scientific reports

Check for updates

OPEN

# A novel gear RUL prediction method by diffusion model generation health index and attention guided multi-hierarchy LSTM

Xinping Chen

**Gears, as indispensable components of machinery, demand accurate prediction of their Remaining Useful Life (RUL). To enhance the utilization of ordered information within time series data and elevate RUL prediction precision, this study introduces the attention-guided multi-hierarchy LSTM (AGMLSTM). This innovative approach leverages attention mechanisms to capture the intricate interplay between high and low hierarchical features of the input data, marking the first application of such a technique in gear RUL prediction. Additionally, a refined health indicator (HI) is introduced, constructed through a diffusion model, to precisely reflect the gears' health condition. The proposed RUL prediction method unfolds as follows: firstly, HIs are computed from gear vibration data. Subsequently, leveraging the known HIs, AGMLSTM predicts future HIs, and the RUL of the gear is determined upon surpassing the failure threshold. Quantitative analysis of experimental results conclusively demonstrates the superiority of the proposed RUL prediction method over existing approaches for gear RUL estimation.**

Recently, with the rapid development of Industry 4.0, engineering equipment has become increasingly complex and intelligent. In practice, the reliability and stability of equipment operation is an important prerequisite for completing the preset tasks, therefore extremely rigorous requirements are put forward. Given that the gear is the critical and costly drivetrain component in wind turbines and aero-engines, whose failure makes downtime, operations, and maintenance (OM) costs, and, to some extent, casualties[1,2]. Thus researchers pay more attention to the approaches to remaining useful life (RUL) prediction[3–5]. RUL refers to the expected continuous normal working time of the gear from the present to the occurrence of potential failures[5]. The RUL prediction, as an important role in prognostics and health management (PHM), enables the predicted OM decision assistance, which helps ensure equipment stability and avoid damage.

After continuous exploration and verification, the prediction of RUL has yielded significant theoretical research results in the academic world and holds vast potential for application in the industrial sector. RUL prediction approaches are broadly categorized into three classes: method-based[6,7], data-driven[3,4], and hybrid[8,9]. These methods exhibit distinct characteristics, but with the rapid advancement of technologies like artificial intelligence[5], sensor technology[10], and signal processing technology[11], the data-driven approach has emerged as the mainstream method for RUL prediction, particularly in complex engineering equipment. Besides data-driven method is more convenient than the model-based method and hybrid method which require a certain expert knowledge of failure mechanism. These characteristics enable data-driven methods suitable for RUL prediction and become a research hot.

Deep learning (DL), being the most popular method in the data-driven approach, has shown remarkable success in machine PHM[12]. Recently, numerous RUL prediction methods based on DL have been proposed by scholars. Ren et al.[3] introduced a simple DL method for machine RUL prediction, incorporating features in the time domain and frequency domain into a fully connection NN. Meanwhile, Yang et al.[4] developed a novel DL method with the first convolution neural network (CNN) as the detector of the initial failure point of rotating machinery, and the second CNN as the RUL predictor. Cheng et al.[5] first designed the CNN to construct the

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

health indicator from the raw data pre-processed by the Hilbert-Huang Transform (HTT), then estimated the machine RUL by SVR regression.

Long short-term memory (LSTM)[13] as the famous recurrent neural network (RNN) variant not only has its recursive properties but also has unique gating mechanisms, which makes it very suitable for processing sequential data compared with other neural networks (NNs). Therefore, LSTM as a RUL predictor is becoming more and more popular in PHM field. Yang et al.[14] executed mounts of experiments to find the advantage of the operation information data in the improvement of RUL prediction by using the LSTM models. Wu et al.[15] proposed a deep LSTM model to estimate bearing RUL via multiple sensor signals. Yuan et al.[16] investigated the prognostic performance of several RNNs for RUL estimation of aero-engines, including normal RNN, gated recurrent unit (GRU), and LSTM. Wang et al.[17] presented a novel RUL prediction approach. Firstly, the bearing degradation curve was classified into multiple stages, and then the RUL was obtained by multi-step prediction according to the stage. For the joint tasks of fault assessment and RUL estimation, Miao et al.[18] designed and established a dual-learning LSTM model. Chen al.[19] adopted an attention mechanism to weigh the data of different time steps in the cellular to improve the predictive ability of the improved LSTM. Qin al.[20] proposed a novel attention mechanism to screen the important information before and after inputting the hidden layer of GRU and further improve the roll bearing's RUL prediction accuracy. The above methods improve the LSTM from multiple angles, e.g. combined with CNN feature extraction, artificial feature construction, attention mechanism selection, and other techniques, to obtain better prediction performance. On the other side, researchers find that there is another feature named ordered information hidden in the sequence information which is helpful for RUL estimation, and LSTM based on ordered information (On-LSTM) is firstly proposed to deal with the feature and applied on gear RUL prediction in literature[21,22]. Based on the angle, literatures[23,24] further explored the usage of ordered information on RUL pre-diction tasks by using attention-guided and mining the mixed zone of hierarchies. The studies about gear RUL prediction have been developed and applied. However, there still exist two main gaps in the methods of gear RUL prediction.

1. One is that the prediction method can not mine ordered information of HIs fully and reasonably, which can decrease the feature extraction ability of models and impact the RUL prediction accuracy.
2. Another is that there is rare work on the construction of HI with clear degradation trends and stable failure theories.

Facing the challenge, the article proposed a novel attention-guided multi-hierarchy LSTM (AGMLSTM) model. AGMLSTM not only can mine the feature of mixed hierarchy but also has the ability which is guided by the attention mechanism reasonably. Thus AGMLSTM is more suitable for gear RUL prediction. Besides, a suitable health index (HI) is beneficial for RUL prediction accuracy. In the paper, a novel HI which is smooth and has a clear trend constructed by the diffusion model is presented. Finally, based on the known HIs, the AGMLSTM is used to predict the future HIs step by step until it exceeds the preset failure value, and the RUL of gear is finally obtained. The outperformance of the presented RUL approach is illustrated by the quantitative evaluation of various indexes during the experiments. Particularly noteworthy is the remarkable achievement of 92% RUL prediction accuracy in the challenging task of predicting gear RUL within one hour, signifying the practical significance of our approach in online RUL prediction.

The main contributions in the article are as follows:

1. The adoption of the diffusion model represents a pioneering approach to constructing the HIs for gears, effectively mitigating fluctuations. Gear HI curves exhibit declining trends, and their failure thresholds are similar.
2. AGMLSTM is proposed for gear RUL prediction. This method demonstrates enhanced capability in extracting ordered information, improving feature extraction, and boosting RUL estimation.
3. Building on the diffusion model and AGMLSTM, the study proposes a novel prediction method, validated through comprehensive assessments of full-life vibration data for gears."

The remainder of the article is arranged as follows. "Theoretical basis" not only introduces the concept of diffusion model but also introduces LSTM. The details of the proposed methods are described in "The proposed methodology". The experiments with results analysis are given in "Experimental analysis". Last, in "Conclusion", the conclusion is summarized.

## Theoretical basis

### Diffusion model

Diffusion model[25] is a novel advanced deep generative model. It gradually transforms data into noise and then learns the de-noising process to generate new samples in both forward and backward directions. Thus The learned de-noising module of diffusion model is adopted to construct gear HIs. Figure 1 illustrates the intuition behind the Diffusion model.

In this study, the de-noising diffusion probabilistic model is employed, which operates through the utilization of two Markov chains. Diffusion Model adopts a progressive nosing and de-noising approach. In the forward process, Gaussian noise is gradually added to the original data layer by layer until it transforms into a simple prior Gaussian distribution. In the reverse process, the noise is gradually eliminated by the deep neural network. The fixed approximate posterior $q(x_{1:N}|x_0)$ in the forward stage is calculated in Eqs. (1) and (2),

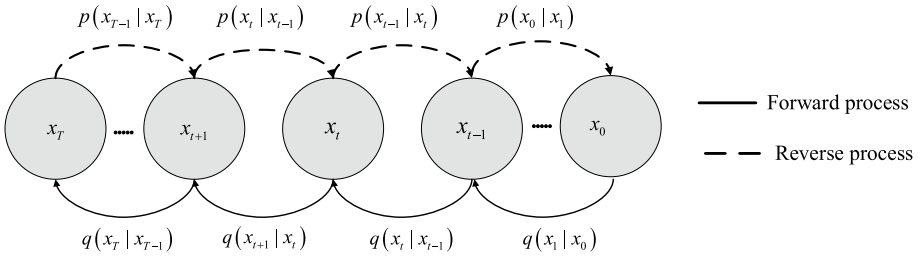


**Figure 1.** The details of diffusion model.

$$q(x_{1:N}|x_0) = \prod_{n=1}^{N} q(x_n|x_{n-1}) \tag{1}$$

$$q(x_n|x_{n-1}) \sim N\left(x_n;\ \sqrt{1-\beta_n}x_{n-1}, \beta_n I\right) \tag{2}$$

where $\beta_n \in (0,1)$, $N$ and $I$ are the added Gaussian noise, sample number, and identity matrix. While at the reverse process, a learnable Gaussian transition which is beginning at $p(x_n)$, with another Markov chain constructs the joint distribution $p_\theta(x_{n-1}|x_n)$, as calculated in Eqs. (3) and (4),

$$p_\theta(x_{0:N}) = P(x_N)\prod_{n=1}^{N} p(x_{n-1}|x_n) \tag{3}$$

$$p_\theta(x_{n-1}|x_n) \sim N(x_{n-1}, \mu_\theta(x_n,n), \delta_\theta(x_n,n)I) \tag{4}$$

where mean $\mu_\theta$ and variance $\delta_\theta$ are obtained from a deep NN.

