weights w_{ij} based on the similarity Score,

$$w_{ij} = soft\ \max\left(\mathrm{s}_{ij}\right) = \frac{\exp\left(\mathrm{s}_{ij}\right)}{\sum_{k=1}^{n} \exp\left(\mathrm{s}_{ik}\right)}. \tag{8}$$

(4) Finally, the output element $\tilde{x}_i$ after PSA is the weighted sum of attention weights with w_{ij} the input element,

$$\tilde{x}_i = \sum_{j=1}^{n} w_{ij}.x_j. \tag{9}$$

Position-sensitive attention mechanism considers both the correlation between elements and the influence of position encoding, resulting in more accurate and position-aware attention weights.

Gated hierarchical long short-term memory network (GHLSTM)

The goal of the Hierarchical LSTM with gating is to further enhance the LSTM model's ability to model long-term dependencies and improve its performance in handling large-scale sequential data. It achieves this by introducing multiple levels of gating to gradually model dependencies at different time scales. The diagram of GHLSTM is shown in Fig. 3. The GHLSTM network is designed to model long-term dependencies across multiple time scales, thereby enhancing the accuracy of RUL prediction. This method consists of two hierarchical LSTM layers: a top-level LSTM for modeling global long-term dependencies and a bottom-level LSTM for capturing medium-term dependencies. The top LSTM processes the entire input sequence to capture long-term dependencies. The bottom LSTM processes half of the input sequence to capture medium-term dependencies. The outputs of the top and bottom LSTMs are concatenated to form the final output as a comprehensive temporal representation. The approach enables the model to adaptively focus on relevant features across different time scales, thereby improving the overall RUL prediction accuracy.

For the top hierarchical LSTM, the whole sequence of the input $\mathbf{x}_{1t}$ is input into the LSTM cellular, the formula is,

$$\mathbf{i}_{1t} = \sigma\left(\mathbf{w}_{1ix}\mathbf{x}_{1t} + \mathbf{w}_{1ih}\mathbf{h}_{1t-1} + \mathbf{b}_{1i}\right), \tag{10}$$

$$\mathbf{f}_{1t} = \sigma\left(\mathbf{w}_{1fx}\mathbf{x}_{1t} + \mathbf{w}_{1fh}\mathbf{h}_{1t-1} + \mathbf{b}_{1f}\right), \tag{11}$$

$$\mathbf{o}_{1t} = \sigma\left(\mathbf{w}_{1ox}\mathbf{x}_{1t} + \mathbf{w}_{1oh}\mathbf{h}_{1t-1} + \mathbf{b}_{1o}\right), \tag{12}$$

$$\bar{\mathbf{c}}_{1t} = \tanh\left(\mathbf{w}_{1cx}\mathbf{x}_{1t} + \mathbf{w}_{1ch}\mathbf{h}_{1t-1} + \mathbf{b}_{1c}\right), \tag{13}$$

$$\mathbf{c}_{1t} = \mathbf{f}_{1t} \odot \mathbf{c}_{1t-1} + \mathbf{i}_{1t} \odot \bar{\mathbf{c}}_{1t}, \tag{14}$$

$$\mathbf{h}_{1t} = \mathbf{o}_{1t} \odot \tanh\left(\mathbf{c}_{1t}\right), \tag{15}$$

For the bottom hierarchical LSTM, half of the whole sequence of the input $\mathbf{x}_{2t}$ is input into the LSTM cellular to extract the hidden feature in another time scale. Noted that the time scale can be deiced by the requirements. The formula of the bottom hierarchical LSTM cellular is,

$$\mathbf{i}_t^2 = \sigma\left(\mathbf{w}_{ix}^2\mathbf{x}_t^2 + \mathbf{w}_{ih}^2\mathbf{h}_{t-1}^2 + \mathbf{b}_i^2\right), \tag{16}$$

$$\mathbf{f}_{2t} = \sigma\left(\mathbf{w}_{2fx}\mathbf{x}_{2t} + \mathbf{w}_{2fh}\mathbf{h}_{2t-1} + \mathbf{b}_{2f}\right), \tag{17}$$

$$\mathbf{o}_{2t} = \sigma\left(\mathbf{w}_{2ox}\mathbf{x}_{2t} + \mathbf{w}_{2oh}\mathbf{h}_{2t-1} + \mathbf{b}_{2o}\right), \tag{18}$$

$$\bar{\mathbf{c}}_{2t} = \tanh\left(\mathbf{w}_{2cx}\mathbf{x}_{2t} + \mathbf{w}_{2ch}\mathbf{h}_{2t-1} + \mathbf{b}_{2c}\right), \tag{19}$$

$$\mathbf{c}_{2t} = \mathbf{f}_{2t} \odot \mathbf{c}_{2t-1} + \mathbf{i}_{2t} \odot \bar{\mathbf{c}}_{2t}, \tag{20}$$

