*Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)*
*Edited by* Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P107-cd

# A multi-scale LSTM with multi-head self-attention embedding mechanism for the remaining useful life prediction of hot strip mill rollers

Ting Zhu

*State Key Laboratory of Mechanical System and Vibration, Department of Industrial Engineering & Management, Shanghai Jiao Tong University, China. E-mail: accounter@sjtu.edu.cn*

Zhen Chen

*State Key Laboratory of Mechanical System and Vibration, Department of Industrial Engineering & Management, Shanghai Jiao Tong University, China. E-mail: chenzhendr@sjtu.edu.cn*

Di Zhou

*College of Mechanical Engineering, Donghua University, China. E-mail: zhoudi@dhu.edu.cn*

Ershun Pan

*State Key Laboratory of Mechanical System and Vibration, Department of Industrial Engineering & Management, Shanghai Jiao Tong University, China. E-mail: pes@sjtu.edu.cn*

Remaining useful life(RUL) prediction of intelligent equipment plays a crucial role in avoiding major safety accidents and substantial economic losses from degradation failures. Recently, many studies focused on deep learning-based data-driven methods, such as long short-term memory (LSTM) neural networks, which used multi-dimensions monitoring signals or features to predict the RULs. However, most existing methods are inability to acquire valid temporal information from long-term time series. Moreover, the input data containing much redundant information leads to imprecise RUL prediction results. To overcome the aforementioned weakness, a multi-scale LSTM neural network with multi-head self-attention embedding mechanism(MLSTM-MHA) is proposed in this article for RUL prediction. Firstly, the memory cells of LSTM are divided into several parts according to different temporal trend types, such as local trends, medium trends, and long trends. Fusing all types of memory cells can capture additional trend information and improve the performance of LSTM in learning time series. Secondly, the multi-head self-attention mechanism is embedded in the forgetting gate and input gate structure of LSTM, which can participate in training the MLSTM-MHA network and adaptively recalculates the network weights. The redundant information is assigned lower weights due to lower values by the attention module. Finally, a hot strip mill roller dataset is used to validate the superiority of the proposed method. Compared with the existing data-driven RUL prediction methods, the proposed method has a more accurate predictive ability.

*Keywords*: Remaining useful life prediction, LSTM, self-attention mechanism, redundant information.

## 1. Introduction

Prognostics and health management (PHM) have received much attention in the field of industrial applications, which is significant for avoiding economic losses from degradation failures and ensuring stable operation of machinery. Remaining useful life (RUL) prediction is always regarded as the most challenging and meaningful task in PHM because deep characteristics of equipment are difficult to excavate from multi-scale degradation data (An 2015). At present, data-driven RUL prediction approaches that use condition monitoring data to estimate equipment health state have become a significant focus.

Generally, data-driven prognostic methods can usually be divided into two types: stochastic process-based methods and machine learning-based methods (Li 2019). For the stochastic process-based RUL prediction methods, Zhang et al. (2018) summarized the current approaches to the Wiener process-based degradation data

analysis and RUL prediction. Wu et al. (2022) proposed an adaptive nonlinear Wiener process model with the degradation drift satisfying the closed skew-normal distribution to estimate bearing RULs. Ling et al. (2019) considered a two-phase degradation model with Gamma process, and a likelihood inference approach with Bayesian theory was proposed to predict the RULs of LEDs. The stochastic process-based methods can describe the random characteristics of equipment degradation process. Nevertheless, complex parameter estimation procedure of stochastic process-based approaches affects their efficiency in online RUL prediction.

