Macroscopic-microscopic attention in LSTM networks based on fusion features for gear remaining life prediction

Yi Qin, *Member, IEEE*, Sheng Xiang, Yi Chai, Haizhou Chen

***Abstract*—In the mechanical transmission system, the gear is one of the most widely used transmission components. The failure of the gear will cause serious accidents and huge economic loss. Therefore, the remaining life prediction of the gear is of great importance. In order to accurately predict the remaining life of the gear, a new type of long-short-term memory neural network with macro-micro attention is proposed. First, some typical time-domain and frequency-domain characteristics of vibration signals are calculated respectively, such as the maximum value, the absolute mean value, the standard deviation, the kurtosis and so on. Then, the principal component of these characteristics is extracted by the isometric mapping method. The importance of fusional characteristic information is filtered via a proposed macro-micro attention mechanism, so that the input weight of neural network data and recursive data can reach multi-level real-time amplification. With the new long short-term memory neural network, the health characteristics of gear vibration signals can be predicted based on the known fusion features. The experimental results show that this method can predict the remaining life of gears and bearings well, and it has higher prediction accuracy than the conventional prediction methods.**

***Index Terms*—remaining life prediction; long short-term memory neural network; macro-micro attention mechanism; vibration signal; feature fusion.**

I. Introduction

Gear is one of the most universally used mechanical parts, which is widely used in mechanical equipment [1]. The reason why gear has such strong and lasting vitality is due to its unique advantages: high transmission efficiency, compact structure, good transmission smoothness, large bearing capacity, long service life, and so on [2]. Under the complex working condition and environment, gears are easily subject to failures, which may result in the catastrophe of the machine running and even threaten the personal safety [3-5]. This is particularly true for the large-scale or extra-large equipment, such as hydro-generators, mine-conveying machinery, power transmission systems for helicopters, heavy-duty machine tools, and so on [6-8]. Compared with commonly-used gear fault diagnosis [9,10], gear remaining life prediction is beneficial to determine the equipment maintenance time reasonably, improve the production efficiency, reduce the accident rate, and prevent the sudden accidents, which is significant for engineering production [11].

Recently, with the rapid developments in sensing, signal processing and artificial intelligence technology, data-driven methods for prognostics and health management (PHM) have gradually become the mainstream solution either in fault diagnosis or remaining useful life（RUL）estimation, instead of physics-based methods which can be expensive and tedious to develop [12-15]. Deep Learning is one of the most advanced method in data-driven methods for the study of RUL prediction. Zhu et al. [16] proposed multiscale convolutional neural network (MCNN) to predict RUL of bearings, which keeps the global and local information synchronously compared to a traditional convolutional neural network (CNN). Xiang et al. [17] proposed a novel deep convolutional neural network-based method using raw data for remaining useful life predictions. Zhang et al. [18] proposed a transfer learning algorithm based on recurrent neural networks for RUL estimation, in which the models can be first trained on different but related datasets and then fine-tuned by the target dataset. Extensive experimental results show that transfer learning can in general improve the prediction models on the dataset with a small number of samples.

Recurrent Neural Network (RNN) differs from traditional artificial Neural Network in that it can utilize the current input information as well as historical information. Based on the particular design, RNN is suitable for processing timing-related information [19]. However, RNN also has its own drawbacks, e.g. the excessive recursion time is indirectly equivalent to increasing the depth of neural network and the

Manuscript received April 26, 2019; revised August 7, 2019; October 9, 2019 and November 11, 2019: accepted November 26, 2019. This work was supported in part by National Key R&D Program of China under grant 2018YFB2001300, in part by the National Natural Science Foundation of China under Grant 51675065, in part by Chongqing Research Program of Basic Research and Frontier Technology under Grant cstc2017jcyjAX0459, and in part by the Fundamental Research Funds for the Central Universities under Grant 2018CDQYJX0011.

Y. Qin and S. Xiang are with the State Key Laboratory of Mechanical Transmission, College of Mechanical Engineering, Chongqing University, Chongqing, 400044, China (e-mail: qy_808@aliyun.com; a1121623518@163.com). Yi Chai is with College of Automation, Chongqing University, Chongqing, 400044, China (e-mail: chaiyi@cqu.edu.cn). H. Chen is with the Qingdao University of Science and Technology, Qingdao 266000, China (e-mail: chenhaizhou84@163.com).

training time; the vanishing gradient problem usually occurs [20]. In order to solve these problems, long short-term memory (LSTM) was proposed by Hochreiter and Schmidhuber in 1997 [21]. It can avoid the long-term dependence problem of RNN, thus it has been widely used. Yuan et al. [22] investigated three RNN models including vanilla RNN, LSTM and gated recurrent unit (GRU) models for fault diagnosis and prognostics of aero engine. They found that these advanced RNN models based on LSTM and GRU performed better than the conventional RNN via a number of experiments. Another interesting observation was that the ensemble model of the above three RNN variants did not boost the performance of LSTM. Wang et al. [23] proposed the residual LSTM network structure to solve the network degradation problem of the deep LSTM model by referring to the residual network, and verified the superiority of the network in text prediction.

Attention mechanism can be regarded as a kind of contribution screening of information which improves the efficiency of neural network by selecting key information for processing. Attention-based recurrent networks have been successfully applied to a wide variety of tasks, such as handwriting synthesis [24], machine translation [25], image caption generation [26] and visual object classification [27]. In the prediction field, attention-based LSTM is getting more and more attention. Ran et al. [28] substituted a tree structure with attention mechanism for the unfolding way of standard LSTM to construct the depth of LSTM and modeling long-term dependence for travel time prediction. Fernando et al. [29] combined two kinds of attention and used LSTM for human trajectory prediction and abnormal event detection. Filtering key information can reduce computing resources, but it can also cause some degree of information loss. But differentiation treatment of input data according to the screening of attention mechanism can not only reflect the focus to important information, but also retain useful information as far as possible.

The commonly-used data-driven prediction methods are essentially pattern recognition approaches, so there is an obvious disadvantage that the training data and the test data are supposed to have the similar health deterioration trends; otherwise, the estimated results of the inference model are unreliable or even nonsensical [30]. And in engineering practice, some equipment health degradation data is limited or even incomplete. So it is necessary to research RUL prediction with a small number of samples. The life prediction method proposed in this paper is to fit the known data of gear health degradation by using the novel neural network we proposed, then predict the remaining life of the gear through its degradation curve.

Different from the application of the above attention mechanism in LSTM network, this paper takes the attention coefficient as the contribution evaluation standard, and the macro-micro attention (MMA) mechanism is proposed to amplify the weights of both input data and recurrent data according to their contribution degree. Instead of only using the input data with the largest attention coefficient in the current methods, all input data and recurrent data are applied to multiple differentiated handling in the proposed LSTM. Finally, with the fused feature of the gear vibration signal based on 13 time domain features, 4 frequency domain features and 4 envelope spectrum features, the gear remaining life prediction method based on the macro-micro attention mechanism LSTM (MMALSTM) is explored. The experiment results show that the proposed method can achieve better performance in gear RUL prediction than the traditional methods. The contributions of this paper can be summarized as follows:

(1) We employ an MMA model structure to deal with the fusion features for RUL prediction. Based on the importance of fusion features got by MMA, the input weight and recurrent weight of fusion features can reach multi-level real-time amplification.
(2) The proposed systematic approach integrates MMA and LSTM into a framework, which automatically estimates RUL and has higher prediction accuracy than conventional methods.

The rest of this paper is organized as follows. Section II not only introduces LSTM, but also introduces attention mechanism. Section III introduces a new type of LSTM with macro-micro attention mechanism and explains the proposed gear remaining useful life prediction method in detail. Section IV conducts an experimental case study on the gear life data and bearing life data to verify the effectiveness of the proposed method, as well as data analysis and discussion. Finally, the conclusion is given in section V.

II. Theoretical background

A. Isometric Mapping Algorithm

A single feature is not enough for an accurate judgement on gear's health state, so a variety of features are needed. Thereupon, dimensionality reduction is applied to fuse these features for obtaining a better health indicator. Isometric Mapping (ISOMAP) [31] algorithm is a nonlinear popular learning algorithm that maintains global properties. ISOMAP can preserve the monitoring information of the original features on the health of the gears as much as possible while reducing the consumption of computing resources by the neural network. Thus, ISOMAP is used for feature fusion.