The objective of the reverse Markov chain, i.e., computing $p_\theta(x_{n-1}|x_n)$, is to remove the Gaussian noise introduced during the forward process. The de-nosing object is $p_\theta(x_{n-1}|x_n)$ for the reverse Markov chain. Supposed that $x_0$ is sampled from the noise $p(x_n)$, repeating the process from $p_\theta(x_{n-1}|x_n)$ until $n = 1$.

For accurate sampling, make the trained reverse Markov chain $p_\theta(x_{n-1}|x_n)$ close to the posterior distribution $q(x_{n-1}|x_n, x_0)$ of the forward process given $x_0$. And Kullback–Leibler (KL) divergence is chosen as the similarity evaluation metric, whose equations are defined as bellows,

$$\begin{aligned} &D_{KL}\left(q(x_{n-1}|x_n,x_0)||p_\theta(x_n-1|x_n)\right) \\ &= E_q\left[\frac{1}{2\sum\theta}||\tilde{\mu}_n(x_n,x_0) - \mu_\theta(x_n,n)||^2\right] + C \end{aligned} \tag{5}$$

In the equation, $C$ is a constant that is independent of θ and $\tilde{\mu}_n$ represents the average value of $q(x_{n-1}|x_n,x_0)$. And the simplified objection is calculated in Eq. (6) by adding the noise NN $\varepsilon_\theta$ with parameters θ,

$$E_{x_{0\sim q(x_0)},\varepsilon\sim N(0,I)} = \left[\lambda(n)||\varepsilon - \varepsilon_\theta\left(\sqrt{\overline{\alpha}_n}x_0 + \sqrt{1-\overline{\alpha}_n}\varepsilon, n\right)||^2\right] \tag{6}$$

where $\lambda(n)$ is the function of positive weight.

## Long short term memory

LSTM[13] is proposed for releasing the limitation by the nonlinear procession of the data based on the gate mechanism as sown in Fig. 2. The mathematical expression of LSTM is as follows:

$$\mathbf{i}_t = \sigma(\mathbf{w}_{ix}\mathbf{x}_t + \mathbf{w}_{ih}\mathbf{h}_{t-1} + \mathbf{b}_i) \tag{7}$$

$$\mathbf{f}_t = \sigma\left(\mathbf{w}_{fx}\mathbf{x}_t + \mathbf{w}_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f\right) \tag{8}$$

$$\mathbf{o}_t = \sigma(\mathbf{w}_{ox}\mathbf{x}_t + \mathbf{w}_{oh}\mathbf{h}_{t-1} + \mathbf{b}_o) \tag{9}$$

$$\bar{\mathbf{c}}_t = \tanh(\mathbf{w}_{cx}\mathbf{x}_t + \mathbf{w}_{ch}\mathbf{h}_{t-1} + \mathbf{b}_c) \tag{10}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t \tag{11}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{12}$$

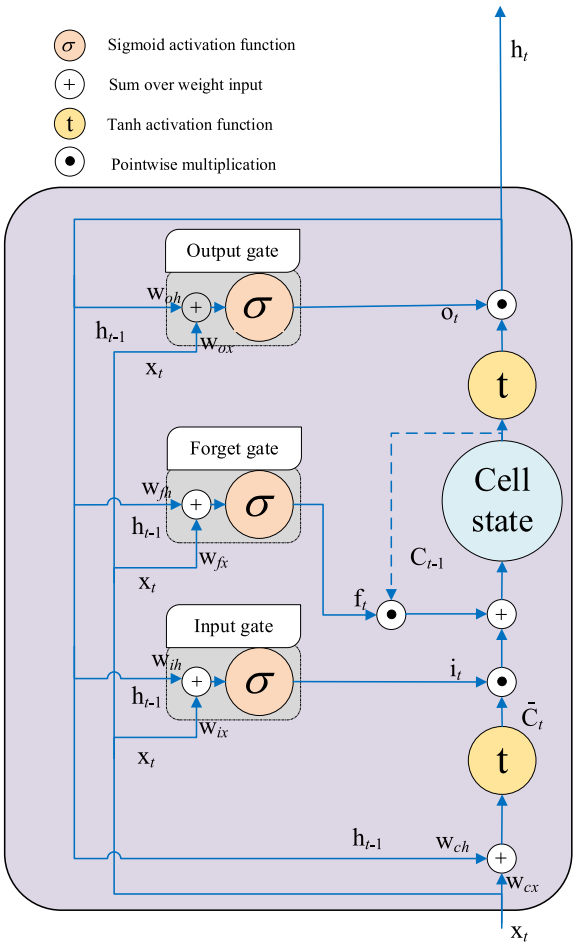


**Figure 2.** The diagram of LSTM neuron structure.

In Eqs. (7–12), the input weight matrixes $\mathbf{w}_{ix}$($\mathbf{w}_{fx}$, $\mathbf{w}_{ox}$, $\mathbf{w}_{cx}$) and the recurrent weight matrixes $\mathbf{w}_{ih}$($\mathbf{w}_{fh}$, $\mathbf{w}_{oh}$, $\mathbf{w}_{ch}$) are defined by the nonlinear transformation of $\mathbf{x}_t$ and $\mathbf{h}_{t-1}$ based on forget (input, output) gate, which decides the forget (input, output) degree of data in the hidden layer; $\mathbf{b}_i$($\mathbf{b}_f$, $\mathbf{b}_o$ and $\mathbf{b}_c$) are the bias of the hidden layer.$\bar{\mathbf{c}}_t$ and $\mathbf{c}_t$ are the internal state and memory state of the cell;$\odot$ denotes the pointwise multiplication. $\sigma$(tanh) is the sigmoid (tanh) activation function.

## The proposed methodology

### Attention-guided multi-hierarchy LSTM

ON-LSTM is first proposed in the NLP field to address the hierarchical structure problem, i.e. "characters, words, and phrases" has a different hierarchy and should be learned in different ways. However, for the vibration signal of mechanical equipment, the hierarchy of order information is difficult to give physical meaning. During the training process, ON-LSTM achieves automatic hierarchy by only providing feedback through the error between predicted and actual results, lacking effective guidance and clear physical interpretation in the hierarchical process. Moreover, the ordered information extracted by ON-LSTM exhibits mixed regions, and the features missed in mixed regions may impact the feature extraction capability. Therefore, this study proposes a new attention-guided multi-hierarchy Long Short-Term Memory (AGMLSTM) neural network that further partitions the mixed hierarchies using the attention mechanism, thereby forming an attention-guided multi-hierarchy information structure. The similarity between the elements of input vectors and recurrent vectors with attention labels determines the segmentation point between input hierarchies and historical hierarchies, which is the index of the most similar element with attention labels. This means that attention is to guide the hierarchical segmentation and give physical meaning to the hierarchy of ordered information of vibration data. Simultaneously, the multi-hierarchy partitioning enables neural networks to fully utilize ordered information. Information that is easily retained over a long period is assigned a high attention hierarchy, while information that is easily replaceable is assigned a low attention hierarchy. The mixed information, representing the intermediate attention hierarchy, is further divided into the sub-hierarchies of low intermediate attention, intermediate attention, and high intermediate attention, which respectively represent short-term information, mid-term information, and long-term information. It should be noted that the intermediate hierarchies (low intermediate, intermediate,

and high intermediate) will be zero when the high and low attention hierarchy information has no interaction. In this case, the information in the zone will not participate in the neural network's update process.

Let $\mathbf{x}_t = \left[ x_{t,1} \ x_{t,2} \ \ldots \ x_{t,n} \right]^T$ and $\mathbf{h}_{t-1} = \left[ h_{t-1,1} \ h_{t-1,2} \ \ldots \ h_{\mathrm{t-1,m}} \right]^T$ denote the input HIs at time step $t$ and the recurrent data at time step $t-1$, respectively. Compared to other networks, the main difference of AGMLSTM lies in the hierarchical information partitioning during the cell unit update process, as illustrated in Fig. 3. The proposed AGMLSTM utilizes attention-guided multi-hierarchy partitioning influenced by attention labels.

By calculating the similarity between input data, recurrent data, and the attention label, the maximum attention coefficient element is identified as the hierarchy segmentation point, so that the model identifies the hierarchy from the largest element to the element that is most similar to the label. Thus, the designed hierarchical structure can be combined with RNN through the attention hierarchies of information. By employing the designed update rules, information with a lower attention hierarchy is more prone to forgetting, while information with a higher attention hierarchy is preserved for a longer duration.

Due to the construction of multi-hierarchy information, let's assume that the main and auxiliary hierarchical positions of the input information $\mathbf{x}_t$ are denoted as $d^1_{t,i}$ and $d^2_{t,i}$, respectively, while the main and auxiliary hierarchical positions of the historical information $\mathbf{h}_{t-1}$ are denoted as $d^1_{t,f}$ and $d^2_{t,f}$. These positions are generated using the following construction functions: $F_1$, $F_2$, $F_3$, and $F_4$, guided by the query vector $q_m$. The auxiliary hierarchical positions are used to refine the interval of hierarchical mixing.

$$d^1_{t,i} = F_1(\mathbf{x}_t, \mathbf{h}_{t-1}, q_m) \tag{13}$$

$$d^2_{t,i} = F_2(\mathbf{x}_t, \mathbf{h}_{t-1}, q_m) \tag{14}$$

$$d^1_{t,f} = F_3(\mathbf{x}_t, \mathbf{h}_{t-1}, q_m) \tag{15}$$

$$d^2_{t,f} = F_4(\mathbf{x}_t, \mathbf{h}_{t-1}, q_m) \tag{16}$$

The memory cell state vector is updated according to certain rules based on the attention hierarchy of input information and recurrent information.