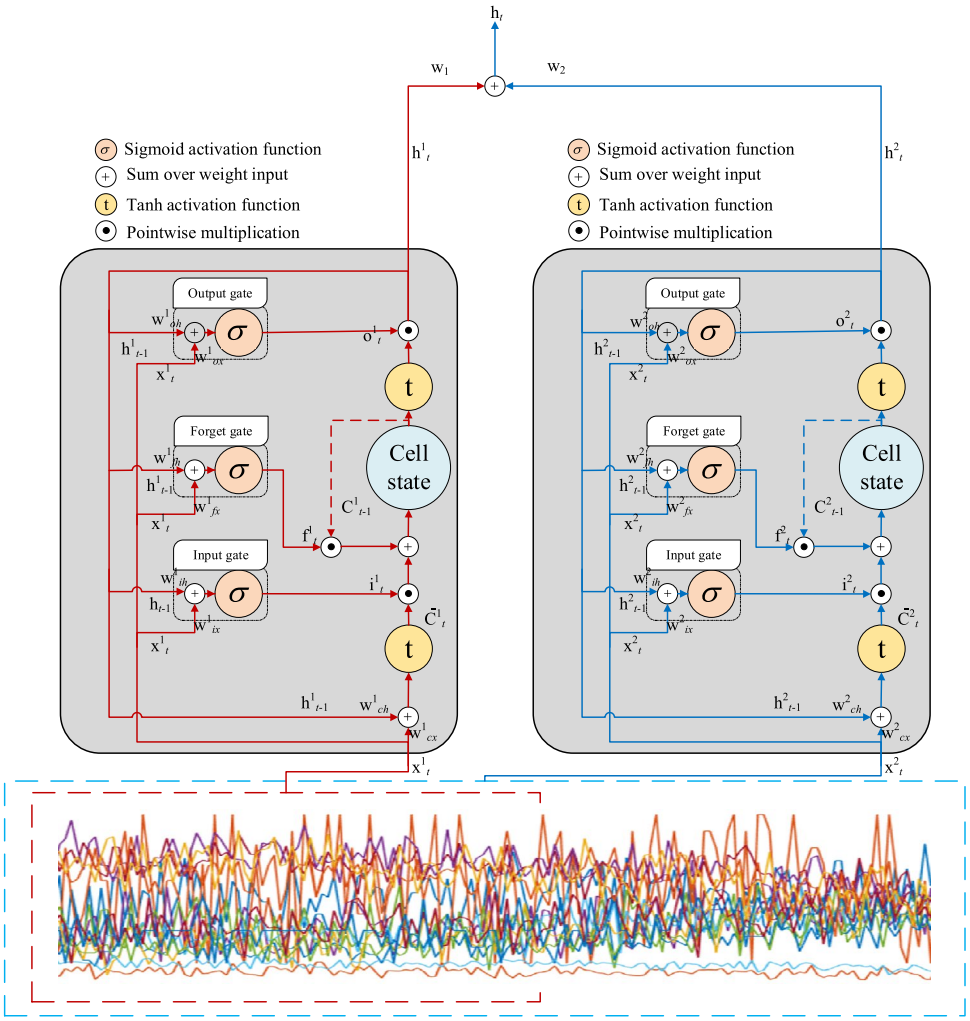

Figure 3. The structure of the proposed GHLSTM.

$$\mathbf{h}_{2t} = \mathbf{o}_{2t} \odot \tanh(\mathbf{c}_{2t}), \tag{21}$$

where $\mathbf{w}_{ix}$($\mathbf{w}_{fx}$, $\mathbf{w}_{ox}$ and $\mathbf{w}_{cx}$), and $\mathbf{w}_{ih}$($\mathbf{w}_{fh}$, $\mathbf{w}_{oh}$, and $\mathbf{w}_{ch}$) are the input and recurrent matrix weights, $\mathbf{b}_i$($\mathbf{b}_f$, $\mathbf{b}_o$, and $\mathbf{b}_c$) are the bias of the hidden layer,$\bar{\mathbf{c}}_t$ denotes the internal state of the cell, $\mathbf{c}_t$ denotes the memory cell state, σ represents the sigmoid function; tanh represents the tanh activation, and $\odot$ represents the pointwise multiplication. And the definitions of the two LSTM cellular are the same.

Then the output of the top hierarchical LSTM cellular $\mathbf{h}_{1t}$ and the bottom hierarchical LSTM cellular $\mathbf{h}_{2t}$ are combined to construct the final output of GHLSTM $\mathbf{h}_t$,

$$\mathbf{h}_t = \mathbf{w}_1\mathbf{h}_{1t} + \mathbf{w}_2\mathbf{h}_{2t}. \tag{22}$$

where $\mathbf{w}$ are the connected weights making the two outputs the same dimension.