B. Long short-term memory

The LSTM neural network is a special RNN neural network proposed by Hochreiter and Schmidhuber [21] in 1997 that can learn long-term dependencies. The LSTM neural network also consists of an input layer, a hidden layer, and an output layer, and the difference lies in using an LSTM structure that includes the input gate, the output gate, the forget gate, and the memory cell as the hidden layer, as shown in Fig. 1.

The forget gate is used to determine whether to keep the historical information stored in the current memory cell. If the door is opened, the historical information stored in the current memory cell is retained, otherwise the historical information is forgotten. The input gate is used to determine whether to allow the input layer information to enter the current memory cell. The open door allows the input layer signal to enter, while the closed door does not allow. The output gate is used to deter-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2959492, IEEE Transactions on Industrial Electronics

mine whether to output the current input layer signal to the next layer, the open door allows signal output while the closed door does not allow.

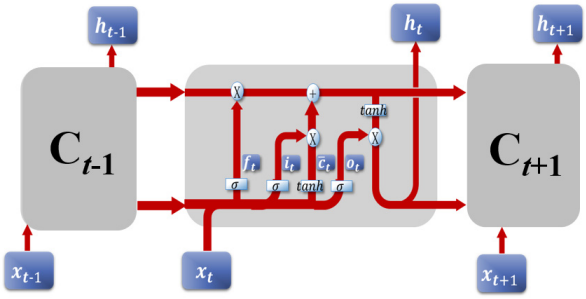

Fig. 1 Hidden layer structure of LSTM neural network

The LSTM network computes a mapping from an input sequence $\mathbf{x}=(x_1,x_2...x_n)$ to an output sequence $\mathbf{h}=(h_1,h_2...h_m)$ by iteratively calculating the network unit activations (t = 1,2,$\cdots$,T):

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{w}_{ix}\mathbf{x}_t + \mathbf{w}_{im}\mathbf{h}_{t-1} + \mathbf{b}_i\right)\\
\mathbf{f}_t &= \sigma\left(\mathbf{w}_{fx}\mathbf{x}_t + \mathbf{w}_{fm}\mathbf{h}_{t-1} + \mathbf{b}_f\right)\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh\left(\mathbf{w}_{cx}\mathbf{x}_t + \mathbf{w}_{cm}\mathbf{h}_{t-1} + \mathbf{b}_c\right)\\
\mathbf{o}_t &= \sigma\left(\mathbf{w}_{ox}\mathbf{x}_t + \mathbf{w}_{om}\mathbf{h}_{t-1} + \mathbf{b}_o\right)\\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)
\end{aligned} \tag{1}$$

where $\mathbf{i}$ is the input gate, $\mathbf{o}$ is the output gate, $\mathbf{f}$ is the forget gate, $\mathbf{c}$ includes cell activation vectors, and $\mathbf{h}$ is the memory cell outputs, $\mathbf{w}$ represents the weight matrix (for example, the weight $\mathbf{w}_{ix}$ matrix representing the input $\mathbf{x}$ to the input gate), $\mathbf{b}$ represents the threshold ($\mathbf{b}_i$ which is the threshold of the input gate), σ is the sigmoid activation function, tanh is the tanh activation function, $\odot$ represents dot product.

C. Attention mechanism

When neural network is used to process a large amount of input information, the attention mechanism of human brain can also be used for reference, and only some key information input is selected for processing, so as to improve the efficiency of neural network [32]. $\mathbf{x}=[x_1 \quad x_2 \quad \cdots \quad x_n]$ is used to represent n input data. In order to save computing resources, it is not necessary to input all n input information into the neural network for calculation. An available way is to select some information related to the task from $\mathbf{x}$ and input it into the neural network. Given a query vector q related to the task, we use the attention variable $z\in[1,n]$ to represent the index position of the selected information, that is, $z=i$ means the ith input information is selected. Given q and $\mathbf{x}$, the probability α_i for selecting the ith input data is calculated by.

$$\alpha_i = \frac{\exp\left(s\left(x_i,\mathrm{q}\right)\right)}{\sum_{j=1}^{n}\exp\left(s\left(x_j,\mathrm{q}\right)\right)} \tag{2}$$

where α_i represents attention coefficient, $s(x_i,q)$ represents the scoring function [33], which is defined as,

$$s\left(x_i,q\right) = \frac{x_i^T q}{\sqrt{d}} \tag{3}$$

where d is the dimension of input information.

In this paper, as the criterion of data contribution, α_i is used to amplify the weight of relevant data. Through the amplification of weights, the neural network will focus on learning the corresponding data .The amplification coefficient is set as η_i, which is defined as

$$\eta_i = 1+\alpha_i \tag{4}$$

and the details of its effect will be discussed in section III. B.

III. THE PROPOSED METHOD

Aiming at the limitations of traditional data-driven methods, a deep feature learning method combining MMA and LSTM is proposed to automatically predict RUL, which we call it MMALSTM. The flowchart of the proposed method is shown in Fig. 2.

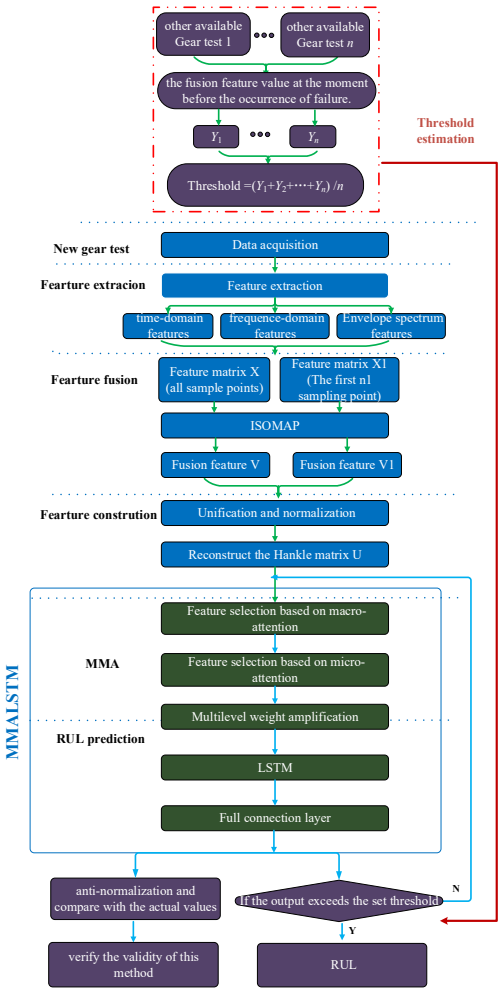

Fig. 2 The flowchart of the proposed gear remaining life prediction method.

The measured degradation signal contains abundant useful information. The time domain and frequency domain features of gear vibration signal can reflect the developing tendency of gear health status to some extent. In order to fully and accurately express the gear degradation process, the pro-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2959492, IEEE Transactions on Industrial Electronics

posed method will calculate all the features. In order to reduce the computational burden of the neural network, these high-dimensional features need to be reduced and fused. Then the fusion feature information after dimension reduction will be used for multi-step prediction of MMALSTM, and fusion features will be processed in a differentiated way, i.e. MMA. In this paper, considering the different amount of available information contained in different feature information, the proposed MMA is used to deal with the fusion feature data at macro level and micro level. Finally, according to the results of MMA, the weight of input information and recurrent information will be amplified, and then the fusion feature data will be processed automatically and differently. In the training process, we tune the parameters by Real-Time Recurrent Learning (RTRL) and minimize mean square error (MSE). After the model is trained, the prediction will continue until the prediction point exceeds the set threshold. The threshold value is the experimental parameter and can be selected according to the usage requirement of gears. In the paper, via several groups of gear tests under the same working condition, the threshold is estimated as the mean value of the fusion features at the moment before the occurrence of failure obtained by these experimental datasets.