1) If $d^1_{t,f} \le d^1_{t,i}$, the main hierarchy of the input information $\mathbf{x}_t$ is higher than the main hierarchy of the historical information $\mathbf{h}_{t-1}$, resulting in an intermediate attention hierarchy. AGMLSTM is capable of further refining the intermediate attention hierarchy and dividing it into sub-hierarchies: low intermediate attention hierarchy, intermediate attention hierarchy, and high intermediate attention hierarchy, shown in Fig. 4. Therefore, when the hierarchical relationship simultaneously satisfies $d^2_{t,f} \le d^2_{t,i}$, the auxiliary hierarchy of the input information $\mathbf{x}_t$ is also higher than the auxiliary hierarchy of the historical information $\mathbf{h}_{t-1}$. There is an interactive space between $d^2_{t,f}$ and $d^2_{t,i}$. The cell unit update rules are as follows: within the cell unit interval $\left[0, d^1_{t,f}\right)$, the candidate memory cell state vector $\bar{\mathbf{c}}_t$ is directly input into the corresponding memory cell, while within the cell unit interval $\left[d^1_{t,i}, d_{max}\right]$, the memory cell state vector $\mathbf{c}_{t-1}$ from the previous time step is directly input into the corresponding memory cell. As for the overlapping region $\left[d^1_{t,f}, d^1_{t,i}\right)$, further refinement updates are performed based on the auxiliary hierarchical positions of the input and historical information. For the overlapping region $\left[d^1_{t,f}, d^2_{t,f}\right)$, the update of $\mathbf{c}_t$ is:

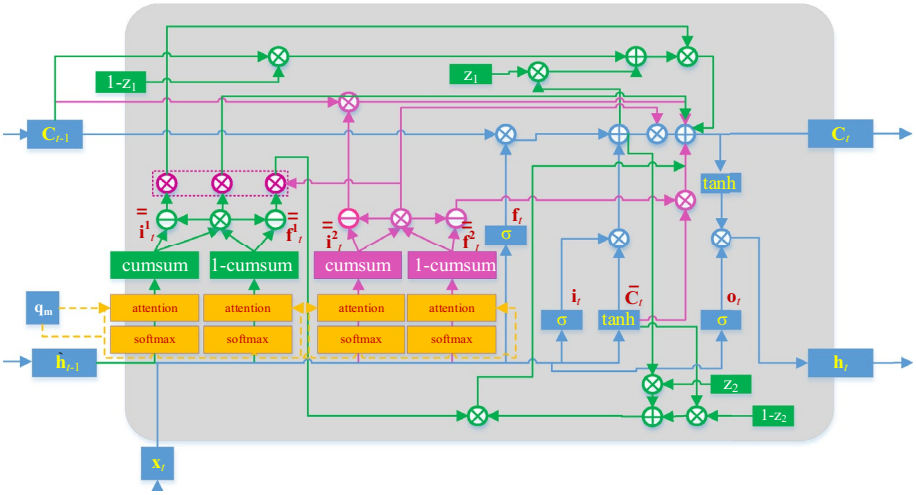


**Figure 3.** The cellular of AGMLSTM.

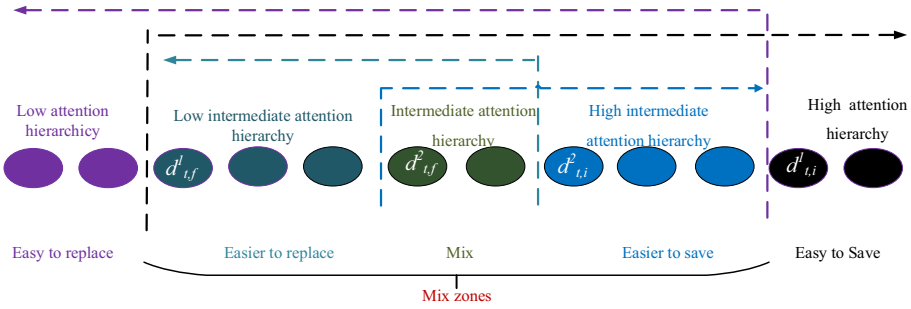


**Figure 4.** The hierarchy division of AGMLSTM when $d^1_{t,f} \leq d^2_{t,f} \leq d^2_{t,i} \leq d^1_{t,i}$.

$$\mathbf{c}_t = \mathbf{s}_1 \odot (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t) + (1 - \mathbf{s}_1) \odot \bar{\mathbf{c}}_t \tag{17}$$

where $1-\mathbf{s}_1$ is the scale of short-term information in the cellular memory at the case.

For the overlapping region $\left[d^2_{t,f}, d^2_{t,i}\right)$, the update rule of $\mathbf{c}_t$ is defined as follows:

$$\mathbf{c}_t = (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t) \tag{18}$$

For the overlapping region $\left[d^2_{t,i}, d^1_{t,i}\right)$,$\mathbf{c}_t$ is updated by,

$$\mathbf{c}_t = \mathbf{s}_2 \odot (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t) + (1 - \mathbf{s}_2) \odot \mathbf{c}_{t-1} \tag{19}$$

where $1-\mathbf{s}_2$ represents the long-term data ratio. Therefore, under this hierarchy distribution $\mathbf{c}_t$ is presented bellows,

$$\mathbf{c}_t = \begin{pmatrix} \mathbf{c}_{t-1} \left[d^1_{t,i}, d_{max}\right] \\ \mathbf{s}_2 \odot (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t) + (1 - \mathbf{s}_2) \odot \mathbf{c}_{t-1} \left[d^2_{t,i}, d^1_{t,i}\right) \\ (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t) \left[d^2_{t,f}, d^2_{t,i}\right) \\ \mathbf{s}_1 \odot (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t) + (1 - \mathbf{s}_1) \odot \bar{\mathbf{c}}_t \left[d^1_{t,f}, d^2_{t,f}\right) \\ \bar{\mathbf{c}}_t \left[0, d^1_{t,f}\right) \end{pmatrix} \tag{20}$$

When the hierarchical relationship simultaneously satisfies $d^2_{t,f} \geq d^2_{t,i}$, and the auxiliary hierarchical level of the input information $\mathbf{x}_t$ is lower than the auxiliary hierarchical level of the historical information $\mathbf{h}_{t-1}$, there is no interactive space between $d^2_{t,f}$ and $d^2_{t,i}$,, shown in Fig. 5. In this case, the update mechanisms within the index ranges $\left[d^1_{t,i}, d_{max}\right]$ and $\left[0, d^1_{t,f}\right)$ remain consistent with the first case. However, within the index range $\left[d^1_{t,f}, d^2_{t,i}\right)$, the update of $\mathbf{c}_t$ is as follows:

$$\mathbf{c}_t = \mathbf{s_1} \odot (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t) + (1 - \mathbf{s_1}) \odot \bar{\mathbf{c}}_t \tag{21}$$

where $1-\mathbf{s_1}$ is the short-term information ratio. For elements in the range $\left[d^2_{t,i}, d^2_{t,f}\right)$, the cell memory state $\mathbf{c}_t$ is zero, while in the range $\left[d^2_{t,f}, d^1_{t,i}\right)\mathbf{c}_t$ is calculated as below,

$$\mathbf{c}_t = \mathbf{s}_2 \odot (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t) + (1 - \mathbf{s}_2) \odot \mathbf{c}_{t-1} \tag{22}$$

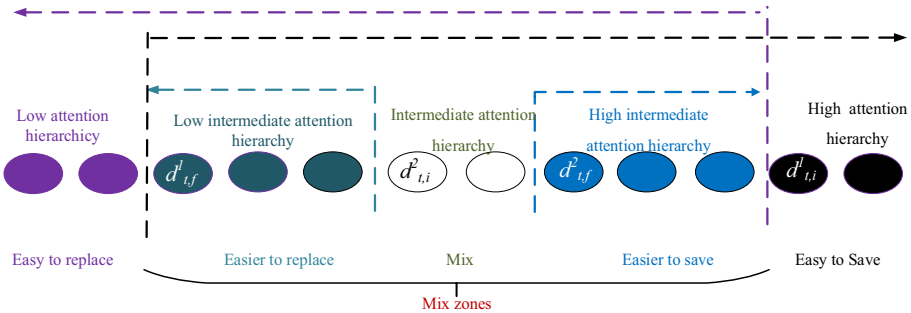


**Figure 5.** The hierarchy division of AGMLSTM when $d^1_{t,f} \leq d^2_{t,i} \leq d^2_{t,f} \leq d^1_{t,i}$.

where $1-\mathbf{s}_2$ denotes the long-term information ratio. In summary, $\mathbf{c}_t$ at the hierarchy is updated by the below rules,

$$\mathbf{c}_t = \begin{pmatrix} \mathbf{c}_{t-1}\left[d_{t,i}^1, d_{max}\right] \\ \mathbf{s}_2 \odot (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t) + (1-\mathbf{s}_2) \odot \mathbf{c}_{t-1}\left[d_{t,f}^2, d_{t,i}^1\right) \\ 0\left[d_{t,i}^2, d_{t,f}^2\right) \\ \mathbf{s}_1 \odot (\mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \bar{\mathbf{c}}_t) + (1-\mathbf{s}_1) \odot \bar{\mathbf{c}}_t\left[d_{t,f}^1, d_{t,i}^2\right) \\ \bar{\mathbf{c}}_t\left[0, d_{t,f}^1\right) \end{pmatrix} \tag{23}$$