Experimental analysis

Evaluation indexes

The widely used evaluation indexes for RUL prediction, i.e. score and root mean square error (RMSE), are adopted for the quantitated demonstration of the model performance. And the formulas of the indexes are given below,

$$A_i = \begin{cases} \exp(-((\overline{Rul_i} - Rul_i)/13)) - 1, & \overline{Rul_i} < Rul_i \\ \exp((\overline{Rul_i} - Rul_i)/10) - 1, & \overline{Rul_i} \geq Rul_i \end{cases} \tag{23}$$

$$Score = \sum_{i=1}^{N} A_i, \tag{24}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Rul_i - \overline{Rul_i}\right)^2}, \tag{25}$$

As shown in Fig. 4, it serves as a graphical representation of the trend of evaluation metrics. The curve's changing trend is easily discernible from the graph. When the error is positive, the Score value increases rapidly, indicating that the Score imposes a stronger penalty on lagged predictions. This characteristic aligns better with practical engineering requirements. Therefore, the Score is considered more reasonable compared to RMSE.

The description of the C-MAPSS dataset

To demonstrate the effectiveness and superiority of the proposed method in predicting the remaining useful life of aircraft engines, we utilized the C-MAPSS dataset provided by NASA[27], whose diagram is shown in Fig. 5. The dataset consists of a collection of aircraft engines, as shown in the figure. Furthermore, to showcase the capabilities of the proposed method under different operating conditions and fault modes, we used the simplest FD001 dataset and the most complex FD004 dataset as validation data.

FD001 dataset consists of 100 engines operating under a single operating condition and a single fault mode. The engines have varying lifespans, with the shortest operational cycle being 128 and the longest being 362. The dataset includes sensor measurements, such as fan speed, compressor speed, oil pressure, and various temperatures, along with operational settings like throttle setting and true airspeed. FD004 is a more complex dataset derived from the same aircraft engine simulations, containing 249 engines operating under 6 different operating conditions and experiencing 2 different fault modes. Similar to FD001, the engines have lifespans ranging from 128 to 543 operational cycles. The sensor measurements and operational settings are also similar to FD001, but the inclusion of multiple conditions and fault modes makes FD004 significantly more challenging for RUL prediction. The data details are presented in Table 3. The tasks of FD001 and FD004 remain the same, to accurately predict the RUL of each engine.

The preprocessing of input

Firstly, we delete the unimportant sensor measurements (sensors 1, 5, 6, 10, 16, 18, and 19), which are stable and have less degradation information. According to the literature[27], operating condition information is also helpful in RUL prediction. Thus the final input matrix consists of the remaining 14 sensor measurements and the three operating condition information. The second step, data segmentation is executed, as shown in Fig. 6 For the ith input with n dimension input and l sequence length (window size), the relevant RUL label is set as Ts–l–(i–1) × m, where m and T are the sliding steps and full-lifecycle value. Through greedy search by the experiments, the hyper-parameters l and m are set to 30 and 1. The last step is the linear piecewise RUL preprocessing for the RUL label $Rul_{\max} = 125$ as below,

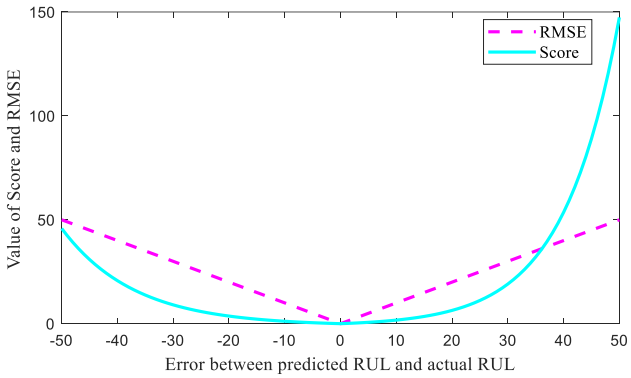

Figure 4. The curves of the two evaluation indexes.

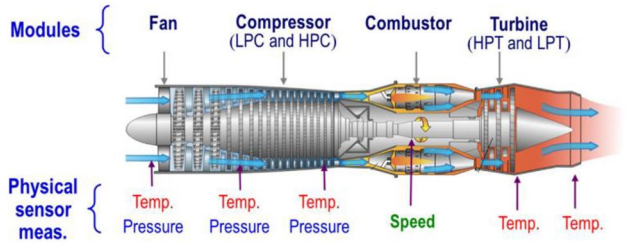

Figure 5. Diagram of the aircraft engine[27].