A. Macroscopic-Microscopic Attention in LSTM

The contribution of different characteristics of gear life signal to the index of gear health is different，scholars construct or screen feature information with large contribution through various methods[34-36]. Based on the above ideas, this paper constructs high-value feature information by dimensionality reduction of 21 kinds of eigenvalues through popular algorithms. Different from the attention mechanism which screens the key information for input, the attention coefficient of eigenvalue matrix is evaluated at macro level and micro level in this paper. Then the input weight and recurrent weight of the input neural network are amplified in real time according to the macroscopic and microscopic attention coefficients, so that the neural network pays more attention to the data with large contribution. The structure of MMALSTM neural network is in Fig. 3. The number of input cells and the number of output cells in MMALSTM are respectively set as 20 and 1, and the learning rate is set as 0.1. The number of hidden layer cells is set as 14 in this study. The initialization method of neural network employs the standard initialization. And the recurrent neural network based on MMA is deduced as follows.

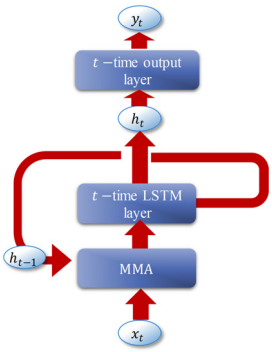

Fig. 3 structure of MMALSTM neural network

Firstly, we deal with the data matrix and calculate its macro and micro attention coefficient by the macro -micro attention mechanism. Suppose that the current whole input data is $\mathbf{X}_t = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_t \end{bmatrix}$. The input segment data at the time instant t' is defined as $\mathbf{x}_{t'} = \begin{bmatrix} x_{t',1} & x_{t',2} & \cdots & x_{t',n} \end{bmatrix}^T$ and the recurrent information at the time instant $t'-1$ is defined as $\mathbf{h}_{t'-1} = \begin{bmatrix} h_{t'-1,1} & h_{t'-1,2} & \cdots & h_{t'-1,m} \end{bmatrix}^T$. Macro attention mechanism (MA) is to process the data in the whole time interval by attention mechanism, resulting that LSTM can perform differentiation treatment on the input data $\mathbf{X}_t$ at the whole time scale, i.e. the improved LSTM has the ability of learning fusion features in a differentiated way according to the macroscopic importance of fusion features. The prediction result at the macro level can be represented as a query vector q_M. At each time instant t' ($t' = 1, 2, \cdots, t$), the mean value $\bar{x}_{t'}$ of input data $\mathbf{x}_{t'}$ is calculated, and its correlation with q_M is evaluated by the macro attention coefficient, which will be given in Eq. (5). According to the degree of similarity, the weights of all input segment data ($\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t$) at different time instants are amplified by different macro attention coefficients. This mechanism makes the neural network focus on the input data at important time instants and learn quickly. Micro attention mechanism is to process the data $\mathbf{x}_{t'}$ and $\mathbf{h}_{t'-1}$ at each time instant by attention mechanism, resulting that LSTM can perform differentiation treatment on the input data at the instant scale, i.e. the improved LSTM has the ability of learning fusion features in a differentiated way according to the microscopic importance of fusion features. At the time instant t', the microscopic prediction result can be represented as a query vector $q_{t',m}$. For the ith input data $x_{t',i}$, its correlation with $q_{t',m}$ is evaluated by a micro attention coefficient, which will be given in Eq. (6). For the ith recurrent data $h_{t'-1,i}$, its correlation with $q_{t',m}$ is evaluated by another micro attention coefficient, which will be given in Eq. (7). According to the degree of similarity, the weights of input data $x_{t',i}$ and recurrent data $h_{t'-1,i}$ are amplified by the two attention coefficients respectively. The micro attention mechanism allows the neural network to focus on the important input data and recurrent data in each segment data and learn quickly. Thus the operation of attention mechanism on data in the whole time dimension and each time dimension is called macro-micro attention mechanism.

In this paper, the macro -micro attention mechanism of input matrix and the micro attention mechanism of recurrent matrix are processed (The recurrent data in the whole time dimension is not known before input into the network, so it cannot be processed at the macro level). In the training process, the mean value $\bar{x}_{t+1}$ of $\mathbf{x}_{t+1}$ is set as query vector q_M at the macro level, $x_{t'+1,n}$ is set as query vector $q_{t',m}$ for the time instant t' at the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2959492, IEEE Transactions on Industrial Electronics

micro level. In the prediction process, $\bar{x}_t$ and $x_{t',n}$ represent q_M and $q_{t',m}$ respectively. And then the attention coefficients deduced from Eqs. (2) and (3) are calculated by:

$$\chi_{t'} = \frac{\exp\left(S\left(\bar{x}_{t'}, q_M\right)\right)}{\sum_{j=1}^{t}\exp\left(S\left(\bar{x}_j, q_M\right)\right)} \tag{5}$$

$$\alpha_{t',i} = \frac{\exp\left(s\left(x_{t',i}, q_{t',m}\right)\right)}{\sum_{j=1}^{n}\exp\left(s\left(x_{t',j}, q_{t',m}\right)\right)+\sum_{p=1}^{m}\exp\left(s\left(h_{t'-1,p}, q_{t',m}\right)\right)} \tag{6}$$

$$\lambda_{t',i} = \frac{\exp\left(s\left(h_{t',i}, q_{t',m}\right)\right)}{\sum_{j=1}^{n}\exp\left(s\left(x_{t',j}, q_{t',m}\right)\right)+\sum_{p=1}^{m}\exp\left(s\left(h_{t'-1,p}, q_{t',m}\right)\right)} \tag{7}$$

where $\chi_{t'}$ is macro attention coefficients corresponding to the whole input data; $\alpha_{t',i}$, $\lambda_{t',i}$ are micro attention coefficients corresponding to input data and recurrent data at the time instant t' respectively. And the three scoring functions are respectively defined as

$$S\left(\bar{x}_j, q_M\right) = \frac{\bar{x}_j\, q_M}{\sqrt{t}} \tag{8}$$

$$s\left(x_{t',j}, q_{t',m}\right) = \frac{x_{t',j}^T q_{t',m}}{\sqrt{n+m}} \tag{9}$$

$$s\left(h_{t',j}, q_{t',m}\right) = \frac{h_{t',j}^T q_{t',m}}{\sqrt{n+m}} \tag{10}$$

Secondly, according to the macro-micro attention coefficient and Eq. (4), the weight of input data is amplified in real time at multiple levels, while the weight of recursive data is amplified in real time based on the micro attention coefficient. To sum up, the weight matrixes used in the proposed LSTM at the time instant t' are given by

$$\begin{aligned}
\mathbf{w}'_{t',ix} &= (1+\chi_{t'})\times\mathbf{w}_{t',ix}\boldsymbol{\alpha}_{t'}\\
\mathbf{w}'_{t',ox} &= (1+\chi_{t'})\times\mathbf{w}_{t',ox}\boldsymbol{\alpha}_{t'}\\
\mathbf{w}'_{t',fx} &= (1+\chi_{t'})\times\mathbf{w}_{t',fx}\boldsymbol{\alpha}_{t'}\\
\mathbf{w}'_{t',ih} &= \mathbf{w}_{t',ih}\boldsymbol{\lambda}_{t'}\\
\mathbf{w}'_{t',oh} &= \mathbf{w}_{t',oh}\boldsymbol{\lambda}_{t'}\\
\mathbf{w}'_{t',fh} &= \mathbf{w}_{t',fh}\boldsymbol{\lambda}_{t'}
\end{aligned} \tag{11}$$

where

$$\boldsymbol{\alpha}_{t'} = \begin{bmatrix} 1+\alpha_{t',1}, 1+\alpha_{t',1}, \cdots, 1+\alpha_{t',1} \\ 1+\alpha_{t',2}, 1+\alpha_{t',2}, \cdots, 1+\alpha_{t',2} \\ \vdots \quad \vdots \quad \vdots \\ 1+\alpha_{t',n}, 1+\alpha_{t',n}, \cdots, 1+\alpha_{t',n} \end{bmatrix}_{n\times n},\quad \boldsymbol{\lambda}_{t'} = \begin{bmatrix} 1+\lambda_{t',1}, 1+\lambda_{t',1}, \cdots, 1+\lambda_{t',1} \\ 1+\lambda_{t',2}, 1+\lambda_{t',2}, \cdots, 1+\lambda_{t',2} \\ \vdots \quad \vdots \quad \vdots \\ 1+\lambda_{t',m}, 1+\lambda_{t',m}, \cdots, 1+\lambda_{t',m} \end{bmatrix}_{m\times m} \tag{12}$$

From the above, the flow chart of MMA, MA, ma are depicted in the Figs. 4 and 5, where Fig. 4(a) represents the flowchart of MMA, Fig. 4(b) represents the flowchart of MA, and Fig. 5 represents the flowchart of ma.