Instead, machine learning-based methods can learn the degradation characteristics of machinery from massive historical monitoring data and thereby predict the RULs automatically (Krot 2020). The potential information in the monitoring data can be mined by the deep structures of neural networks such as convolutional neural networks (CNN) and long short-term memory (LSTM). Yang et al. (2019) proposed an RUL prediction method based on a double-CNN model architecture that can intelligently extract critical features from original vibration signals. Chen et al. (2019) presented a general two-step solution for RUL prediction based on the nonlinear degradation process with KPCA and gate recurrent unit (GRU). Ma and Mao. (2020) predicted the ball bearing RULs by a LSTM network integrated with convolutional operation. Han et al. (2021) combined stacked autoencoder and recurrent neural network (RNN) to obtain health indicators without human interference, which was validated by a public bearing faults dataset. Liu et al. (2021) integrated the advantages of multi-stage LSTM model and clustering analysis, which improved the prediction accuracy of RULs of the aero-engine. These RNN variants such as GRU and LSTM can remember a part of temporal information to improve prediction accuracy (Elsheikh 2019). However, they can not retain the global temporal information in a long term and renew the local temporal information in time. And the contributions of temporal information at different times has not received attention.

To break the bottleneck of existing RNN-based RUL prediction methods, this paper proposed a multi-scale LSTM neural network with multi-head self-attention embedding mechanism (MLSTM-MHA) to extract different temporal degradation information from raw monitoring data. First, three multi-head self-attention modules extract significant temporal information after the forget gate and input gate. Second, improved memory cells can capture various time-scale information by setting three memory sub-cell units. Finally, the comparative results with some state-of-art methods prove that the proposed method can predict RUL more accurately.

The rest of this article is organized as follow. Section 2 presents a brief introduction of basic theoretical background. Then, the proposed method and the detail construction process are elaborated in Section 3. Next, the experimental implementation and comparison results are introduced in Section 4. Finally, the conclusion and future work is given in Section 5.

## 2. Basic theoretical background

The functions of self-attention mechanism prompt the deep network to learn the concealed knowledge in time-series data closely related to the equipment degradation process. In practice, given the same set of queries, keys, and values, the model is expected to learn different behaviours based on the exact self-attention mechanism and then combine the other behaviours as knowledge to capture the various ranges of dependencies within the sequence. The structure of multi-head self-attention module is shown in Fig. 1.

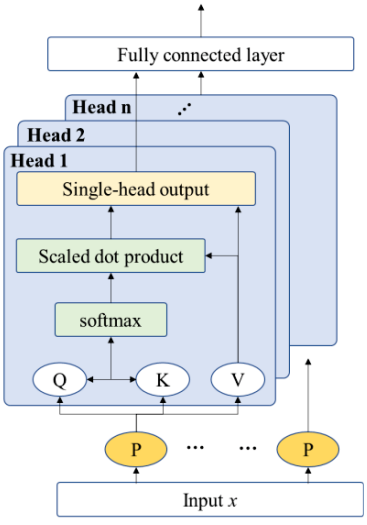


Fig. 1. Structure of multi-head self-attention module.

Since several self-attentions layers are stacked in parallel to form multi-head attention, the idea is to obtain multiple queries by linearly transforming the query using different weight matrices. Each newly formed query essentially requires a different type of relevant information, thus allowing the attention model to introduce more information into the computation.

## 3. MLSTM-MHA network

The MLSTM-MHA network is composed of the multi-head self-attention unit and multi-scale memory cell.

The forget gate and the input gate in the LSTM network are responsible for selective forgetting and remembering of the hidden layer input and the current input, respectively. Both the forgetting ability to forget gate and the retention ability to input gate use the *sigmoid* function to adjust the hidden input and current input to between 0 and 1, and then decide which parts are input to the memory cell by:

$$i_t = \sigma(u_{hi} h_{t-1} + w_{xi} x_t + b_i) \tag{1}$$

$$\mathrm{f}_t = \sigma(u_{h\mathrm{f}} h_{t-1} + w_{x\mathrm{f}} x_t + b_{\mathrm{f}}) \tag{2}$$

where $u = \{u_{h\mathrm{f}}, u_{hi}\}$ and $w = \{w_{x\mathrm{f}}, w_{xi}\}$ denote the weight matrixes of current input data $\boldsymbol{x_t}$ and hidden input $\boldsymbol{h_{t-1}}$, respectively; $b = \{b_{\mathrm{f}}, b_i\}$ denote the bias vectors; and $\sigma$ represents sigmoid function.