Subset	FD001	FD004
Total number of engines	100	249
Operating condition	1	6
Type of fault	1	2
Maximum cycles	362	543
Minimum cycles	128	128

Table 3. The details of dataset C-MAPSS.

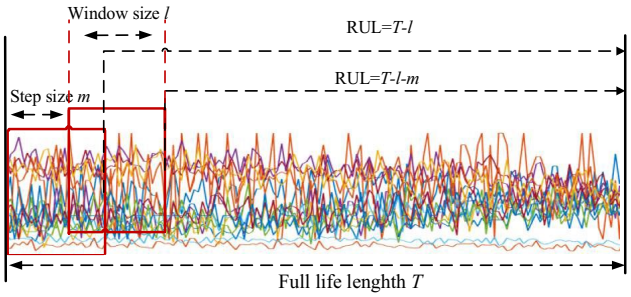

Figure 6. Processing of data segmentation.

$$Rul = \begin{cases} Rul, \text{ if } Rul \leq Rul_{\max} \\ Rul_{\max}, \text{ if } Rul > Rul_{\max} \end{cases} \tag{26}$$

The analysis and comparison of RUL prediction results

RUL prediction performance of the proposed method

The predicted results of the proposed model on the FD001 and FD004 subsets are shown in the figures below. In Figs. 7, 8, the value of the x-axis denotes the test engine number of the subset, while the y-axis represents the remaining useful life values (in cycles). The predicted remaining useful life and the actual remaining useful life of the test engines are represented by the red solid line and the purple dashed line, respectively. Overall, the predicted remaining useful life values of the test set engines in both subsets roughly align with the actual values, indicating the effectiveness of the proposed method in predicting the remaining useful life in these two subsets. Additionally, the error between the predicted life and the actual life in Fig. 7 is smaller than in Fig. 8. This

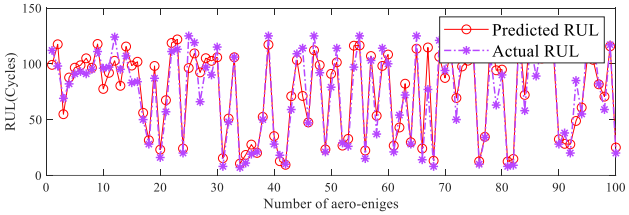

Figure 7. RUL prediction performance on FD001.

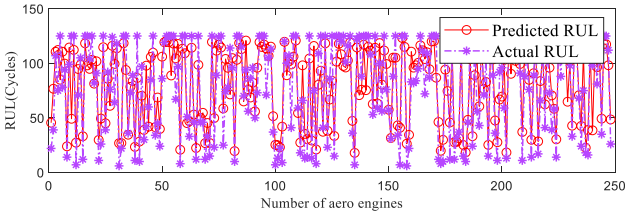

Figure 8. RUL prediction performance on FD004.

indicates that the proposed model performs better on the FD001 subset compared to the FD004 dataset. This is because the degradation trend of the aerospace engine under a single operating condition is relatively simpler. Moreover, there is a significant overlap between the degradation trends of the training set and the test set of aerospace engines. Therefore, the proposed method achieves higher accuracy in predicting the remaining useful life of aerospace engines under a single operating condition and a single fault compared to complex operating conditions and compound faults.

The proposed model's predictions on the complete degradation process are shown in Figs. 9a–d and 10a–d for four randomly selected aerospace engine engines from each subset The predicted RUL (PR) and actual RUL (AR) are represented by the blue line and red line, the absolute error (AE), calculated by based on PR and AR at each time instant, is denoted by the green bar. Thus the average error is represented by the mean of all AE values (MAE). The overall remaining useful life prediction results for FD001 are significantly better than for FD004, as indicated by the average MAE value. As the number of cycles increases, the degradation trend of aerospace engine engines becomes apparent. The proposed model exhibits higher accuracy in predicting the remaining useful life of most aerospace engine engines in the later stages of degradation compared to the earlier stages, as shown in Figs. 9a, c, d and 10a–d.

Ablation experiments

To validate the superiority of the proposed method, namely the effectiveness of PSA and GHLSTM, a series of erosion experiments were conducted. Assuming model m1 represents the proposed enhanced Transformer model, model m2 uses the same model architecture except for the attention part, which adopts the traditional multi-head self-attention module. Similarly, model m3 employs the same deep learning module except for the regression module, which uses the traditional LSTM model. Model m4 also uses the same deep learning module, but both its regression module and self-attention module adopt traditional models. All models are fine-tuned