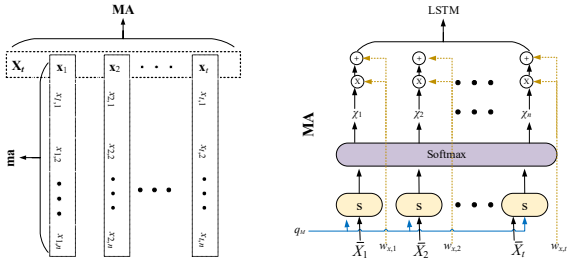

Fig. 4 (a)The flowchart of MMA. (b) The flowchart of MA.

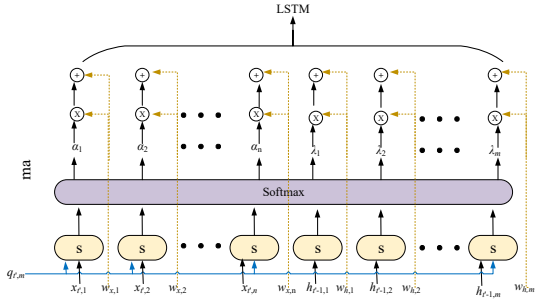

Fig. 5 The flowchart of ma.

With the weight amplified based on MMA, a new variant of LSTM is proposed, which is named as MMALSTM. Via Eq. (1), we can derive the calculation formula of MMALSTM as follows:

$$\begin{aligned}
\mathbf{i}_{t'} &= \sigma\left(\mathbf{w}'_{t',ix}\mathbf{x}_{t'} + \mathbf{w}'_{t',ih}\mathbf{h}_{t'-1} + \mathbf{b}_i\right)\\
\mathbf{f}_{t'} &= \sigma\left(\mathbf{w}'_{t',fx}\mathbf{x}_{t'} + \mathbf{w}'_{t',fh}\mathbf{h}_{t'-1} + \mathbf{b}_f\right)\\
\mathbf{c}_{t'} &= \mathbf{f}_{t'}\odot\mathbf{c}_{t'-1} + \mathbf{i}_{t'}\odot\tanh\left(\mathbf{w}_{t',cx}\mathbf{x}_{t'} + \mathbf{w}_{t',ch}\mathbf{h}_{t'-1} + \mathbf{b}_c\right)\\
\mathbf{o}_{t'} &= \sigma\left(\mathbf{w}'_{t',ox}\mathbf{x}_{t'} + \mathbf{w}'_{t',oh}\mathbf{h}_{t'-1} + \mathbf{b}_o\right)\\
\mathbf{h}_{t'} &= \mathbf{o}_{t'}\odot\tanh\left(\mathbf{c}_{t'}\right)\\
\mathbf{y}_{t'} &= g\left(\mathbf{w}_{t',yh}\mathbf{h}_{t'} + \mathbf{b}_y\right)
\end{aligned} \tag{13}$$

where g is the linear activation function.

B. Differentiated Learning of Weight

Differentiated learning ability of MMALSTM is reflected in the update of weights, and the weights related to data with high attention can be learned faster. MMALSTM is a special RNN structure with a more complex hidden layer. In MMALSTM, for any time instant t', the updating process of hidden layer weights is given below:

$$\frac{\partial\mathbf{y}_{t'}}{\partial\mathbf{w}_{t',x}} = \sum_{k=0}^{t'}\frac{\partial\mathbf{y}_{t'}}{\partial\mathbf{h}_{t'}}\frac{\partial\mathbf{h}_{t'}}{\partial\mathbf{c}_{t'}}\left(\prod_{j=k+1}^{t'}\tanh'(\mathbf{c}_{t'})\delta((1+\chi_{t'})\times\mathbf{w}_{t',fx}\boldsymbol{\alpha}_{t'}\mathbf{x}_{t'} + \mathbf{w}_{t',fh}\boldsymbol{\lambda}_{t'}\mathbf{h}_{t-1'} + \mathbf{b}_f)\right)\frac{\partial\mathbf{c}_k}{\partial\mathbf{w}_{t',x}}\boldsymbol{\alpha}_{t'}\times(1+\chi_{t'})$$

$$\frac{\partial\mathbf{y}_{t'}}{\partial\mathbf{w}_{t',h}} = \sum_{k=0}^{t'}\frac{\partial\mathbf{y}_{t'}}{\partial\mathbf{h}_{t'}}\frac{\partial\mathbf{h}_{t'}}{\partial\mathbf{c}_{t'}}\left(\prod_{j=k+1}^{t'}\tanh'(\mathbf{c}_{t'})\delta((1+\chi_{t'})\times\mathbf{w}_{t',fx}\boldsymbol{\alpha}_{t'}\mathbf{x}_{t'} + \mathbf{w}_{t',fh}\boldsymbol{\lambda}_{t'}\mathbf{h}_{t-1'} + \mathbf{b}_f)\right)\frac{\partial\mathbf{c}_k}{\partial\mathbf{w}_{t',h}}\boldsymbol{\lambda}_{t'}$$

$$\begin{bmatrix}\mathbf{w}_{t'+1,x}\\ \mathbf{w}_{t'+1,h}\end{bmatrix} = \begin{bmatrix}\mathbf{w}_{t',x}\\ \mathbf{w}_{t',h}\end{bmatrix} - \alpha\begin{bmatrix}\frac{\partial\mathbf{y}_{t'}}{\partial\mathbf{w}_{t',x}}\\ \frac{\partial\mathbf{y}_{t'}}{\partial\mathbf{w}_{t',h}}\end{bmatrix} \tag{14}$$

where $\mathbf{w}_{t',x}$ and $\mathbf{w}_{t',h}$ represent the input weights and the recurrent weights at the time instant t' respectively.

As δ is monotonically increasing and $(1+\chi_{t'})>1$, and so are $(1+\alpha_{t',i})$ and $(1+\lambda_{t',i})$, MMALSTM has a larger learning step size and a faster learning speed in each iteration compared

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2959492, IEEE Transactions on Industrial Electronics

with LSTM. To be specific, different weights can be learned in different ways according to the similarity of their prediction targets, and the weight of the data with high correlation converges faster, so that the neural network pays more attention to the data with large contribution. Namely, the weight of the data with greater attention coefficient converges in priority.

Moreover, when learning tends to be saturated, there is a larger weight gradient than LSTM, which can alleviate the learning saturation problem to a certain extent.

C. Prediction method based on MMALSTM and ISOMAP

In summary, the steps of the proposed method for predicting the gear remaining life based on the combination of the isometric mapping algorithm and MMALSTM neural network are given below:

1. During the gear life cycle, collect the gear vibration samples in specified sampling windows according to the interval T_s between two adjacent samples. Suppose that the number of sample is n.

2. Calculate the 21 kinds of time and frequency characteristics of these vibration samples separately, then the eigenvalue matrix $\mathbf{X}$ of the size $n \times 21$ can be obtained.

3. Select the eigenvalue matrix $\mathbf{X1}$ of the $n1$ samples from $\mathbf{X}$ as the training matrix.

4. The training matrix $\mathbf{X1}$ and original matrix $\mathbf{X}$ are respectively processed by the ISOMAP algorithm, and the obtained eigenvectors with the largest eigenvalues $\mathbf{V1} = \left(v1_1, v1_2, \ldots, v1_{n1}\right)^T$ and $\mathbf{V} = \left(v_1, v_2, \ldots, v_n\right)^T$ are used as their principal components respectively. The matrix $\mathbf{X}$ and the vector $\mathbf{V}$ are only used to verify the validity of this method, which don't participate in neural network training and prediction.