However, in a time series of degradation data, the trend data at each time point possesses different importance and contains both valuable and redundant information. They provide different degrees of contribution to the predicted results of the equipment's remaining useful life. Therefore, valuable information should be retained and redundant information should be eliminated in the forget gate and input gate according to the importance of the whole degradation sequence. Considering the good properties of the multi-head self-attention mechanism for time series learning, this paper proposed an LSTM-MHA network that embeds it into forget gate and input gate to achieve self-adaptive selection of the output. In each single-head self-attention module, the input $\boldsymbol{x}$ is processed initially in the embedding layer to acquire feature vectors $\boldsymbol{P}$ with $d_P$ dimension. Then, the query $\boldsymbol{Q_i}$, the key $\boldsymbol{K_i}$, and the value $\boldsymbol{V_i}$ can be obtained as follows:

$$Q_i = W_Q^i P,\ K_i = W_K^i P,\ V_i = W_V^i P \tag{3}$$

where $W_Q^i \in R^{d_Q \times d_P}$, $W_K^i \in R^{d_K \times d_P}$, and $W_V^i \in R^{d_V \times d_P}$ are the three linear projection matrices.

The three vectors $\boldsymbol{Q_i}$, $\boldsymbol{K_i}$, and $\boldsymbol{V_i}$ are then put into the scaled dot product calculation part to calculate the attention weights and a reconstructed self-attention matrix, whose implementation process is as follows:

$$\begin{cases} head_i = Attention(Q,K,V) = \mathrm{softmax}\left(\dfrac{QK^T}{\sqrt{d_k}}\right)V \\ \mathrm{softmax} = \dfrac{e^{z_i^T o_i}}{\sum_j e^{z_j^T o_i}} \end{cases} \tag{4}$$

where $d_k$ is the dimension of $\boldsymbol{K}$, and T denotes the transpose operation. In the end, all single-head self-attention modules are concatenated together to obtain the multi-head features based on (5).

$$multihead = concat\left(head_1, \cdots, head_h\right) W_P \tag{5}$$

where concat(·) denotes the splicing operation on these single-head self-attention modules and then multiplies by the linear transformation matrix $\boldsymbol{W_p}$. The resulting output of multi-head self-attention module has the same scale as the input $\boldsymbol{x}$.

Therefore, the formulas (4)-(5) have been updated as follows:

$$i_t = multihead\left(\sigma(u_{hi} h_{t-1} + w_{xi} x_t + b_i)\right) \tag{6}$$

$$\mathrm{f}_t = multihead\left(\sigma(u_{h\mathrm{f}} h_{t-1} + w_{x\mathrm{f}} x_t + b_{\mathrm{f}})\right) \tag{7}$$

Because of the embedding structure, the multi-head self-attention module can participate in the training of the MLSTM-MHA network and self-adaptively recalculates the weights as the network parameters are updated.

The degradation trend of the equipment performance is a long-time process, and it can not only contain a local trend related to short-term fluctuations but also a global trend related to long-term degradation. Therefore, different time series data of equipment degradation data can contain additional trend information. To acquire a better

prediction model in RUL prediction work, the model needs to record the overall change in equipment life using global trend information and update the current fluctuations using local trend information.

Based on improved forget gate and input gate, a multi-scale memory cell containing three memory sub-cell units $c_s$, $c_m$, and $c_l$ is proposed. With the updated results of three sub-cell units, the multi-scale memory cell is renewed by local, medium, and global information, and the updating rule elaborates as follows.

The local trend in equipment degradation represents the effect of changes in current work conditions on equipment performance. Therefore, the local memory sub-cell units need to capture the fluctuations of temporal degradation information rapidly, and the formula is defined as follows:

$$c_s = \bar{c}_t = multihead\left(\tanh\left(u_{hc}h_{t-1} + w_{xc}x_t + b_c\right)\right) \quad (8)$$

where $\boldsymbol{u_{hc}}$ and $\boldsymbol{w_{xc}}$ denote the weight matrixes of current input data $\boldsymbol{x_t}$ and hidden input $\boldsymbol{h_{t-1}}$, respectively; $\boldsymbol{b_c}$ denote the bias vector; and tanh(·) is the activation function.