2) If $d_{t,f}^1 \leq d_{t,i}^1$, the main hierarchical level of the input information $\mathbf{x}_t$ is higher than the main hierarchical level of the historical information $\mathbf{h}_{t-1}$, indicating that the attention focus on the input data than the recurrent data, there are no overlapping cell unit regions. Therefore, $\mathbf{c}_t$ within the intermediate attention level, there is no need for the mixing of short-term and mid-term memory $\mathbf{f}_t \odot \mathbf{c}_{t-1} + i_t \odot \bar{\mathbf{c}}_t$ to update. Within the cell unit interval $\left[d_{t,i}, d_{t,f}\right)$, the current time step's cell activation vector is set to zero. $\bar{\mathbf{c}}_t$ is the direct input within the cell unit interval $\left[0, d_{t,i}\right)$, and for $\mathbf{c}_{t-1}$ is the interval $\left[d_{t,f}, d_{max}\right]$. At the situation, $\mathbf{c}_t$ is updated by Eq. (24) with its hierarchical partition shown in Fig. 6.

$$\mathbf{c}_{t'} = \begin{pmatrix} \bar{\mathbf{c}}_t, < d_{t,i} \\ \mathbf{0}, \left[d_{t,i}, d_{t,f}\right) \\ \mathbf{c}_{t-1}, \geq d_{t,f} \end{pmatrix} \tag{24}$$

The construction functions $F_1$, $F_2$, $F_3$ and $F_4$ are derived as follows. We first normalize the input data $\mathbf{x}_t$ and historical data $\mathbf{h}_{t-1}$ using softmax function, introducing four $m$-dimensional vectors $\bar{\mathbf{f}}_t^1$, $\bar{\mathbf{i}}_t^1$, $\bar{\mathbf{f}}_t^2$ and $\bar{\mathbf{i}}_t^2$.

$$\bar{\mathbf{f}}_t^1 = \text{softmax}\left(\mathbf{w}_{f_1}\mathbf{x}_t + \mathbf{w}_{f_1}\mathbf{h}_{t-1} + \mathbf{b}_{f_1}\right) \tag{25}$$

$$\bar{\mathbf{i}}_t^1 = \text{softmax}\left(\mathbf{w}_{i_1}\mathbf{x}_t + \mathbf{w}_{i_1}\mathbf{h}_{t-1} + \mathbf{b}_{i_1}\right) \tag{26}$$

$$\bar{\mathbf{f}}_t^2 = \text{softmax}\left(\mathbf{w}_{f_2}\mathbf{x}_t + \mathbf{w}_{f_2}\mathbf{h}_{t-1} + \mathbf{b}_{f_2}\right) \tag{27}$$

$$\bar{\mathbf{i}}_t^2 = \text{softmax}\left(\mathbf{w}_{i_2}\mathbf{x}_t + \mathbf{w}_{i_2}\mathbf{h}_{t-1} + \mathbf{b}_{i_2}\right) \tag{28}$$

where $\mathbf{w}_f$ and $\mathbf{w}_i$ represent the weight matrices of the softmax layers for historical data and input data, respectively, while $\mathbf{b}_f$ and $\mathbf{b}_i$ represent the thresholds of the softmax layers for historical data and input data.

Next, the attention coefficients $\alpha_{t,i}^1$, $\alpha_{t,i}^2$, $\lambda_{t,i}^1$ and $\lambda_{t,i}^2$ for the input data, and recurrent data are calculated using Eqs. (29–32), respectively:

$$\alpha_{t,i}^1 = \frac{\exp\left(s\left(\bar{i}_{t,i}^1, q_{t,m}\right)\right)}{\sum_{j=1}^{m}\exp\left(s\left(\bar{i}_{t,j}^1, q_{t,m}\right)\right)} \tag{29}$$

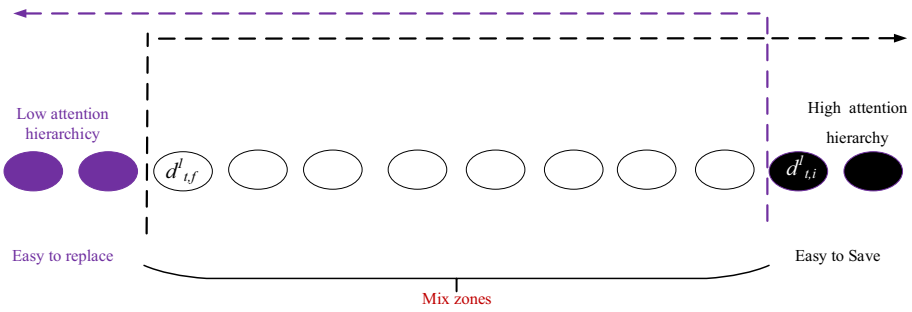


**Figure 6.** The hierarchy division of AGMLSTM when $d_{t,f}^1 \leq d_{t,i}^1$.

$$\lambda_{t,i}^{1} = \frac{\exp\left(s\left(\bar{f}_{t,i}^{1}, q_{t,m}\right)\right)}{\sum_{j=1}^{m} \exp\left(s\left(\bar{f}_{t,j}^{1}, q_{t,m}\right)\right)} \tag{30}$$

$$\alpha_{t,i}^{2} = \frac{\exp\left(s\left(\bar{i}_{t,i}^{2}, q_{t,m}\right)\right)}{\sum_{j=1}^{m} \exp\left(s\left(\bar{i}_{t,j}^{2}, q_{t,m}\right)\right)} \tag{31}$$

$$\lambda_{t,i}^{2} = \frac{\exp\left(s\left(\bar{f}_{t,i}^{2}, q_{t,m}\right)\right)}{\sum_{j=1}^{m} \exp\left(s\left(\bar{f}_{t,j}^{2}, q_{t,m}\right)\right)} \tag{32}$$

During the training process, the query vector $q_{t,m}$ at this time step $t$ is set as $x_{t+1,n}$, while during the inference process, it is set as $x_{t,n}$.

The four scoring functions $s\left(\bar{i}_{t,i}^{1}, q_{t,m}\right)$, $s\left(\bar{i}_{t,i}^{2}, q_{t,m}\right)$, $s\left(\bar{f}_{t,f}^{1}, q_{t,m}\right)$ and $s\left(\bar{f}_{t,f}^{2}, q_{t,m}\right)$ are defined as follows:

$$s\left(\bar{i}_{t,i}^{1}, q_{t,m}\right) = \frac{\bar{i}_{t,i}^{1\,T} q_{t,m}}{\sqrt{m}} \tag{33}$$

$$s\left(\bar{f}_{t,i}^{1}, q_{t,m}\right) = \frac{\bar{f}_{t,i}^{1\,T} q_{t,m}}{\sqrt{m}} \tag{34}$$

$$s\left(\bar{i}_{t,i}^{2}, q_{t,m}\right) = \frac{\bar{i}_{t,i}^{2\,T} q_{t,m}}{\sqrt{m}} \tag{35}$$

$$s\left(\bar{f}_{t,i}^{2}, q_{t,m}\right) = \frac{\bar{f}_{t,i}^{2T} q_{t,m}}{\sqrt{m}} \tag{36}$$

The maximum positions of the attention coefficients $d_{t,i}^{1}$ $d_{t,i}^{2}$ are set as the main and auxiliary hierarchical positions of the input information $\mathbf{x}_t$; and the maximum positions of the attention coefficients $d_{t,f}^{1}$ and $d_{t,f}^{2}$ are set as the main and auxiliary hierarchical positions of the historical information $\mathbf{h}_{t-1}$, respectively:

$$d_{t,i}^{1} = \text{index}[\max(\alpha_{t,i}^{1})] \tag{37}$$

$$d_{t,f}^{1} = \text{index}[\max(\lambda_{t,i}^{1})] \tag{38}$$

$$d_{t,i}^{2} = \text{index}[\max(\alpha_{t,i}^{2})] \tag{39}$$

$$d_{t,f}^{2} = \text{index}[\max(\lambda_{t,i}^{2})] \tag{40}$$

where index() denotes as the element position extraction function.

To achieve the automatic hierarchical update as described above, the cumulative sum function cumsum() is used to compute the cumulative sums of the attention coefficients, resulting in the main and auxiliary input gates $\bar{\bar{\mathbf{i}}}_t^{1}$ and $\bar{\bar{\mathbf{i}}}_t^{2}$, as well as the main and auxiliary forget gates $\bar{\bar{\mathbf{f}}}_t^{1}$ and $\bar{\bar{\mathbf{f}}}_t^{2}$, which can be written as follows:

$$\bar{\bar{\mathbf{i}}}_t^{1} = 1 - \text{cumsum}\left(\boldsymbol{\alpha}_t^{1}\right) \tag{41}$$

$$\bar{\bar{\mathbf{f}}}_t^{1} = \text{cumsum}\left(\boldsymbol{\lambda}_t^{1}\right) \tag{42}$$

$$\bar{\bar{\mathbf{i}}}_t^{2} = 1 - \text{cumsum}\left(\boldsymbol{\alpha}_t^{2}\right) \tag{43}$$

$$\bar{\bar{\mathbf{f}}}_t^{2} = \text{cumsum}\left(\boldsymbol{\lambda}_t^{2}\right) \tag{44}$$

Then, the attention hierarchy structure is partitioned using the following equations:

$$\mathbf{w}_t^0 = \bar{\bar{\mathbf{i}}}_t^1 \circ \bar{\bar{\mathbf{f}}}_t^1 \tag{45}$$

$$\mathbf{w}_t^1 = \bar{\bar{\mathbf{f}}}_t^1 - \mathbf{w}_t^0 \tag{46}$$

$$\mathbf{w}_t^2 = \mathbf{w}_t^0 \circ \left(\bar{\bar{\mathbf{f}}}_t^2 - \mathbf{w}_t^3\right) \tag{47}$$

$$\mathbf{w}_t^3 = \mathbf{w}_t^0 \circ \left(\bar{\bar{\mathbf{f}}}_t^2 \circ \bar{\bar{\mathbf{i}}}_t^2\right) \tag{48}$$

$$\mathbf{w}_t^4 = \mathbf{w}_t^0 \circ \left(\bar{\bar{\mathbf{i}}}_t^2 - \mathbf{w}_t^3\right) \tag{49}$$

$$\mathbf{w}_t^5 = \bar{\bar{\mathbf{i}}}_t^1 - \mathbf{w}_t^1 \tag{50}$$

In the equation,$\mathbf{w}_t^1$ , $\mathbf{w}_t^2$,$\mathbf{w}_t^3$ , $\mathbf{w}_t^4$ and $\mathbf{w}_t^5$ represent the high attention hierarchy, high intermediate attention hierarchy, intermediate attention hierarchy, low intermediate attention hierarchy, and low attention hierarchy, respectively.