5. Since the size of the matrix $\mathbf{X}$ is larger than that of the matrix $\mathbf{X1}$ and both the sums of $\mathbf{V}$ and $\mathbf{V1}$ are equal to zero according to the characteristic of ISOMAP algorithm, the two vectors may have different starting value, even though their trends are same. Therefore, it is necessary to unify them. By minimizing

$$E = \sum_{i=1}^{n1} \left(v1_i - a v_i - b\right)^2 \tag{15}$$

a and b can be computed, and then all the elements of $\mathbf{V}$ are unified through the following equation

$$v_i^{'} = a v_i + b \tag{16}$$

6. Linearly normalize the vector $\mathbf{V1}$ to obtain the normalized vector $W = \left(w_1, w_2, \ldots, w_{n1}\right)^T$.

7. Reconstruct matrix $\mathbf{U}$:

$$\mathbf{U} = \begin{bmatrix} w_1 & w_2 & \cdots & w_{n1-p} \\ w_2 & w_3 & \cdots & w_{n1-p+1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p+1} & w_{p+2} & \cdots & w_{n1} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_{p+1} \end{bmatrix} \tag{17}$$

where p is the cell number of input layer and

$$\mathbf{u}_i = \begin{bmatrix} w_i & w_{i+1} \cdots & w_{n1-p+i-1} \end{bmatrix} \tag{18}$$

8. Set the first p vectors of the matrix $\mathbf{U}$ as the input of the MMALSTM neural network and the last vectors as the output respectively, then train the MMALSTM neural network.

9. With the trained MMALSTM neural network, the mapping function f for prediction is defined. By inputting the last p vectors of the matrix $\mathbf{U}$ into the trained MMALSTM, the output $\overline{u}_{p+2}$ at the first prediction time (FPT) can be calculated as

$$\overline{\mathbf{u}}_{p+2} = f\left(\mathbf{u}_2, \mathbf{u}_3 \cdots, \mathbf{u}_{p+1}\right) \tag{19}$$

Then $\mathbf{U}$ is updated by

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_{p+1} & \overline{\mathbf{u}}_{p+2} \end{bmatrix}^T \tag{20}$$

10. Repeating step 9 several times, we can obtain the predicted data series $\overline{u}_i$.

$$\begin{aligned} \overline{\mathbf{u}}_{p+3} &= f\left(\mathbf{u}_3, \mathbf{u}_4 \cdots, \mathbf{u}_{p+1}, \overline{\mathbf{u}}_{p+2}\right) \\ \overline{\mathbf{u}}_{p+4} &= f\left(\mathbf{u}_4, \mathbf{u}_5 \cdots, \mathbf{u}_{p+1}, \overline{\mathbf{u}}_{p+2}, \overline{\mathbf{u}}_{p+3}\right) \\ \overline{\mathbf{u}}_{p+5} &= f\left(\mathbf{u}_5, \mathbf{u}_6 \cdots, \mathbf{u}_{p+1}, \overline{\mathbf{u}}_{p+2}, \overline{\mathbf{u}}_{p+3}, \overline{\mathbf{u}}_{p+4}\right) \\ &\vdots \\ \overline{\mathbf{u}}_i &= f\left(\mathbf{u}_{i-p}, \mathbf{u}_{i-p+1} \cdots, \mathbf{u}_{p+1}, \overline{\mathbf{u}}_{p+2} \cdots \overline{\mathbf{u}}_{p+(i-p-2)}, \overline{\mathbf{u}}_{p+(i-p-1)}\right) \end{aligned} \tag{21}$$

When the predicted characteristic data $\overline{w}_{n1-p+i-1}$ ($\overline{w}_{n1-p+i-1} \in \overline{\mathbf{u}}_i$) exceeds a preset threshold, the gear remaining life can be calculated by $\left(i - n1\right) \times T_s$. Then we anti-normalize the predicted data series and get $\overline{v}_i$ ($i = n1+1, n1+2, \cdots, n$), which can be compared with the actual vector $\mathbf{V}^{'} = \left(v_{n1+1}^{'}, v_{p+2}^{'}, \ldots, v_n^{'}\right)^T$ to verify the validity of this method.

IV. Experimental Verification

A. Experimental Description

The gear life test was performed by a gear contact fatigue testing machine. As shown in Fig. 6, the testing machine is composed of a torque controller, a cooling and lubrication controller, an experimental operating platform and a gear operating platform. The experimental gearbox is mainly composed of two stages of gear transmissions, whose structure is shown in Fig. 7. Shaft 1 is the driving shaft and shaft 2, 3 are the driven shafts. The teeth numbers of gears 1, 2, 3, 4 are respectively 31, 25, 25 and 31. It is easy to note that its input speed is equal to the output speed.

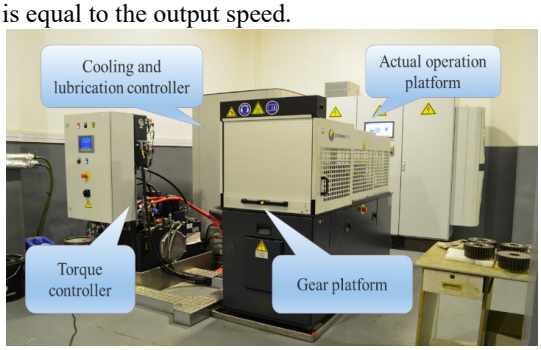

Fig. 6 Gear contact fatigue testing machine.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2959492, IEEE Transactions on Industrial Electronics

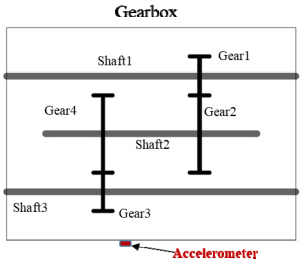

Fig. 7 Structure of gearbox in the gear contact fatigue testing machine.

In this experiment, the gear material was 40Cr; the module of gear was 5; the oil flow in the experimental gear box was 4L/h; and the cooling temperature was 70 degrees centigrade. The vibration signals were acquired by an accelerometer placed on the case of the experimental gearbox. The sampling frequency was set as 50000 Hz. Each sample had a duration of 20 s, which meant each sample had 1000000 points. The record interval was 40 s. The experiment stopped when the amplitude of the collected sample surpassed a certain level. There are totally 4 run-to-failure data sets with two different operating conditions as shown in Table I. Moreover, Most of the samples were acquired under the steady stage, and their characteristics were almost the same, which were of little significance to use these data for prediction. Therefore, only the last 400 samples were used for gear life prediction.

TABLE I
SPECIFICATIONS OF THE EXPERIMENT DATA

	Data set 1	Data set 2	Data set 3	Data set 4
Load(N)	1400	1400	1300	1300
Speed(rpm)	500	500	1000	1000
Experimental time(min)	814	820	789	796
Number of sample points	1221	1230	1183	1193

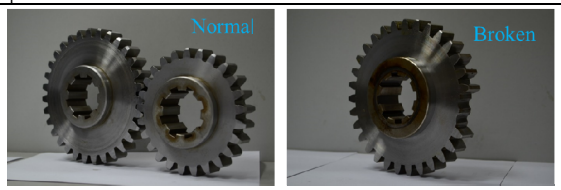

Fig. 8 Gears before experiment and after experiment in operation condition 1.

B. RUL Prediction by the Proposed Method

With the dataset, 21 time-domain and frequency-domain features can be calculated, and then they are applied to RUL prediction. In order to reduce the computational burden of the neural network, these high-dimensional features need be fused to form a health indicator. Except for the neural network, manifold learning is a frequently-used method for dimension reduction. The feature fusion results obtained by Principal Component Analysis (PCA), multidimensional scaling (MDS) and ISOMAP algorithms, Laplacian Eigenmaps (LE), Locally Linear Embedding (LLE) and Local Tangent Space Alignment (LTSA) are illustrated in Fig. 9. It can be seen from this figure that only PCA, ISOMAP and MDS are consistent with the degradation trend of gear life from normal to damage. However, among the three methods, the contribution rate c [37] of fusion feature values after ISOMAP and PCA procession is the largest, as shown in table II, which contains the most high-dimensional information. Compared with PCA, ISOMAP is a type of nonlinear manifold learning. To better characterize the nonlinearity, ISOMAP is used for generating the health indicator.