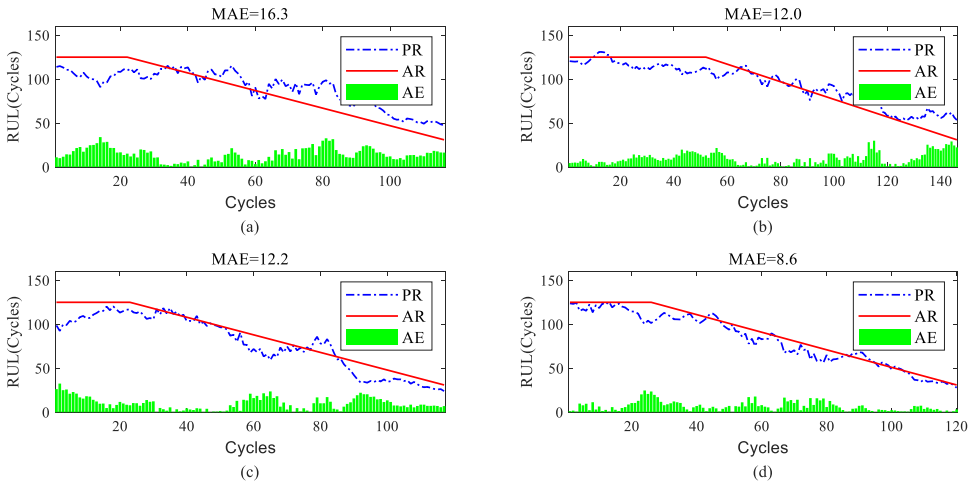

Figure 9. RUL prediction performance of engines of FD001 ((**a**) # 46, (**b**) # 58, (**c**) # 66, and (**d**) # 92).

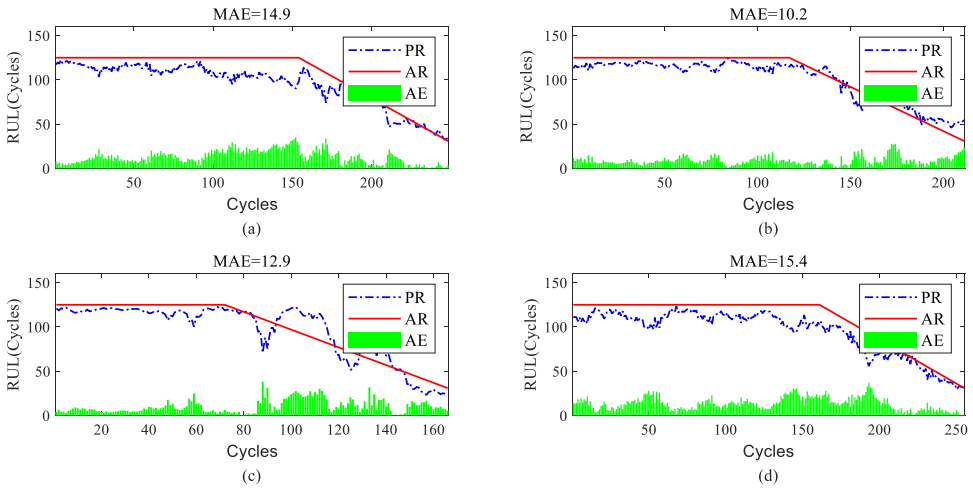

Figure 10. RUL prediction performance of engines of FD004 ((**a**) # 35, (**b**) # 68, (**c**) # 100, and (**d**) # 151).

and ten parallel experiments are conducted. The mean values and standard deviation (std) value of RMSE and score across all experiments are calculated, as shown in Table 4. The mean value is taken as the final predicted RUL, while std is used to quantify the robustness of the RUL prediction. It is evident from the table that the proposed model exhibits the lowest metric values and demonstrates the best predictive performance compared to other models. The std value is significantly lower than the mean value. Moreover, the predictive performance of m2 and m3 is superior to that of m4. These observations indicate that the proposed techniques contribute to the improvement of the accuracy in predicting the RUL.

Compared with state-of-arts

To further highlight the advantages of the proposed enhanced Transformer in predicting remaining useful life, a comparative experiment was conducted between the proposed model and several state-of-the-art methods[42–49]. To provide a comprehensive evaluation, the training set and testing set are fixed as the same for all compared models, each model was fine-tuned with the optimization goal of maximizing the accuracy in predicting remaining useful life, and 10 parallel experiments were conducted on FD001 and FD004 subsets. Subsequently, the scores and RMSE values based on the prediction results of all the aforementioned methods are listed in Table 5. From the table, it can be observed that all methods perform best in the FD001 subset and worst in the FD004 subset. This is because FD001 has the simple operating condition and fault type, while FD004 is the most complex subset with a larger number of tested engines.