Finally, with the above equations, the propagation equation of AGMLSTM can be written as follows:

$$\begin{cases} \mathbf{f}_t = \sigma\left(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f\right) \\ \mathbf{i}_t = \sigma\left(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i\right) \\ \mathbf{o}_t = \sigma\left(\mathbf{W}_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o\right) \\ \bar{\mathbf{c}}_t = \tanh\left(\mathbf{W}_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c\right) \\ \hat{\mathbf{c}}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \\ \mathbf{c}_t = \mathbf{w}_t^1 \odot \mathbf{c}_{t-1} + \mathbf{w}_t^2 \odot \left[z_2 \odot \hat{\mathbf{c}}_t + (1 - z_2) \odot \mathbf{c}_{t-1}\right] \\ + \mathbf{w}_t^3 \odot \hat{\mathbf{c}}_t + \mathbf{w}_t^4 \odot \left[z_1 \odot \hat{\mathbf{c}}_t + (1 - z_1) \odot \bar{\mathbf{c}}_t\right] + \mathbf{w}_t^5 \odot \bar{\mathbf{c}}_t \\ \mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right) \end{cases} \tag{51}$$

where other parameters are the same as LSTM.

## RUL prediction approach

A health indicator (HI) that can accurately show the degradation process of gears is crucial to the performance of the prediction model. Therefore, the HI of the vibration signal obtained by the trained diffusion model is used in the article for gear RUL prediction, whose superiority has been demonstrated. Considering that most DL approaches for gear RUL prediction are pattern recognition methods, which are influenced by the quantity and quality of data, an RUL prediction approach under limited samples[20] is used in the article, whose flowchart is shown in Fig. 7 and the details are presented as follows:

1. The HI data $\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \dots & z_{n-1} & z_n \end{bmatrix}$ is calculated based on the full-lifecycle vibration data by the sampling approach whose sampling time is $T$ and sample interval is $\Delta t$.

2. Then the first part $\mathbf{z}' = \begin{bmatrix} z_1 & z_2 & \dots & z_{m-1} & z_m \end{bmatrix}$ of $\mathbf{z}$ is chosen and linearly normalized to obtain $\mathbf{V} = \begin{bmatrix} v_1 & v_2 & \dots & v_{m-1} & v_m \end{bmatrix}$.

3. Training pair, containing the model input $\begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 & \dots & \mathbf{G}_{l-1} & \mathbf{G}_l \end{bmatrix}^{\mathrm{T}}$ and output $\mathbf{G}_{l+1}$, is reconstructed by:

$$\mathbf{G} = \begin{bmatrix} v_1 & v_2 & \cdots & v_{m-l} \\ v_2 & v_3 & \cdots & v_{m-l+1} \\ \vdots & \vdots & \ddots & \vdots \\ v_{l+1} & v_{l+2} & \cdots & v_m \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \\ \vdots \\ \mathbf{G}_{l+1} \end{bmatrix} \tag{52}$$

where the value of $l$ is equal to the neural numbers of the input layer and $\mathbf{G}_i$ is denoted by:

$$\mathbf{G}_i = \begin{bmatrix} v_i & v_{i+1} & \cdots & v_{m-l+i-1} \end{bmatrix} \tag{53}$$

4. The training loss $L$ of the proposed model is denoted as the mean square error (MSE) between the last row $\mathbf{G}_{l+1}$ and the predicted $\hat{\mathbf{G}}_{l+1}$ based on the first $l$ rows of the matrix $\mathbf{G}$.

$$\mathbf{y}_t = \hat{\mathbf{G}}_{l+1} = f(\mathbf{G}_1, \mathbf{G}_2 \cdots, \mathbf{G}_l) \tag{54}$$

$$\min\left[L(\mathbf{w}, \mathbf{b}, \mathbf{s})\right] = \frac{1}{2}\left[\hat{\mathbf{G}}_{l+1} - \mathbf{G}_{l+1}\right]^2 \tag{55}$$

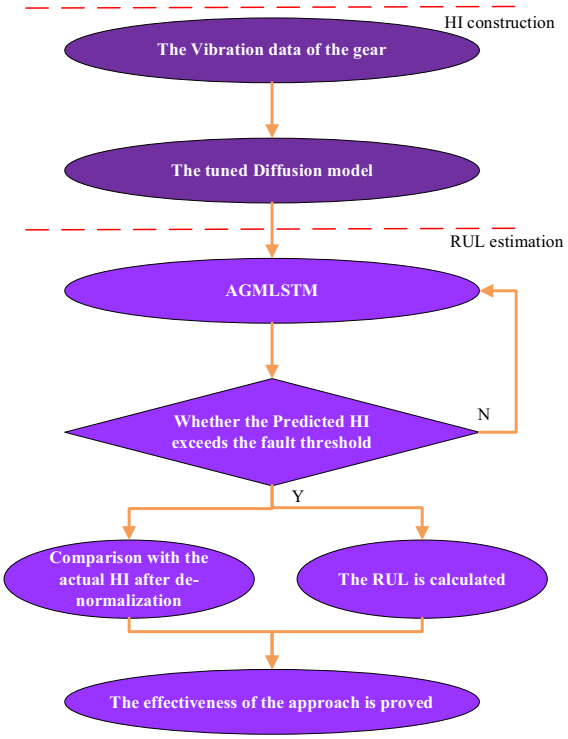


**Figure 7.** The flowchart of the proposed RUL prediction approach.

where $f$ denotes the model transaction function; $\mathbf{w}$, $\mathbf{b}$, and $\mathbf{s}$ separately denote the learning matrix.

5. After the trained proposed method is obtained, the last $l$ is set as the model input to estimate the HI in the next point. Then the step-by-step prediction is executed by:

$$\begin{aligned}
\mathbf{G}_{l+2} &= f\left(\mathbf{G}_2, \mathbf{G}_3 \cdots, \mathbf{G}_{l+1}\right) \\
\mathbf{G}_{l+3} &= f\left(\mathbf{G}_3, \mathbf{G}_4 \cdots, \mathbf{G}_{l+1}, f\left(\mathbf{G}_2, \mathbf{G}_3 \cdots, \mathbf{G}_{l+1}\right)\right) \\
&\vdots \\
\mathbf{G}_n &= f\left(\mathbf{G}_{n-l}, \mathbf{G}_{n-l+1} \cdots, \mathbf{G}_{n-2}, f\left(\mathbf{G}_{l-k-1}, \mathbf{G}_{n-l} \cdots, \mathbf{G}_{n-2}\right)\right)
\end{aligned} \tag{56}$$

6. At last, once the failure threshold is lower than the inversely normalized predicted HIs, the estimated RUL $\overline{Rul}$ is finally obtained by Eq. (57):

$$\overline{Rul} = n_1 \times \Delta t \tag{57}$$

where $n_1$ is the number of predicted HI points before exceeding the threshold. And the actual RUL shows the effectiveness of the proposed method.

## Model optimization

The configuration exploration of the predictive model is executed based on grid search. The hyper-parameters, namely, candidates of learning rate $\alpha$ and neuron number in each layer, are constructed as each grid note, which is searched for optimal predictive performance parameters.

The weight matrix $\mathbf{w}$, the bias matrix $\mathbf{b}$, and the proportion matrix $\mathbf{s}$ of the model are trained during the training stage based on the loss function Eq. (55) and updated on Eq. (58) by Adam optimizer.

$$\begin{bmatrix} \mathbf{w}_{t+1} \\ \mathbf{b}_{t+1} \\ \mathbf{s}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_t \\ \mathbf{b}_t \\ \mathbf{s}_t \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L_t}{\partial \mathbf{w}_t} \\ \frac{\partial L_t}{\partial \mathbf{b}_t} \\ \frac{\partial L_t}{\partial \mathbf{s}_t} \end{bmatrix} \tag{58}$$

## Experimental analysis

Several fatigue full-life experiments are executed by a gear contact fatigue test rig to investigate the lifespan of gears from normal conditions to failure (tooth broken and pitting). The material of the gear for the tooth fracture case was 40Cr, while the gear material for the pitting case was 20CrMnTi. The gear module was set to 5, and the experimental gear case had an oil flow rate of 4 L/h with a cooling temperature of 70 °C. The gears that experienced tooth-broken failures (Dataset 1 and Dataset 2) had tooth counts of 31, 25, 25, and 31, respectively. On the other hand, the gears that suffered from pitting failures (Dataset 3 and Dataset 4) had tooth counts of 26, 24, 24, and 26, as shown in Table 1.