The global trend majorly obtains long-term temporal information. Thus, the global memory sub-cell units should always be delivered in the MLSTM-MHA network, and the formula is written as follows:

$$c_l = c_{t-1} \quad (9)$$

where $c_{t-1}$ denotes the cell state at the last moment. It means that the global memory sub-cell units selectively contain all previous temporal degradation information at the current time point.

The medium trend is between the local trend and the global trend. It required a balance between the retention of historical information and the volatility of current trends. he structure of the medium memory sub-cell units is similar to the structure of memory cell in traditional LSTM, and the updated formula is defined as follow:

$$\begin{aligned} c_m &= \mathrm{f}_t \odot c_l + i_t \odot c_s \\ &= multihead\left(\sigma(u_{hf}h_{t-1} + w_{xf}x_t + b_f)\right) \odot c_{t-1} \\ &+ multihead\left(\sigma(u_{hi}h_{t-1} + w_{xi}x_t + b_i)\right) \\ &\odot multihead\left(\tanh\left(u_{hc}h_{t-1} + w_{xc}x_t + b_c\right)\right) \end{aligned} \quad (10)$$

Combine all the above operations, a novel multi-scale LSTM neural network with multi-head self-attention embedding mechanism is constructed, and its topological structure is illustrated in Fig. 2.

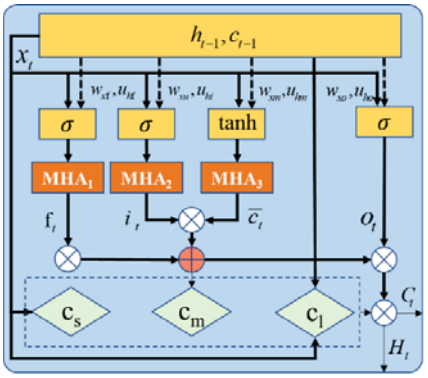


Fig. 2. Neuron structure of the MLSTM-MHA

The iterative formulas of MLSTM-MHA are derived as follows:

$$\begin{aligned} i_t &= multihead\left(\sigma(u_{hi}h_{t-1} + w_{xi}x_t + b_i)\right) \\ \mathrm{f}_t &= multihead\left(\sigma(u_{hf}h_{t-1} + w_{xf}x_t + b_f)\right) \\ o_t &= \sigma(u_{ho}h_{t-1} + w_{xo}x_t + b_o) \\ \bar{c}_t &= multihead\left(\tanh\left(u_{hc}h_{t-1} + w_{xc}x_t + b_c\right)\right) \\ c_t &= \left[c_s, c_m, c_l\right]^T \\ &= \begin{bmatrix} multihead\left(\tanh\left(u_{hc}h_{t-1} + w_{xc}x_t + b_c\right)\right) \\ \left(\mathrm{f}_t \odot c_{t-1} + i_t \odot multihead\right. \\ \left.\left(\tanh\left(u_{hc}h_{t-1} + w_{xc}x_t + b_c\right)\right)\right) \\ c_{t-1} \end{bmatrix} \\ h_t &= o_t \odot \tanh\left(c_t\right) \end{aligned} \quad (11)$$

where $o_t$ is the output gate; $\boldsymbol{u_{ho}}$, $\boldsymbol{w_{xo}}$, and $\boldsymbol{b_o}$ denote the two weight matrixes and bias vector, respectively; $\bar{c}_t$ is the candidate memory cell; and $\boldsymbol{h_t}$ is the hidden output of MLSTM-MHA.

## 4. Experimental results and discussion

### 4.1. *Dataset description and prerecession*

The industrial hot strip mill roller dataset by Baoshan Iron & Steel Co., Ltd. is composed of several subsets which has shown in Fig.3. Since

the roller inspection is mainly arranged after the processing of relevant products in the process of rolling steel, the run batch is used to characterize the rolling time information (Jiao 2021).