Dimension reduction is not our main research content, and we focus on the improvement of neural network, thus we just select one of these methods for pre-processing.

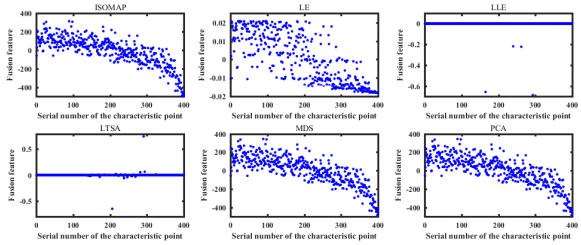

Fig. 9 Comparison of different manifold learning methods.

TABLE II
COMPARISON DIAGRAM OF CONTRIBUTION

Method	c
MDS	0.9797
PCA	1
ISOMAP	1

The characteristic matrix of 380 experimental samples are used as the training matrix to predict the data trend of the next 20 time instants. With the proposed method, the training, predictive and actual curves are respectively illustrated in Fig. 10. It can be seen from this figure that the predicted value is very close to the actual value and their trends are similar.

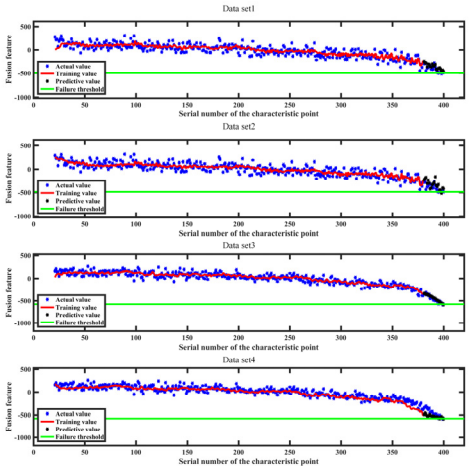

Fig. 10 Failure threshold, training curve, predicted curve and actual curve for 380 experimental samples of real gears.

The prediction ability was tested at the different given number of sampling points of data set 1. Each group was tested for 20 times, and the mean percentage error of predicted RUL was calculated, as shown in the Fig. 11.

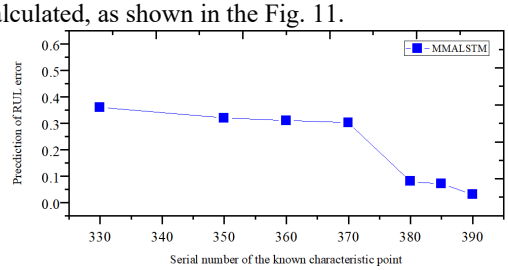

Fig. 11 The obtained prediction errors under the various numbers of the known characteristic point of data set 1.

It can be seen from Fig. 11 that the closer the prediction point is to the failure point, the more accurate the prediction of gear remaining life will be. The prediction ability of MMALSTM has been greatly improved, with the increase of the known characteristic points, reaching 97% accuracy when the number of characteristic points is 390.

To comprehensively assess the performance of the RUL estimation approach, given 380 known characteristic points, each dataset has been tested for 5 times, and a score function in the IEEE PHM 2012 challenge [38] is used

$$Score = \frac{1}{20}\sum_{i=1}^{20} A_i \tag{22}$$

where

$$A_i = \begin{cases} \exp(-\ln(0.5)\cdot(Er_i/5)), & Er_i \le 0 \\ \exp(+\ln(0.5)\cdot(Er_i/20)), & Er_i \ge 0 \end{cases} \tag{23}$$

and Er_i denotes the percent error for the i th experiment

$$Er_i\% = \frac{Rul_i - \overline{Rul_i}}{Rul_i} \times 100 \tag{24}$$

where Rul and $\overline{Rul}$ mean the actual RUL and the estimated RUL. In addition to the score function, Mean Absolute Deviation (MAE) and Root Mean Square Error (NRMSE) are also used for further comparison.

$$MAE = \frac{1}{20}\sum_{i=1}^{20} | Rul - \overline{Rul} | \tag{25}$$

$$NRMSE = \frac{\sqrt{\frac{1}{20}\sum_{i=1}^{20}\left(Rul - \overline{Rul}\right)^2}}{\frac{1}{20}\sum_{i=1}^{20}\overline{Rul}} \tag{26}$$

C. Comparison with Other Attention Mechanism

For comparison, four types of LSTMs based on attention mechanism are also applied. The first method called malstm relies on micro-attention to process both input data and recurrent data. The second method named as malstm1 relies on micro-attention to process only input data. The third method named as Mmalstm relies on macro-micro attention to process input data. The last method relies on macro-attention to process input data which we call it MAlstm. Then the next step are same as the method we proposed. The comparative results of real-time prediction obtained by the above methods are listed in Table III, where Test *a*, *b* represents the *a*th experiment of dataset *b*.

TABLE III
RUL ESTIMATION RESULTS AND COMPARISON I

	MMA LSTM	malstm	malstm1	Mmalstm	MAlstm
Test 1, 1	800s	720s	480s	480s	640s
Test 2, 1	800s	1080s	640s	640s	600s
Test 3, 1	720s	1080s	280s	760s	1480s
Test 4 ,1	720s	800s	440s	440s	440s
Test 5, 1	800s	1080s	800s	800s	640s
Test 1, 2	720s	1040s	1080s	920s	1040s
Test 2, 2	720s	1080s	920s	800s	1200s
Test 3, 2	1000s	800s	1400s	280s	1160s
Test 4, 2	1080s	800s	1200s	760s	800s
Test 5, 2	800s	1080s	960s	720s	720s
Test 1, 3	680s	800s	920s	1040s	800s
Test 2, 3	800s	800s	1080s	800s	880s
Test 3, 3	600s	720s	720s	720s	920s
Test 4, 3	800s	920s	920s	800s	720s
Test 5, 3	800s	720s	920s	880s	680s
Test 1, 4	800s	800s	920s	800s	800s
Test 2, 4	800s	960s	720s	800s	880s
Test 3, 4	600s	800s	720s	960s	800s
Test 4, 4	720s	720s	1080s	960s	1080s
Test 5, 4	720s	720s	840s	720s	920s
Score	0.834	0.639	0.442	0.746	0.576
MAE	70	114	214	110	172
NRMSE	0.142	0.183	0.308	0.243	0.279

From Table III, we can see that the proposed method acquires the highest score and the lowest error among these methods, which indicates that the proposed method provides more accurate RUL estimation results. Different from that other methods only consider the input data or only magnify the weight of data at a single level, the proposed method amplifies the weight of data more comprehensively and deeply, which enables the neural network to make full use of the contribution of data to the prediction results, and focus on processing data with large contribution. Under this situation, the effort for model selection and parameter tuning is greatly reduced and the prediction accuracy is increased.

D. Comparison with Other LSTMs

MMALSTM is actually an improved LSTM neural network. To further illustrate the advantage of the proposed MMALSTM method, traditional LSTM and other LSTM-based prediction methods are introduced. These methods have the similar structures as the MMALSTM except the structure of hidden layer. The traditional LSTM utilizing the basic model structure without the multiscale layer is first compared. The second method is composed of two same LSTM structures stacked on top of each other, we called DLSTM. The third method is LSTM with a projection layer (LSTMP). The forth method is GRU. The comparative results of real-time prediction obtained by the above methods are listed in Table IV.

From Table IV, the proposed method achieves the best performance among all the methods. Because of multi-level differentiated data processing, MMALSTM has obvious advantage compared to traditional LSTMs. For comprehensive comparison, the estimation results of all methods in table IV and III are illustrated in Fig. 12.