As depicted in Fig. 8, the experimental setup comprises a torque controller, a cooling and lubrication controller, an experimental operation platform, and a gear operation platform. The sampling frequency for the experimental setup is fixed at 50,000 Hz. To minimize data volume, this study sets the recording interval, and the sampling length are 60 s and 10 s. And Part of the healthy state data at the beginning of the run is deleted. Data sets 1 and 3 are used to train the Diffusion model for calculating gear HIs. Then, the trained Diffusion model is used to encode the health indicator points of data sets 2 and 4. To test the prediction ability of the predictive model, this study conducts experiments using the health indicator points from all data sets. Through grid search, optimal hyper-parameters for the AGMLSTM are obtained. For data sets 1, 3, and 4, the number of neurons in the input, hidden, and output layers of AGMLSTM are set to 100, 35, and 1. For data set 2, they are set to 60, 20, and 1. The learning rates for the models on data sets 1, 2, 3, and 4 are set to 0.02, 0.03, 0.05, and 0.05.

Appropriate health indicators can effectively reflect the health condition of mechanical equipment and improve the RUL prediction capability[26–30]. Due to the limitations of single features such as root mean square, kurtosis, and frequency centroid, they may not adequately capture the degradation trend of mechanical equipment in most data sets. Therefore, this study develops a health indicator based on diffusion model that can be used in most cases. Since the signals collected during the steady-state phase contain less degradation information,

| Dataset | Data 1 | Data 2 | Data 3 | Data 4 |
|---|---|---|---|---|
| Load (KN) | 1.4 | 1.4 | 1.2 | 1.2 |
| Speed (rpm) | 500 | 500 | 1000 | 1000 |
| Test time (min) | 814 | 820 | 696 | 951 |
| Number of samples | 814 | 820 | 696 | 951 |
| Failure mode | Broken | Broken | Pitting | Pitting |

**Table 1.** Description of data.

**Figure 8.** Gear contact fatigue testing machine.

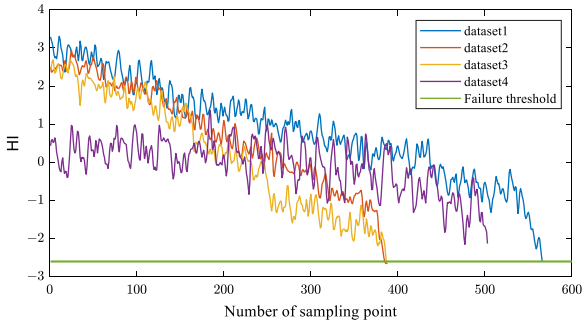


**Figure 9.** HI of four datasets.

only a portion of the samples from the lifecycle data set is used to calculate the health indicator points using diffusion model and then applied to remaining useful life prediction. Figure 9 displays the obtained health indicator points for all four gear data sets. The constructed health indicator point curves can effectively reflect the degradation trend of gear health, which is highly beneficial for RUL prediction. All gear health indicator curves exhibit a declining trend, and their failure thresholds are similar. This aids in setting a unified failure threshold for different experimental setups, thereby enhancing the robustness of gear RUL prediction.

The study undertook comparative experiments employing distinct optimization algorithms to underscore the superior performance of the chosen optimizer. Specifically, SGDM[31], RMSprop[32], and Adam were deliberately selected for comparison within a consistent structural framework, and subsequent optimization was applied across all models. The evaluation process involved ten parallel experiments for each model, focusing on a one-hour prediction task. Model performance was rigorously assessed using key performance indicators, namely the mean absolute error (MAE), the normalized root mean square error (NRMSE), the mean absolute percentage error (MAPE), and Score[23], as presented in Fig. 10.

It can be concluded that the model adopted by Adam has the lowest values of MAE, NRMSE, and MAPE, and the highest Score value. This means that with the Adam optimizer, the proposed method has better RUL prediction performance. Thus, Adam is more suitable for the proposed method when it deals with gear RUL prediction.

The evaluation indicators of different HIs for different gear datasets are respectively calculated and the mean value of evaluation indicators are listed in Table 2. First the two widely used statistical features such as RMSE and Kurtosis in PHM[20] are chosen as HIs. Then HI based on popularity learning is constructed, i.e. PCA. Finally,

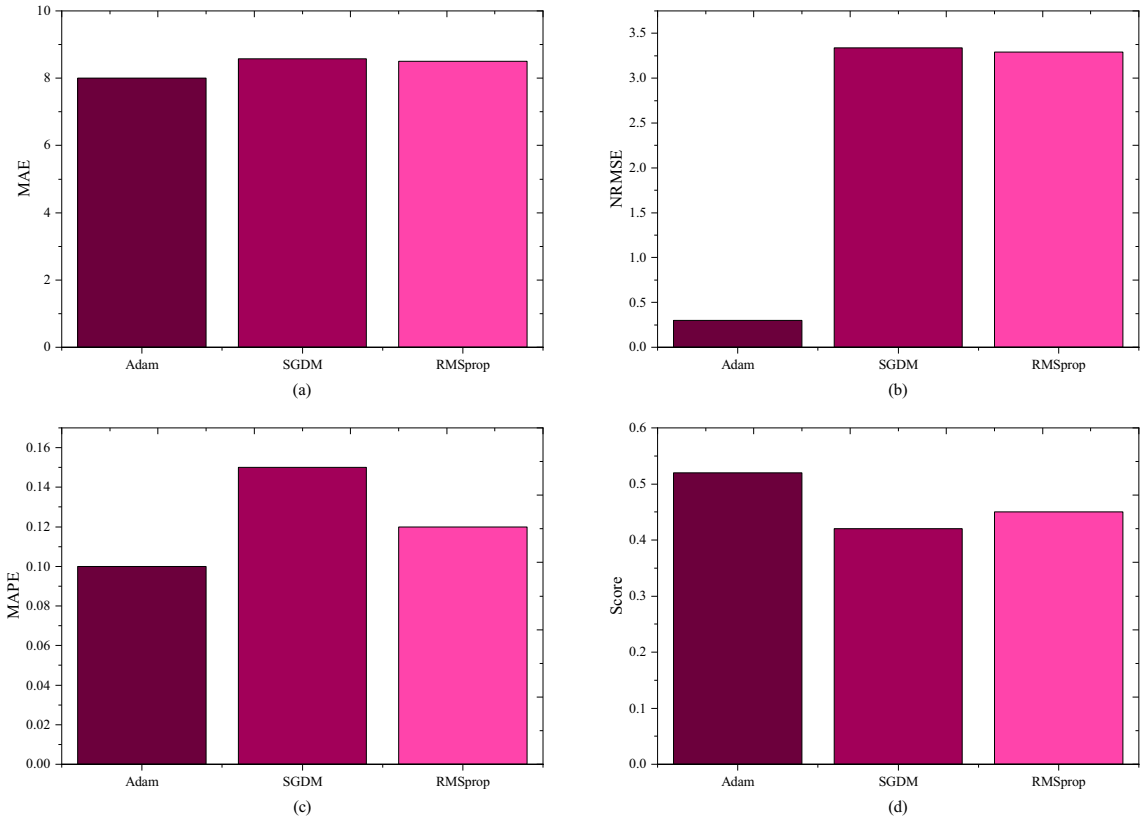


**Figure 10.** Comparison of predictive ability under different optimizers.

| | Mon | Corr | Rob | CI |
|---|---|---|---|---|
| RMSE | 0.519 | 0.706 | 0.826 | 0.636 |
| Kurtosis | 0.414 | 0.618 | 0.737 | 0.540 |
| PCA | 0.754 | 0.863 | 0.674 | 0.771 |
| DBN | 0.916 | 0.802 | 0.653 | 0.829 |
| VAE | 0.829 | 0.853 | 0.830 | 0.837 |
| Diffusion model | 0.955 | 0.921 | 0.890 | 0.932 |

**Table 2.** The evaluation indicators of different HIs for datasets.

HIs are constructed by other Unsupervised networks deep belief network (DBN)[33], and variational autoencoder (VAE)[24].

In Table 2, a comprehensive analysis of the evaluation indicators for HIs reveals that those generated by the diffusion model consistently outperform other HIs across gear datasets. Notably, the values of monotonicity and the comprehensive indicator for the diffusion model-reconstructed HI stand out, reaching impressive scores of 0.955 and 0.921, respectively. This signifies that the HI constructed through the diffusion model adeptly captures and reflects the degradation trend in gear datasets. The comparison across different HIs reveals that those generated by DBN, VAE, and the diffusion model surpass those based on PCA, RMSE, and Kurtosis. This suggests that HIs constructed by neural networks exhibit greater flexibility when dealing with HIs under fixed patterns, although they may not be ideal for reflecting the degradation trend in gear datasets. Besides, the diffusion model stands out by delivering strong performance evaluation results. This highlights its superior generalization ability, indicating that the HIs produced by the diffusion model are well-suited for assessing health status in gear datasets. Consequently, the HIs constructed by the diffusion model effectively and reliably capture the degradation trend in gear systems.

Using the small-sample life prediction method, the proposed AGMLSTM is compared with classical models (LSTM, GRU) and published deep learning models, i.e. Gated dual attention unit (GDAU)[20], On-LSTM[21], Coctail LSTM (CLSTM)[24], for RUL prediction on the four gear data sets. To compare the prediction accuracy and robustness of each method, grid search is used to obtain the optimal hyper-parameters for each model, and then all tuned networks are tested 10 times on each gear data set. The prediction task is set as predicting 60 HI points (1-h RUL) for the comparative experiment, comparing the prediction capabilities of the benchmark models. Based on the experimental prediction results, MAE, NRMSE, MAPE and Score are used to quantitatively evaluate the prediction accuracy, as shown in Fig. 11.