On the FD001 dataset, which contains a single operating condition and fault mode, the proposed model achieved a Score of 220 ± 23 and an RMSE of 13.14 ± 0.21. This performance showed an improvement of 4% in Score compared to the best-performing existing method (acyclic graph network), which obtained a Score of 229 and an RMSE of 11.96. While the RMSE value of the proposed method is lower than acyclic graph network, the evaluation index Score is more in line with the actual engine and the Score value is lower than Acyclic Graph Network. This means that the comprehensive performance of the proposed method is best compared with other models.

Furthermore, on the more complex FD004 dataset, which encompasses multiple operating conditions and faults, the proposed model achieved a Score of 1420 ± 125 and an RMSE of 14.25 ± 0.25. This performance demonstrated an improvement of 10% in Score and 6% in RMSE compared to the best-performing existing method (SIGRNNDWI), which obtained a Score of 1568 and an RMSE of 15.12. Overall, the proposed model exhibited improved RUL prediction accuracy on both datasets, particularly on the more complex FD004 subset. These results validate the effectiveness of the proposed PSA and GHLSTM techniques in enhancing RUL prediction for aircraft engines.

Model	FD001		FD004	
	Score	RMSE	Score	RMSE
m4 (mean ± std)	301 ± 33	14.58 ± 0.41	2310 ± 199	16.35 ± 0.63
m3 (mean ± std)	265 ± 28	13.45 ± 0.35	1765 ± 165	15.89 ± 0.54
m2 (mean ± std)	244 ± 23	13.65 ± 0.28	1580 ± 136	15.76 ± 0.42
m1 (mean ± std)	220 ± 23	13.14 ± 0.21	1420 ± 125	14.25 ± 0.25

Table 4. The RUL prediction comparisons of different methods on subset FD001and FD002.

Model	FD001		FD004	
	Score	RMSE	Score	RMSE
MONBNE[42]	334	15.04	6558	28.66
LSTM + attention + handscraft feature[43]	322	14.53	5649	27.08
Acyclic graph network[44]	229	11.96	3370	22.43
AEQRNN[45]	N/A	N/A	4597	20.60
MCLSTM-based[46]	260	13.21	2926	22.10
SMDN[47]	240	13.72	1591	18.24
SIGRNNDWI[48]	229	13.14	1568	15.12
MSBLS[49]	–	–	1785	17.75
Proposed (mean ± Std)	220 ± 23	13.14 ± 0.21	1420 ± 125	14.25 ± 0.25
Improvement	4%	–	10%	6%

Table 5. The RUL prediction comparisons of different methods on subset FD001and FD002.

Conclusion

For accurately predicting the RUL of aero-engines, this article proposed a novel enhanced transformer-based DL method with the PSA mechanism and GHLSTM method. The main contributions of the article are as follows. One is the proposed PSA mechanism, PSA can solve the problem of the traditional attention mechanism that the extracted high-level features are insensitive to their local context at each time step. Another is the development of GHLSTM, GHLSTM can learn the hidden features at different time scales, which helps to improve RUL. The effect of the proposed technologies has been validated by the ablation experiments. Through the quantitative evaluation of common indicators, the proposed method has an average improvement of 7% in Score and 11% in RMSE compared with other methods on the RUL prediction tasks of FD001 and FD004.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 26 July 2023; Accepted: 8 April 2024
Published online: 02 May 2024