For experimental datasets, the fusion features between known 300 points and 350 points show no degradation, namely, gears are in normal operation. In the absence of deterioration information, the predictive ability of above models is limited. The 50-100 steps ahead prediction of traditional LSTMs is poor, and the prediction curve has no obvious downward trend. In such case, RUL also lose efficacy. However, MMALSTM still has certain predictive ability, reaching 62% accuracy when the number of known characteristic points is 300, as shown in Fig. 13.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2959492, IEEE Transactions on Industrial Electronics

TABLE IV
RUL ESTIMATION RESULTS AND COMPARISON II

	MMA LSTM	LSTM	DLSTM	LSTMP	GRU
Test 1, 1	800s	560s	1120s	720s	800s
Test 2, 1	800s	560s	960s	720s	720s
Test 3, 1	720s	1840s	520s	720s	720s
Test 4 ,1	720s	1840s	880s	840s	720s
Test 5, 1	800s	560s	1520s	720s	720s
Test 1, 2	720s	720s	320s	840s	760s
Test 2, 2	720s	1080s	520s	840s	760.s
Test 3, 2	1000s	520s	560s	920s	720s
Test 4, 2	1080s	1360s	560s	1080s	1720s
Test 5, 2	800s	800s	1840s	880s	880s
Test 1, 3	680s	400s	280s	440s	280s
Test 2, 3	800s	1080s	720s	440s	360s
Test 3, 3	600s	400s	680s	480s	160s
Test 4, 3	800s	440s	520s	400s	200s
Test 5, 3	800s	2720s	280s	360s	240s
Test 1, 4	800s	2680s	320s	440s	280s
Test 2, 4	800s	1040s	520s	440s	360s
Test 3, 4	600s	400s	680s	400s	360s
Test 4, 4	720s	440s	720s	400s	280s
Test 5, 4	720s	960s	680s	520s	280s
Score	0.834	0.428	0.513	0.608	0.577
MAE	70	522	322	226	348
NRMSE	0.142	0.724	0.565	0.424	0.804

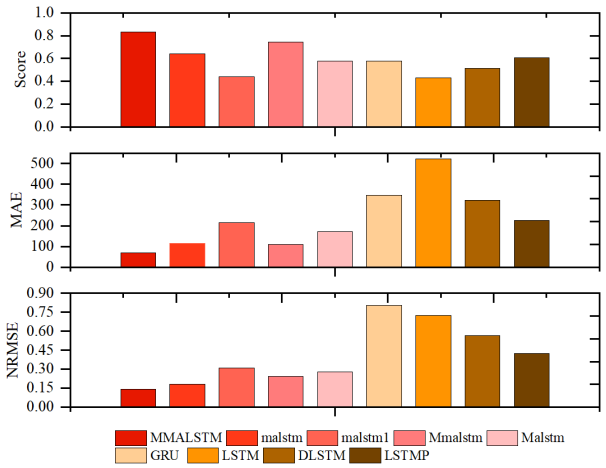

Fig. 12 Comparison chart of ability to predict RUL of different models.

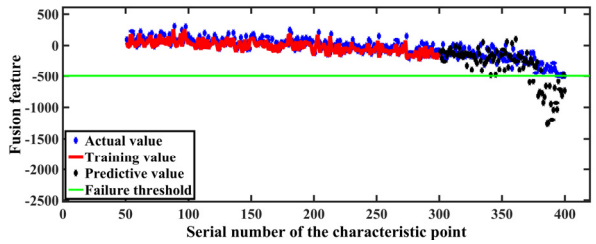

Fig. 13 Failure threshold, training curve, predicted curve and actual curve for 300 experimental samples of dataset 1.

Amplifying the data weight according to the data contribution based on attention mechanisms can improve the prediction ability of neural network. The prediction accuracy of malstm1 is only higher than that of LSTM. The prediction ability of Malstm is better than LSTM, DLSTM and GRU, whereas the GRU structure is only better than LSTM. Enlarging the range of weight amplification is helpful to improve the prediction performance. Obviously, MMALSTM is better than Mmalstm, and malstm is better than malstm1. Furthermore, enlarging the depth of weight amplification is helpful to improve the prediction performance. Mmalstm is better than Malstm and malstm1, whereas the deep LSTM structure is not as good as LSTM with the projection layer. From the above comparative results, it can be concluded that MMAlstm is indeed superior to traditional LSTMs and other attention-based LSTMs.

To show the superiority of the convergence performance of MMALSTM, the lifetime dataset 1 is used for comparison. The curves of prediction error versus iteration number obtained by various recurrent neural networks (LSTM, LSTMP, DLSTM, malstm, malstm1, Mmalstm, MAlstm, MMALSTM, GRU) are illustrated in Fig. 14. It can be easily observed from Fig.14 that MMALSTM is easier to get the local optimum due to MMA, so that its convergence is the fastest. Moreover, the differentiation treatment of relevant weights can indeed accelerate the learning of neural networks. MMALSTM and other lstms with attention mechanism are better than LSTM. Therefore, MMALSTM has the fastest convergence speed due to its multi-level differentiation treatment mechanism.

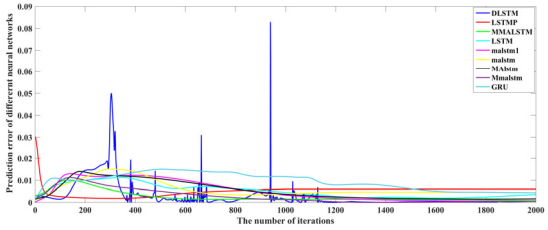

Fig. 14 Comparison of the convergence speed of different neural networks for dataset 1.

E. RUL Prediction of Bearings

The experimental data comes from PRONOSTIA in the IEEE PHM 2012 Data Challenge. The proposed method is applied to the data set bearing1-1.The run-to-failure bearing1-1 contains 2803 samples. In this case, five features (mean absolute difference，standard deviation and root-mean-square, mean of frequency distribution, envelope spectrum: mean of frequency distribution) are calculated and fused by ISAOMAP. The unknown data of the bearing 1-1 does not participate in model training, but only participates in the verification of the prediction results. The data set bearing1-1 is obtained under the constant operation condition 1. To obtain an accurate threshold, the mean value of fusion features of bearing1-2 and bearing1-3 with the same operation condition before the apparent amplitude jump is calculated as the failure thresh-old of bearing 1-1.

In this dataset, numerous samples were acquired during the steady stage, and their characteristics were almost the same, which were of little significance to use these data for prediction. Therefore, only the last 1750 samples were used for bearing life prediction. In this experiment, the number of input cells, hidden layer cells and output cells in MMALSTM are respectively set as 60, 17 and 1, and the learning rate is set as 0.05.

The characteristic matrix of 1600 experimental samples is used as the training matrix to predict the 150 characteristic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2959492, IEEE Transactions on Industrial Electronics

points corresponding to the next 150 time instants. Via the proposed method, the training, predictive and actual curves are illustrated in Fig. 15. We can see from this figure that the predicted value begins to exceed the threshold from the serial number of 1736. As the serial number of actual failure point is 1750, the prediction error *Er* is calculated as 9.3%. It then follows that the proposed MMALSTM can be also applied to predict the RUL of bearings effectively.

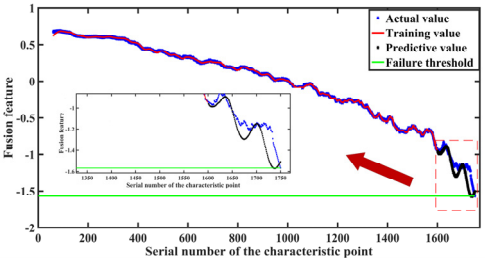

Fig. 15 Failure threshold, training curve, predicted curve and actual curve for 1600 experimental samples of real bearing.

V. Conclusions

This paper explored a new method for gear life prediction by feature infusion and MMALSTM. First, an isometric mapping algorithm is used to fuse 13 time-domain features, 4 frequency-domain features and 4 envelope spectrum features of gear vibration signal into a new feature. With the fused feature, a new type of LSTM based on MMA (MMALSTM) is then proposed to predict the gear remaining useful life. MMALSTM can improve the prediction accuracy of the LSTM, and it is the main contribution of this study. Via the experimental data, the performance of MMALSTM is verified and compared to GRU, traditional LSTMs and several attention-based LSTMs. The comparative results show that MMALSTM achieved the best prediction accuracy, and the proposed method is more suitable for gear RUL prediction. Besides, the proposed MMALSTM can be effectively applied to predict the RUL of bearings.

In the future, we will study the depth structure model of MMALSTM and its application into gear life prediction. Furthermore, how to predict the gear remaining useful life under variable working condition by the deep MMALSTM is also worth exploring.