As illustrated in Fig. 11, the superiority of the proposed AGMLSTM model over other counterparts is evident, showcasing exceptional performance in predicting RUL. This observation underscores the significant impact of incorporating comprehensive ordered information, especially when employing attention mechanisms at the hidden layer level. The strategic utilization of attention mechanisms facilitates the network models in effectively navigating data heterogeneity, leading to a remarkable enhancement in RUL estimation accuracy. Based on the actual gear tests, the outperformance of the prosed RUL prediction method is proven by MAE, NRMSE, MAPE, and Score, with improvement of 33%, 40%, 17%, and 8% respectively compared with the state-of-art. Consequently, the proposed models emerge as highly apt for the precise prediction of gear remaining useful life, attributing their success to the adept utilization of ordered information and attention-guided learning mechanisms.

AGMLSTM and CLSTM refine the mixed hierarchy through fine-grained processing based on the introduced main and auxiliary gating mechanisms. The distinction lies in the fact that AGMLSTM employs an attention mechanism for hierarchical localization. Consequently, while AGMLSTM and CLSTM achieve better RUL prediction accuracy compared to ON-LSTM and GDAU, they come with an increased parameter count. With the

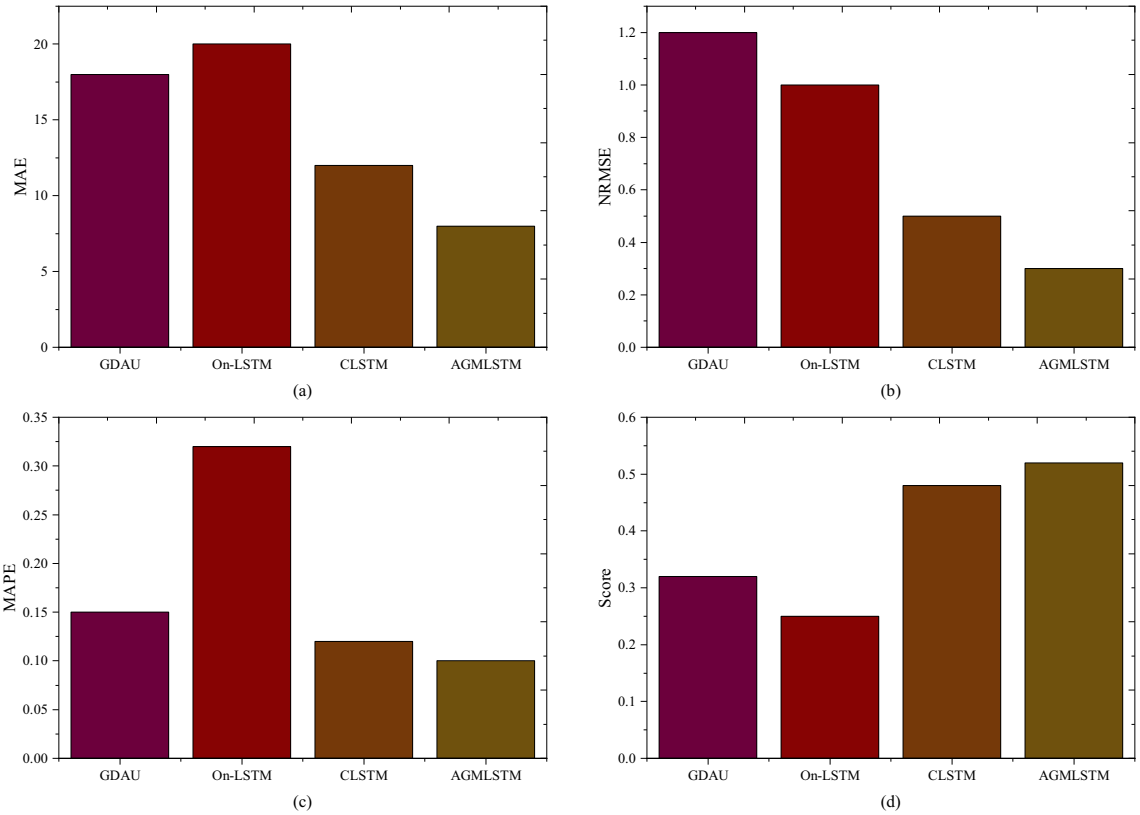


**Figure 11.** The gear RUL estimation performance of different methods.

| | On-LSTM | GDAU | CLSTM | AGMLSTM |
|---|---|---|---|---|
| Train time in each epoch(s) | 0.12 | 0.15 | 0.17 | 0.18 |

**Table 3.** The complexity analysis of models.

same number of hidden layer neurons $L_n$, AGMLSTM increases the parameter count compared to $8 \times L_n$ ON-LSTM and $16 \times L_n$ GDAU, and is approximately equivalent to CLSTM. To provide a more intuitive representation of the network's computational complexity, we calculated the time required for each iteration during the training process on the same computer device, as shown in Table 3.

From Table 3, it is evident that AGMLSTM and CLSTM incur a higher time cost than On-LSTM. This is attributed to the different hierarchical learning mechanisms these models employ for input processing, with additional gating units introducing more network parameters. The GDAU, which incorporates dual attention gates, exhibits a similar phenomenon. Additionally, it is crucial to note that the training phase is offline, and during the online prediction phase, the trained AGMLSTM incurs a prediction time of only $7.8 \times 10^{-5}$ s. Hence, the prediction time overhead of AGMLSTM is deemed acceptable considering its superior long-term RUL prediction accuracy.

Based on the above analysis, the rational and comprehensive use of ordered information is crucial for enhancing the accuracy of gear RUL prediction, especially in cases where known samples contain less gear degradation information. Therefore, the proposed method AGMLSTM, guided by an attention mechanism for multi-hierarchy partitioning, effectively extracts more gear state degradation information, resulting in superior overall RUL prediction performance compared to other methods.

Illustrating the robustness of our proposed small-sample intelligent prediction method, we employ data set 3 as a paradigmatic case study, harnessing the AGMLSTM model for an insightful exploration of RUL prediction across diverse forecast horizons. The delineation of the training set, consisting of known data from the initial segment, and the validation set, featuring unknown data from the subsequent portion, lays the groundwork for a comprehensive evaluation. Intriguingly, the AGMLSTM model's prowess is vividly showcased through an in-depth analysis of its predictive capabilities on data set 1, where the focus is squarely on anticipating 90, 70, and 50 HIs. As delineated in Figs. 12, 13 and 14, a compelling narrative unfolds, elucidating a direct correlation between the increasing number of known HIs and the model's augmentation in prediction proficiency. The figures distinctly reveal a convergence of estimated health indicator points towards their true counterparts, affirming the method's precision and efficacy. Crucially, the overarching alignment between prediction values and actual values across a spectrum of forecast instances underscores the AGMLSTM model's unparalleled effectiveness in gear RUL prediction. This nuanced ability to predict with heightened precision as our understanding of health indicators expands substantiates the model's robustness and underscores its potential for real-world applications.