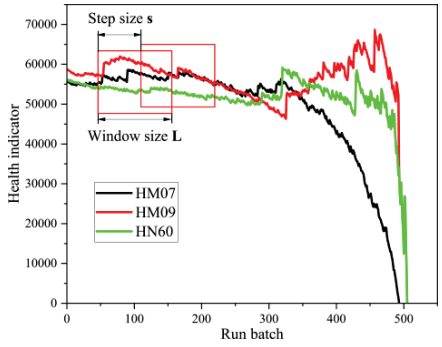


Fig. 3. The several subsets from industrial hot strip mill rollers dataset.

This dataset mainly records roller operation and maintenance information from 1580 hot strip finish mill equipment F1 to equipment F7. This paper chooses HM07, HM09 and HN60 to validate the superiority of the proposed method considering the integrity of data. As shown in Table I, these three subsets such as HM07 include condition monitoring data closely related to rollers performance, such as up and down time, steel production, rollers' diameter, operation mileage and other vital information between December 2017 and August 2020. In this paper, rollers' health indicators (HI) constructed by their diameter and remaining mileage are used to describe rollers' degradation process.

Table I. Implement details of the MLSTM-MHA

| Roller Num | Up time | ⋯ | Diameter(mm) |
|---|---|---|---|
| HM07 | 2017/12/1 1:06:00 | ⋯ | 824.94 |
| HM07 | 2017/12/1 5:04:00 | ⋯ | 824.67 |
| ⋯ | ⋯ | ⋯ | ⋯ |
| HM07 | 2020/08/31 23:54:02 | ⋯ | 739.4 |

And a sliding window with window size $L$ and step size $s$ is adopted to segment data. For the train and valid data, a sliding window moves with a step size $s$ from the life-start point to the life-end point. This paper chooses the same window size $L$ and step size $s$ considering the data consistency. The rollers' RUL are specified as the time remaining between the current time and the time when the health indicator first reaches the failure threshold, i.e.,

$$RUL(t) = Ti\{t \,|\, HI(t) = \lambda, t > 0\} - t_0 \quad (12)$$

where $\lambda$ denotes the failure threshold; $t_0$ indicates the current time.

The min-max standardization is performed for the measured values to get health indicators of the HM07 dataset, as shown in (12):

$$HI_k^*(t) = \frac{x_k(t) - \min\{x_k\}}{\max\{x_k\} - \min\{x_k\}} \quad (13)$$

where $HI_k(t)$ is the health indicator collected by the $\boldsymbol{k}$-th roller at time point $\boldsymbol{t}$.

In order to compare the performance of the proposed MLSTM-MHA method with previous methods, several evaluation indexes such as root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination ($R^2$) are employed, i.e.,

$$RMSE = \sqrt{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \Big/ N} \quad (14)$$

$$MAE = \sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right| \Big/ N \quad (15)$$

$$R^2 = 1 - \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \Big/ \sum_{i=1}^{N}(Y_i - \bar{Y}_i)^2 \quad (16)$$

where $N$ denotes the total sample number; $Y_i$, $\hat{Y}_i$ and $\bar{Y}_i$ denote the true RUL, the predicted RUL, and the mean predicted RUL, respectively.

### 4.2. *Comparison results and discussions*

The experiment arranges three full connect layers after MLSTM-MHA to acquire final RUL prediction results. The hyperparameters of MLSTM-MHA network chosen by the grid search are described as follows: the layer number of MLSTM-MHA is 2, the neural number of MLSTM is 64 and the head number of MHA is 2. After cross validation, other implementation details are described in Table II.