References

1. Lei, Y. *et al.* Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process.* **104**, 799–834 (2018).
2. Gebraeel, N., Lei, Y., Li, N., Si, X. & Zio, E. Prognostics and Remaining Useful Life Prediction of Machinery: Advances, Opportunities and Challenges. *Journal of Dynamics, Monitoring and Diagnostics* **2**(1), 1–12 (2023).
3. Zhang, J. *et al.* A variational local weighted deep sub-domain adaptation network for remaining useful life prediction facing cross-domain condition. *Reliab. Eng. Syst. Saf.* **231**, 108986 (2023).
4. Ma, F., Zhang, H., Gong, Q. & Hon, K. K. B. A novel energy efficiency grade evaluation approach for machining systems based on inherent energy efficiency. *Int. J. Prod. Res.* **59**, 6022–6033 (2021).
5. Shu, H., Zou, C., Chen, J. & Wang, S. Research on micro/nano surface flatness evaluation method based on improved particle swarm optimization algorithm. *Front. Bioeng. Biotechnol.* **9**, 775455 (2021).
6. Yu, T., Chen, W., Junfeng, G. & Poxi, H. Intelligent detection method of forgings defects detection based on improved EfficientNet and memetic algorithm. *IEEE Access* **10**, 79553–79563 (2022).
7. Duan, L. *et al.* State of charge estimation of lithium-ion batteries based on second-order adaptive extended Kalman filter with correspondence analysis. *Energy* **280**, 128159 (2023).
8. Al-Greer, M. & Bashir, I. Physics-based model informed smooth particle filter for remaining useful life prediction of lithium-ion battery. *Measurement* **214**, 112838 (2023).
9. Fordal, J. M. *et al.* Application of sensor data based predictive maintenance and artificial neural networks to enable Industry 4.0. *Adv. Manuf.* **11**, 248–263 (2023).
10. Yousaf, M. Z., Khalid, S., Tahir, M. F., Tzes, A. & Raza, A. A novel dc fault protection scheme based on intelligent network for meshed dc grids. *Int. J. Electr. Power Energy Syst.* **154**, 109423 (2023).
11. Shi, Y., Mao, Y., Xu, X. & Xia, J. Machine learning-assisted dual fiber Bragg grating-based flexible direction sensing. *IEEE Sens. J.* **23**, 25572–25578 (2023).
12. Dong, S. *et al.* Deep transfer learning based on Bi-LSTM and attention for remaining useful life prediction of rolling bearing. *Reliab. Eng. Syst. Saf.* **230**, 108914 (2023).
13. Liu, X. *et al.* A hybrid multi-stage methodology for remaining useful life prediction of control system: Subsea Christmas tree as a case study. *Expert Syst. Appl.* **215**, 119335 (2023).
14. Jiang, B. *et al.* A holistic feature selection method for enhanced short-term load forecasting of power system. *IEEE Trans. Instrum. Meas.* **72**, 1–11 (2022).
15. Liu, Q., Jia, M., Gao, Z., Xu, L. & Liu, Y. Correntropy long short term memory soft sensor for quality prediction in industrial polyethylene process. *Chemometr. Intell. Lab. Syst.* **231**, 104678 (2022).
16. Xiang, S., Li, P., Luo, J. & Qin, Y. Micro transfer learning mechanism for cross-domain equipment RUL prediction. *IEEE Trans. Autom. Sci. Eng.* https://doi.org/10.1109/TASE.2024.3366288 (2024).
17. Wu, Q., Zhou, X. & Pan, X. Cutting tool wear monitoring in milling processes by integrating deep residual convolution network and gated recurrent unit with an attention mechanism. *Proc. Inst. Mech. Eng. B J. Eng. Manuf.* **237**, 1171–1181 (2023).
18. Zhang, Y., Hutchinson, P., Lieven, N. A. & Nunez-Yanez, J. Remaining useful life estimation using long short-term memory neural networks and deep fusion. *IEEE Access* **8**, 19033–19045 (2020).
19. Miao, H., Li, B., Sun, C. & Liu, J. Joint learning of degradation assessment and RUL prediction for aeroengines via dual-task deep LSTM networks. *IEEE Trans. Ind. Inform.* **15**(9), 5023–5032 (2019).
20. Liu, J., Lei, F., Pan, C., Hu, D. & Zuo, H. Prediction of remaining useful life of multi-stage aero-engine based on clustering and LSTM fusion. *Reliab. Eng. Syst. Saf.* **214**, 107807 (2021).
21. Zhang, Y., Xin, Y., Liu, Z.-W., Chi, M. & Ma, G. Health status assessment and remaining useful life prediction of aero-engine based on BiGRU and MMoE. *Reliab. Eng. Syst. Saf.* **220**, 108263 (2022).
22. Ma, M. & Mao, Z. Deep wavelet sequence-based gated recurrent units for the prognosis of rotating machinery. *Struct. Health Monit.* **20**(4), 1794–1804 (2021).
23. Xiao, L., Duan, F., Tang, J. & Abbott, D. A noise-boosted remaining useful life prediction method for rotating machines under different conditions. *IEEE Trans. Instrum. Meas.* **70**, 1–12 (2021).
24. Song, Y., Shi, G., Chen, L., Huang, X. & Xia, T. Remaining useful life prediction of turbofan engine using hybrid model based on autoencoder and bidirectional long short-term memory. *J. Shanghai Jiaotong Univ. (Sci.)* **23**, 85–94 (2018).
25. Gu, J. *et al.* Recent advances in convolutional neural networks. *Pattern Recognit.* **77**, 354–377 (2018).
26. Zhu, J., Chen, N. & Peng, W. Estimation of bearing remaining useful life based on multiscale convolutional neural network. *IEEE Trans. Ind. Electron.* **66**(4), 3208–3216 (2018).
27. Li, X., Ding, Q. & Sun, J.-Q. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliab. Eng. Syst. Saf.* **172**, 1–11 (2018).
28. Yang, B., Liu, R. & Zio, E. Remaining useful life prediction based on a double-convolutional neural network architecture. *IEEE Trans. Ind. Electron.* **66**(12), 9521–9530 (2019).
29. Jiang, J.-R., Lee, J.-E. & Zeng, Y.-M. Time series multiple channel convolutional neural network with attention-based long short-term memory for predicting bearing remaining useful life. *Sensors* **20**(1), 166 (2019).
30. Vaswani, A. *et al.* Attention is all you need. *Adv. Neur. Inp. Syst.* **30**, 1–11 (2017).