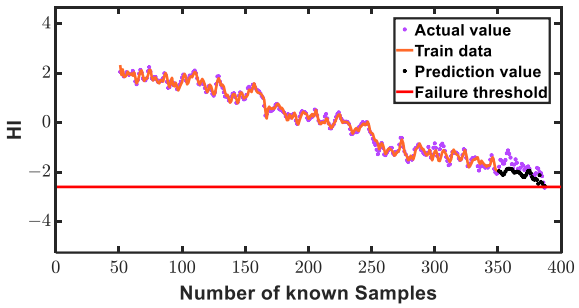


**Figure 12.** Prediction illustration for 30 predicted points of data 3.

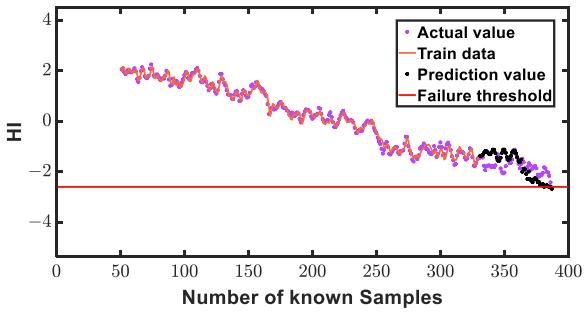


**Figure 13.** Prediction illustration for 60 predicted points of data 3.

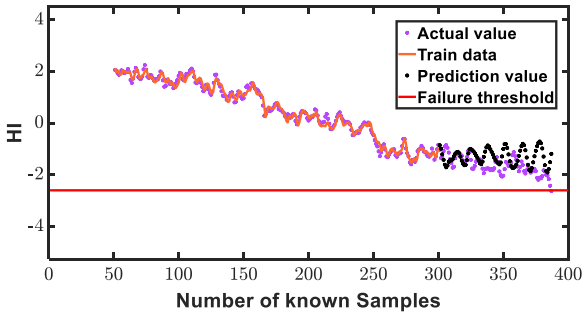


**Figure 14.** Prediction illustration for 90 predicted points of data 3.

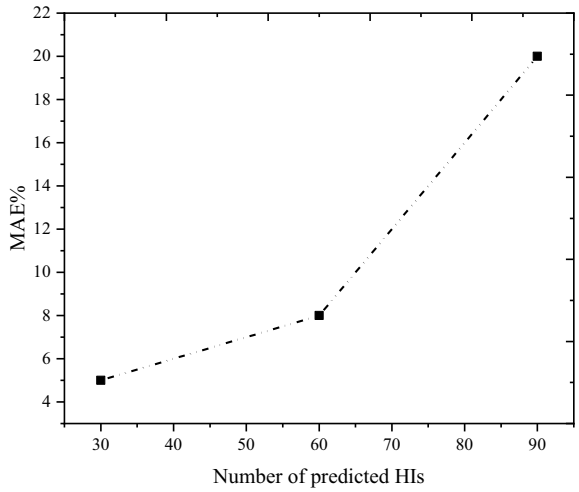


**Figure 15.** MAEs of RUL prediction results under different known HI points.

In Fig. 15, the prowess of AGMLSTM in predicting RUL at varying known health indicator points is rigorously assessed using the MAE. A compelling trend unfolds, revealing a noteworthy inverse correlation: the MAE values exhibit a consistent decline as the number of health indicator points rises. This observation underscores the model's heightened proficiency with an expanding set of health indicators. Examining specific instances, for a prediction involving 30 health indicator points, the RUL prediction boasts a mere 5% percentage error. Intriguingly, with an escalation to 60 health indicator points, the percentage error marginally increases to 8%. The augmentation of known HIs entails the incorporation of expanding HIs encompassing fault information into the model training process. This influx of HIs allows the model to assimilate a broader spectrum of fault trends, leading to a progressive enhancement in its predictive capabilities. These outcomes signify AGMLSTM's commendable performance in protracted RUL prediction, showcasing its capacity for sustained accuracy. To further underscore the model's prowess in long-term RUL estimation, a bold attempt is made to predict 90 health indicator points, as illustrated in Fig. 14 Despite a 25% error in the computed result, this endeavor unequivocally establishes AGMLSTM's formidable predictive aptitude for enduring gear RUL scenarios.

## Conclusion

Revolutionizing gear RUL prediction, our groundbreaking approach introduces a novel methodology by constructing HIs through a diffusion model, coupled with the innovative AGMLSTM predictor. Leveraging the temporal and frequency characteristics of vibration measurements, the diffusion model lays the foundation for a distinctive gear HI. This HI, in turn, serves as the linchpin for AGMLSTM, a pioneering predictor designed to comprehensively and judiciously mine ordered information for precise gear RUL forecasts. The strategic incorporation of rich ordered information significantly amplifies the feature extraction capabilities of our predictor, leading to a substantial enhancement in RUL prediction accuracy. Validation through rigorous real-world gear tests unequivocally demonstrates the superior performance of our proposed RUL prediction method. Employing widely accepted evaluation metrics, our approach realizes 8 on MAE, 0.3 on NRMSE, 0.1 on MAPE, and 0.52 on Score, showcasing an impressive improvement of 33%, 40%, 17%, and 8% respectively, compared to state-of-the-art methods. In essence, our proposed approach emerges as the pinnacle of gear RUL prediction methodologies, providing not only heightened accuracy but also unparalleled effectiveness in real-world scenarios.

The proposed methodology in this study primarily addresses the RUL under conditions of single-tooth breakage or pitting failure. However, in practical engineering applications, failures frequently involve the coupling of multiple faults. Therefore, the development of a methodology for predicting the RUL in cases of complex gearbox failure is of significant importance. This aspect will be a key focus of our future research endeavors.

## Data availability

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

Received: 26 July 2023; Accepted: 15 January 2024
Published online: 20 January 2024

## References

1. Cheng, F., Qu, L. & Qiao, W. Fault prognosis and remaining useful life prediction of wind turbine gearboxes using current signal analysis. *IEEE Trans. Sustain. Energy* **9**, 157–167 (2017).
2. Zhang, H., Chen, X., Chen, W. & Shen, Z. Collaborative sparse classification for aero-engine's gear hub crack diagnosis. *Mech. Syst. Signal Process* **141**, 106426 (2020).
3. Ren, L., Cui, J., Sun, Y. & Cheng, X. Multi-bearing remaining useful life collaborative prediction: A deep learning approach. *J. Manuf. Syst.* **43**, 248–256 (2017).
4. Yang, B., Liu, R. & Zio, E. Remaining useful life prediction based on a double-convolutional neural network architecture. *IEEE Trans. Ind. Electron.* **66**, 9521–9530 (2019).
5. Cheng, C. *et al.* A deep learning-based remaining useful life prediction approach for bearings. *IEEE ASME Trans. Mech.* **25**, 1243–1254 (2020).
6. Wei, J., Dong, G. & Chen, Z. Remaining useful life prediction and state of health diagnosis for lithium-ion batteries using particle filter and support vector regression. *IEEE Trans. Ind. Electron.* **65**, 5634–5643 (2017).
7. Cui, L., Wang, X., Wang, H. & Ma, J. Research on remaining useful life prediction of rolling element bearings based on time-varying Kalman filter. *IEEE Trans. Instrum. Meas.* **69**, 2858–2867 (2019).
8. Ahmad, W., Khan, S. A. & Kim, J.-M. A hybrid prognostics technique for rolling element bearings using adaptive predictive models. *IEEE Trans. Ind. Electron.* **65**, 1577–1584 (2017).
9. Song, Y., Liu, D., Yang, C. & Peng, Y. Data-driven hybrid remaining useful life estimation approach for spacecraft lithium-ion battery. *Microelectron. Reliab.* **75**, 142–153 (2017).
10. Huang, Y., Tang, B., Deng, L. & Zhao, C. Fuzzy analytic hierarchy process-based balanced topology control of wireless sensor networks for machine vibration monitoring. *IEEE Sens. J.* **20**, 8256–8264 (2020).
11. Qin, Y. A new family of model-based impulsive wavelets and their sparse representation for rolling bearing fault diagnosis. *IEEE Trans. Ind. Electron.* **65**, 2716–2726 (2017).
12. Zhao, R. *et al.* Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process* **115**, 213–237 (2019).
13. Graves, A. & Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*. 37–45 (2012).
14. Huang, C.-G., Huang, H.-Z. & Li, Y.-F. A bidirectional LSTM prognostics method under multiple operational conditions. *IEEE Trans. Ind. Electron.* **66**, 8792–8802 (2019).
15. Wu, J. *et al.* Data-driven remaining useful life prediction via multiple sensor signals and deep long short-term memory neural network. *ISA Trans.* **97**, 241–250 (2020).
16. Yuan, M., Wu, Y. & Lin, L. *2016 IEEE International Conference on Aircraft Utility Systems (AUS)*. 135–140 (IEEE, 2016).
17. Wang, C., Lu, N., Cheng, Y. & Jiang, B. A data-driven aero-engine degradation prognostic strategy. *IEEE Trans. Cybern.* **51**, 1531–1541 (2019).
18. 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. Indus. Inform.* **15**, 5023–5032 (2019).
19. Chen, Z. *et al.* Machine remaining useful life prediction via an attention-based deep learning approach. *IEEE Trans. Ind. Electron.* **68**, 2521–2531 (2020).
20. Qin, Y., Chen, D., Xiang, S. & Zhu, C. Gated dual attention unit neural networks for remaining useful life prediction of rolling bearings. *IEEE Trans. Indus. Inform.* **17**, 6438–6447 (2020).
21. Shen, Y., Tan, S., Sordoni, A. & Courville, A. Ordered neurons: Integrating tree structures into recurrent neural networks. arXiv preprint arXiv:1810.09536 (2018).
22. Yan, H., Qin, Y., Xiang, S., Wang, Y. & Chen, H. Long-term gear life prediction based on ordered neurons LSTM neural networks. *Measurement* **165**, 108205 (2020).
23. Xiang, S., Qin, Y., Zhu, C., Wang, Y. & Chen, H. LSTM networks based on attention ordered neurons for gear remaining life prediction. *ISA Trans.* **106**, 343–354 (2020).
24. Xiang, S., Zhou, J., Luo, J., Liu, F. & Qin, Y. Cocktail LSTM and its application into machine remaining useful life prediction. *IEEE ASME Trans. Mech.* **2**, 23 (2023).
25. Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **99**, 1–20 (2023).
26. Guo, L., Li, N., Jia, F., Lei, Y. & Lin, J. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing* **240**, 98–109 (2017).
27. Pan, Y., Wu, T., Jing, Y., Han, Z. & Lei, Y. Remaining useful life prediction of lubrication oil by integrating multi-source knowledge and multi-indicator data. *Mech. Syst. Signal Process* **191**, 110174 (2023).
28. Li, X. *et al.* Feature fusion model based health indicator construction and self-constraint state-space estimator for remaining useful life prediction of bearings in wind turbines. *Reliab. Eng. Syst. Saf.* **233**, 109124 (2023).
29. Peng, Z., Huang, X., Tang, D. & Quan, Q. Health indicator construction based on multisensors for intelligent remaining useful life prediction: A reinforcement learning approach. *IEEE Trans. Instrum. Meas.* **72**, 1–13 (2023).
30. Lei, Y. *et al.* Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process* **104**, 799–834 (2018).
31. Postalcıoğlu, S. Performance analysis of different optimizers for deep learning-based image recognition. *Int. J. Pattern Recognit. Artif. Intell.* **34**(2), 2051003 (2020).
32. Elshamy, R., Abu-Elnasr, O., Elhoseny, M. & Elmougy, S. Improving the efficiency of RMSProp optimizer by utilizing Nestrove in deep learning. *Sci. Rep.* **13**(1), 8814 (2023).
33. Hu, C.-H. *et al.* A prognostic model based on DBN and diffusion process for degrading bearing. *IEEE Trans. Indus. Electron.* **67**(10), 8767–8777 (2019).

## 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
X.C. wrote the main manuscript text.

## Competing interests
The author declares 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.