Table II. Implement details of the MLSTM-MHA

| Parameters | Value | Parameters | Value |
|---|---|---|---|
| Learning rate | 0.001 | Fc1 layer size | (64,16) |
| Epoch | 1000 | Fc2 layer size | (16,4) |
| Window size L | 5 | Fc3 layer size | (4,1) |
| Step size s | 1 | Activation layers | 2 |

To demonstrate the superiority of the proposed MLSTM-MHA method, some state of art methods such as RNN, LSTM, BILSTM, and GRU are employed for comparison. The LSTM-MHA also participated in comparative test to verify the effectiveness of multi-scale memory cell. For these time series prediction methods, the number of neurons, hidden layers, learning rate, and other hyperparameters are set by cross validation. All the models were implemented on a same workstation that had a GPU AMD Ryzen 7 5800H. In order to reduce the training time of all the models, we removed some data of the health stage. Then they are divided into training data, valid data, and test data in 8:1:1. According to the engineering experience, the fault thresholds of HM07, HM09, and HN60 are set to 0.1, 0.3, and 0.2 , respectively.

The RUL prediction results for roller HM07 of different models are shown in Fig. 4. It can be obviously noted that MLSTM-MHA and LSTM-MHA have better performance when other prediction methods have large fitting errors. And the health indicators of these two methods are 334 and 335, respectively, at the last run batch point. It suggests that the proposed MLSTM-MHA and LSTM-MHA methods have high prediction accuracy for ultimate roller lifetime and roller RUL prediction.

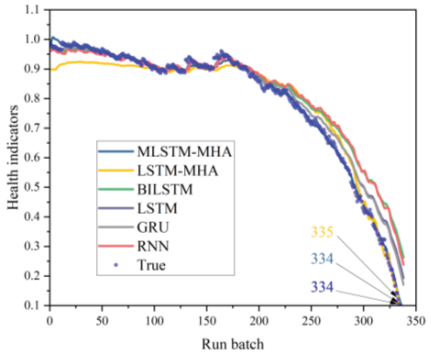


Fig. 4. The RUL prediction results for roller HM07 of different models.

Comparing MLSTM-MHA with LSTM-MHA, we can find that MLSTM-MHA is more consistent with the true health indicators in the first half of the roller run batch. The redundant information in input data makes LSTM-MHA unable to extract sufficient temporal information from the long time series data relying on single memory cell. However, the proposed novel structure of memory cells can deal with multi-scale temporal information such as long trend, medium trend, and short trend. Hence the proposed MLSTM-MHA method can retain outstanding performance in the first half of the roller run batches. After 100 roller run batches, the fitness of LSTM-MHA is improved by multi-head self-attention embedding mechanism. And comparing with traditional LSTM, the proposed LSTM-MHA can acquire better RUL prediction results at the last run batch point. It indicates that multi-head self-attention embedding mechanism can make the LSTM unit pay more attention to important information rather than redundant information.

With the same HM07 datasets, the evaluation indexes for different roller RUL prediction methods are calculated and listed in Table III. As for the RMSE evaluation metric, compared with LSTM-MHA, BILSTM, LSTM, GRU, and RNN, the improvement of MLSTM-MHA for HM07 datasets are 61.94%, 78.81%, 66.29%, 63.69%, and 78.68%. For the MAE evaluation metric, the improvements of MLSTM-MHA for HM07 datasets are 65.42%, 73.56%, 56.71%, 55.15%, and 74.86%. For the $R^2$ evaluation metric, the improvements of MLSTM-MHA for HM07 datasets are 2.37%, 9.08%, 3.15%, 2.65%, and 8.96%. It can be obviously found that MLSTM-MHA achieves the best performance above other prediction methods.

Table III. Evaluation indexes for different roller RUL prediction methods in HM07 datasets

| | RMSE | MAE | $R^2$ |
|---|---|---|---|
| MLSTM-MHA | **0.01409** | **0.01071** | **0.9961** |
| LSTM-MHA | 0.03702 | 0.03097 | 0.97308 |
| BILSTM | 0.06648 | 0.0405 | 0.91322 |
| LSTM | 0.0418 | 0.02474 | 0.96569 |
| GRU | 0.03881 | 0.02388 | 0.97043 |
| RNN | 0.06609 | 0.0426 | 0.91422 |