Refrences

[1] T. Wang, Q. Han, F. Chu, and Z. Feng, "Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review," *Mech. Syst. Signal Proc.,* vol. 126, pp. 662-685, 2019.

[2] S. Liu, C. Song, C. Zhu *et al.*, "Investigation on the influence of work holding equipment errors on contact characteristics of face-hobbed hypoid gear," *Mech. Mach. Theory,* vol. 138, pp. 95-111, Aug, 2019.

[3] N. Saravanan, V. K. Siddabattuni, and K. Ramachandran, "A comparative study on classification of features by SVM and PSVM extracted using Morlet wavelet for fault diagnosis of spur bevel gear box," *Expert Syst. Appl.,* vol. 35, no. 3, pp. 1351-1366, 2008.

[4] Y. Wang, R. Markert, J. Xiang *et al.*, "Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system," *Mech. Syst. Signal Proc.,* vol. 60, pp. 243-251, 2015.

[5] S. Lu, Q. He, and J. Wang, "A review of stochastic resonance in rotating machine fault detection," *Mech. Syst. Signal Proc.,* vol. 116, pp. 230-260, 2019.

[6] Z. Chen, W. Zhai, and K. Wang, "Vibration feature evolution of locomotive with tooth root crack propagation of gear transmission system," *Mech. Syst. Signal Proc.,* vol. 115, pp. 29-44, Jan 15, 2019.

[7] C. Shen, Y. Qi, J. Wang *et al.*, "An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder," *Eng. Appl. Artif. Intell.,* vol. 76, pp. 170-184, 2018.

[8] Y. Qin, J. Zou, B. Tang *et al.*, "Transient feature extraction by the improved orthogonal matching pursuit and K-SVD algorithm with adaptive transient dictionary," *IEEE Trans. Ind. Inform.*, 2019.

[9] S. Lu, R. Yan, Y. Liu *et al.*, "Tacholess speed estimation in order tracking: A review with application to rotating machine fault diagnosis," *IEEE Trans. Instrum. Meas.*, 2019.

[10] Z. Feng, X. Chen, and M. J. Zuo, "Induction Motor Stator Current AM-FM Model and Demodulation Analysis for Planetary Gearbox Fault Diagnosis," *IEEE Trans. Ind. Inform.,* vol. 15, no. 4, pp. 2386-2394, 2018.

[11] D. Wang, Q. Miao, Q. Zhou *et al.*, "An intelligent prognostic system for gear performance degradation assessment and remaining useful life estimation," *J. Vib. Acoust.-Trans. ASME,* vol. 137, no. 2, pp. 021004, 2015.

[12] M. Yuan, Y. Wu, and L. Lin, "Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network." pp. 135-140.

[13] R. Zhao, R. Yan, Z. Chen *et al.*, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Proc.,* vol. 115, pp. 213-237, 2019.

[14] Y. Qin, X. Wang, and J. Zou, "The optimized deep belief networks with improved logistic Sigmoid units and their application in fault diagnosis for planetary gearboxes of wind turbines," *IEEE Trans. Ind. Electron.,* vol. 66, no. 5, pp. 3814-3824, 2018.

[15] Y. Zeng, and Z. Zhu, "Study on Rear Axle Gear Residual Life Predication based on RBF Network," *Acta Oto-Laryngol.*, 2012.

[16] J. Zhu, N. Chen, and W. Peng, "Estimation of bearing remaining useful life based on multiscale convolutional neural network," *IEEE Trans. Ind. Electron.,* vol. 66, no. 4, pp. 3208-3216, 2018.

[17] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Reliab. Eng. Syst. Saf.,* vol. 172, pp. 1-11, 2018.

[18] A. Zhang, H. Wang, S. Li *et al.*, "Transfer Learning with Deep Recurrent Neural Networks for Remaining Useful Life Estimation," *Applied Sciences,* vol. 8, no. 12, pp. 2416, 2018.

[19] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks." pp. 6645-6649.

[20] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 5, no. 2, pp. 157-166, 1994.

[21] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural Comput.,* vol. 9, no. 8, pp. 1735-1780, 1997.

[22] M. Yuan, Y. Wu, and L. Lin, "Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network." pp. 135-140.

[23] J. Wang, B. Peng, and X. Zhang, "Using a stacked residual LSTM model for sentiment intensity prediction," *Neurocomputing*, vol. 322, pp. 93-101, 2018.

[24] K. B. Charbonneau, and O. Shouno, "Neural Trajectory Analysis of Recurrent Neural Network In Handwriting Synthesis," *arXiv preprint arXiv:1804.04890*, 2018.

[25] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing,* vol. 284, pp. 171-176, 2018.

[26] P. Anderson, X. He, C. Buehler *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering." pp. 6077-6086.

[27] L. Wu, Y. Wang, X. Li *et al.*, "Deep attention-based spatially recursive networks for fine-grained visual recognition," *IEEE T. Cybern.,* vol. 49, no. 5, pp. 1791-1802, 2018.

[28] X. Ran, Z. Shan, Y. Fang *et al.*, "An LSTM-Based Method with Attention Mechanism for Travel Time Prediction," *Sensors,* vol. 19, no. 4, pp. 861, 2019.

[29] T. Fernando, S. Denman, S. Sridharan *et al.*, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks,* vol. 108, pp. 466-478, 2018.

[30] Y. Cui, J. Shi, and Z. Wang, "Quantum assimilation-based state-of-health assessment and remaining useful life estimation for electronic systems,"

IEEE Trans. Ind. Electron., vol. 63, no. 4, pp. 2379-2390, 2015.
[31] A. Najafi, A. Joudaki, and E. Fatemizadeh, "Nonlinear dimensionality reduction via path-based isometric mapping," *IEEE Trans. Pattern Anal. Mach Intell,* vol. 38, no. 7, pp. 1452-1464, 2015.
[32] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention." pp. 2204-2212.
[33] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.
[34] L. Ren, J. Cui, Y. Sun *et al.*, "Multi-bearing remaining useful life collaborative prediction: A deep learning approach," *J Manuf Syst.,* vol. 43, pp. 248-256, 2017.
[35] J. Zhang, P. Wang, R. Yan *et al.*, "Long short-term memory for machine remaining life prediction," *J Manuf Syst.,* vol. 48, pp. 78-86, 2018.
[36] L. Guo, N. Li, F. Jia *et al.*, "A recurrent neural network based health indicator for remaining useful life prediction of bearings," *Neurocomputing,* vol. 240, pp. 98-109, 2017.
[37] P. Nectoux, R. Gouriveau, K. Medjaher *et al.*, "PRONOSTIA: An experimental platform for bearings accelerated degradation tests." pp. 1-8.
[38] J. Zhu, N. Chen, and W. Peng, "Estimation of bearing remaining useful life based on multiscale convolutional neural network," *IEEE Trans. Ind. Electron.,* vol. 66, no. 4, pp. 3208-3216, 2018.

Yi Qin received the B.Eng. and Ph.D. degrees in mechanical engineering from Chongqing University, Chongqing, China, in 2004 and 2008 respectively.

Since January 2009, he has been with the Chongqing University, Chongqing, China, where he is currently a Professor in the College of Mechanical Engineering. His current research interests include signal processing, fault prognosis, mechanical dynamics and smart structure.

Sheng Xiang received the B.Eng. degree in mechanical engineering from Yangtze University, Hubei, China, in 2017.

He is currently working toward the Ph.D. degree in mechanical engineering of Chongqing University, Chongqing, China.

His research interests mainly include signal processing, mechanical fault diagnosis and remaining useful life prediction.

Yi Chai received the MA.SC and Ph.D. degrees in control Theory and control engineering from Chongqing University, Chongqing, China, in 1994 and 2001 respectively.

Since September 2003, he has been the Professor of College of Automation in the Chongqing University. His current research interests include safety assessment, fault diagnosis and health prognosis measurement, intelligent optimization and autonomous control.

Haizhou Chen received the MA.Eng and Ph.D. degrees in mechanical engineering from Chongqing University, Chongqing, China, in 2010 and 2017 respectively.

Since July 2017, he has been with the Qingdao University of Science and Technology, Qingdao, China, where he is currently a Lecturer in the College of Electromechanical Engineering. His current research interests include failure mechanism analysis and fault prognosis.