31. Jin, X.-B. *et al.* End-to-end GPS tracker based on switchable fuzzy normalization codec for assistive drone application. *IEEE Trans. Consum. Electron.* **8**, 33 (2023).
32. Kong, J. *et al.* ADCT-Net: Adaptive traffic forecasting neural network via dual-graphic cross-fused transformer. *Inf. Fusion* **103**, 102122 (2024).
33. Tetko, I. V., Karpov, P., Van Deursen, R. & Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Commun.* **11**(1), 5575 (2020).
34. Naseer, M. M. *et al.* Intriguing properties of vision transformers. *Adv. Neur. Inp. Syst.* **34**, 23296–23308 (2021).
35. Zhou, H.-Y. *et al.* A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* **7**, 1–13 (2023).
36. Zhang, Z., Song, W. & Li, Q. Dual-aspect self-attention based on transformer for remaining useful life prediction. *IEEE ASME Trans. Mechatron.* **71**, 1–11 (2022).
37. Su, X., Liu, H., Tao, L., Lu, C. & Suo, M. An end-to-end framework for remaining useful life prediction of rolling bearing based on feature pre-extraction mechanism and deep adaptive transformer model. *Comput. InE Comp.* **161**, 107531 (2021).
38. Chadha, G. S., Shah, S. R. B., Schwung, A. & Ding, S. X. Shared temporal attention transformer for remaining useful lifetime estimation. *IEEE Access* **10**, 74244–74258 (2022).
39. Chang, Y., Li, F., Chen, J., Liu, Y. & Li, Z. Efficient temporal flow Transformer accompanied with multi-head probsparse self-attention mechanism for remaining useful life prognostics. *Reliab. Eng. Syst. Saf.* **226**, 108701 (2022).
40. Ren, L., Jia, Z., Wang, X., Dong, J. & Wang, W. A $ T^{2} $-tensor-aided multiscale transformer for remaining useful life prediction in IIoT. *IEEE Trans. Ind. Inform.* **18**(11), 8108–8118 (2022).
41. Ding, Y. & Jia, M. Convolutional transformer: An enhanced attention mechanism architecture for remaining useful life estimation of bearings. *IEEE ASME Trans. Mechatron.* **71**, 1–10 (2022).
42. Zhang, C., Lim, P., Qin, A. K. & Tan, K. C. Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics. *IEEE Trans. Neural Netw. Learn. Syst.* **28**(10), 2306–2318 (2016).
43. Zhao, R. *et al.* Machine health monitoring using local feature-based gated recurrent unit networks. *IEEE Trans. Ind. Electron.* **65**(2), 1539–1548 (2017).
44. Li, J., Li, X. & He, D. A directed acyclic graph network combined with CNN and LSTM for remaining useful life prediction. *IEEE Access* **7**, 75464–75475 (2019).
45. Cheng, Y., Hu, K., Wu, J., Zhu, H. & Shao, X. Autoencoder quasi-recurrent neural networks for remaining useful life prediction of engineering systems. *IEEE ASME Trans. Mechatron.* **27**(2), 1081–1092 (2021).
46. Xiang, S., Qin, Y., Luo, J., Pu, H. & Tang, B. Multicellular LSTM-based deep learning model for aero-engine remaining useful life prediction. *Reliab. Eng. Syst. Saf.* **216**, 107927 (2021).
47. Xiang, S., Qin, Y., Luo, J. & Pu, H. Spatiotemporally multidifferential processing deep neural network and its application to equipment remaining useful life prediction. *IEEE Trans. Ind. Inform.* **18**(10), 7230–7239 (2021).
48. Xiang, S., Li, P., Huang, Y., Luo, J. & Qin, Y. Single gated RNN with differential weighted information storage mechanism and its application to machine RUL prediction. *Reliab. Eng. Syst. Saf.* **242**, 109741 (2024).
49. Xu, T., Han, G., Zhu, H., Lin, C. & Peng, J. Multiscale BLS-based lightweight prediction model for remaining useful life of aero-engine. *IEEE Trans. Reliab.* https://doi.org/10.1109/TR.2023.3349201 (2024).

Acknowledgements

This paper was supported by Chongqing Technical Innovation and Application Development Special General Project (cstc2019jscx-msxmX0168, cstc2019jscx-msxmX0312, cstc2019jscx-msxmX0008, cstc2020jscx-msxmX0119), and partially supported by school level research projects (120777), supported by the experimental conditions of Chongqing University of Posts and Telecommunications and Chongqing University.

Author contributions

Chen wrote the main manuscript text.

Competing interests

The author declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.C.

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

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