As shown in Fig. 5 and Fig. 6, the number of points to be predicted is set to 340 and 310 for HM09 and HN60 datasets, respectively. The true RUL for rollers is 33 and 37 run batches, respectively. And the RUL prediction results acquired by MLSTM-MHA and LSTM-MHA are {33,37} run batches and {38,38} run batches, respectively. The evaluation indexes for different roller RUL prediction methods are calculated and listed in Table IV and Table V. It can be seen that MLSTM-MHA still achieves the smallest value of RMSE and MAE error evaluation indexes and the biggest value of the $R^2$ score. For the RMSE evaluation metrics, compared with LSTM-MHA, BILSTM, LSTM, GRU, and RNN, the improvements of MLSTM-MHA for HM09 and HN60 datasets are 43.76%, 58.13%, 56.48%, 58.73%, and 62.24% and 11.54%, 51.28%, 44.84%, 44.38%, and 43.10%, respectively. For the MAE evaluation metrics, the improvements of MLSTM-MHA for HM09 and HN60 datasets are 42.65%, 57.31%, 55.91%, 60.66%, and 63.41% and 11.17%, 38.26%, 30.93%, 31.66%, and 36.23%, respectively. In the $R^2$ evaluation metrics, the improvement of MLSTM-MHA for HM09 and HN60 datasets is 7.32%, 17.42%, 15.60%, 18.17%, and 23.43%, and 1.11%, 14.60%, 9.97%, 9.72%, and 9.03%, respectively. Therefore, the proposed MLSTM-MHA method can be appropriately applied to datasets of different rollers' degradation types.

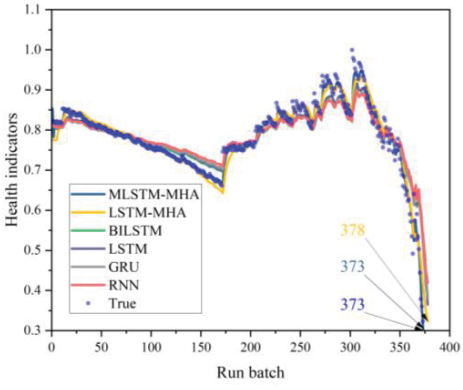


Fig. 5. The RUL prediction results for roller HM09 of different models.

Table IV. Evaluation indexes for different roller RUL prediction methods in HM09 datasets

| | RMSE | MAE | $R^2$ |
|---|---|---|---|
| MLSTM-MHA | **0.01956** | **0.01116** | **0.96943** |
| LSTM-MHA | 0.03478 | 0.01946 | 0.90332 |
| BILSTM | 0.04672 | 0.02614 | 0.82558 |
| LSTM | 0.04494 | 0.02531 | 0.8386 |
| GRU | 0.0474 | 0.02837 | 0.8204 |
| RNN | 0.05181 | 0.0305 | 0.78543 |

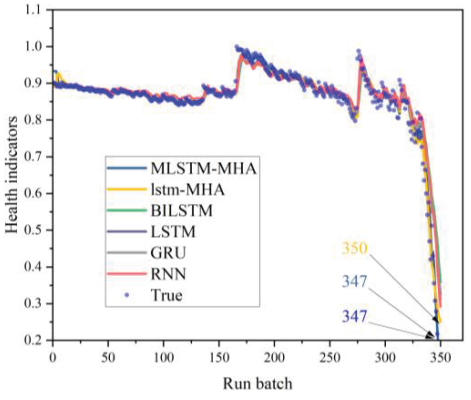


Fig. 6. The RUL prediction results for roller HN60 of different models.

Table V. Evaluation indexes for different roller RUL prediction methods in HN60 datasets

| | RMSE | MAE | $R^2$ |
|---|---|---|---|
| MLSTM-MHA | **0.02238** | **0.01304** | **0.96186** |
| LSTM-MHA | 0.0253 | 0.01468 | 0.95127 |
| BILSTM | 0.04594 | 0.02112 | 0.8393 |
| LSTM | 0.04057 | 0.01888 | 0.87465 |
| GRU | 0.04024 | 0.01908 | 0.87668 |
| RNN | 0.03933 | 0.02045 | 0.88219 |

## 5. Conclusion

In this paper, a novel RUL prediction method for hot strip mill rollers on the principle of time series forecasting was proposed. A new memory cell was designed to extract different scale information from long time series data. By dividing the memory cell into three parts, MLSTM can capture local trends, medium trends, and global trends information from degradation process. Then MLSTM-MHA was proposed by integrating forget gate and input gate with multi-head self-attention mechanism in MLSTM, , which could pay more attention to important

information by improving the weights of gate units. Via the industrial hot strip mill roller dataset by Baoshan Iron & Steel Co., Ltd., the performance of the proposed method was verified and compared to other methods based on BILSTM, LSTM, GRU, and RNN. The comparative results demonstrate that the MLSTM-MHA method possesses the stronger temporal data extracting and prediction ability, and thereby it is more suitable for hot strip mill rollers RUL prediction.

In future work, a weight adjustment strategy to multi-scale memory cell will be explored to improve the performance of online rollers' RUL prediction with multi-stage degradation types.

**Acknowledgement**

The authors would like to thank the Editor and anonymous referees for their careful work and remarkable comments which considerably help to improve this manuscript substantially. This work is supported in part by the National Key Research and Development Program of China under Grant 2020YFB1711100, and in part by National Natural Science Foundation of China under grant number 72001138 and grant number 52005327.

**References**

Zhang, Z., Si, X., Hu, C., and Lei, Y. (2018). Degradation data analysis and remaining useful life estimation: A review on Wiener-process-based methods. European Journal of Operational Research, 271(3), 775-796.

Wu, D., Jia, M., Cao, Y., Ding, P., and Zhao, X. (2022). Remaining useful life estimation based on a nonlinear Wiener process model with CSN random effects. Measurement, 205, 112232.

Ling, M. H., Ng, H. K. T., and Tsui, K. L. (2019). Bayesian and likelihood inferences on remaining useful life in two-phase degradation models under gamma process. Reliability Engineering & System Safety, 184, 77-85.

Yang, B., Liu, R., and Zio, E. (2019). Remaining useful life prediction based on a double-convolutional neural network architecture. IEEE Transactions on Industrial Electronics, 66(12), 9521-9530.

Chen, J., Jing, H., Chang, Y., and Liu, Q. (2019). Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process. Reliability Engineering & System Safety, 185, 372-382.

Ma, M., and Mao, Z. (2020). Deep-convolution-based LSTM network for remaining useful life prediction. IEEE Transactions on Industrial Informatics, 17(3), 1658-1667.

Han, T., Pang, J., and Tan, A. C. (2021). Remaining useful life prediction of bearing based on stacked autoencoder and recurrent neural network. Journal of Manufacturing Systems, 61, 576-591.

Liu, J., Lei, F., Pan, C., Hu, D., & Zuo, H. (2021). Prediction of remaining useful life of multi-stage aero-engine based on clustering and LSTM fusion. Reliability Engineering & System Safety, 214, 107807.

Jiao, R., Peng, K., Dong, J. (2021). Remaining useful life prediction for a roller in a hot strip mill based on deep recurrent neural networks. IEEE/CAA Journal of Automatica Sinica, 8, 1345-1354.

An, D., Kim, N. H., Choi, J. H. (2015). Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. Reliability Engineering & System Safety, 133, 223-236.

Li, Z., Wang, Y., Wang, K. (2019). A deep learning driven method for fault classification and degradation assessment in mechanical equipment. Computers in Industry, 104, 1-10.

Elsheikh, A., Yacout, S., Ouali, M. S. (2019). Bidirectional handshaking LSTM for remaining useful life prediction. Neurocomputing, 323, 148-156.

Krot, P., Prykhodko, I., Raznosilin, V., and Zimroz, R. (2020). Model based monitoring of dynamic loads and remaining useful life prediction in rolling mills and heavy machinery. Advances in Asset Management and Condition Monitoring: COMADEM, 2020, 399-416.