Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 59-20 (2025) 2614–2618

Research on Mechanical Equipment Remaining Useful Life Prediction Method Based on Attention Mechanism and Feature Fusion

Song Hongjian*, Xiao Fuhong*, Dong Yunjia*, Feng Xiaoen*, Lei Mingjia*, Li Yuqing*

**Harbin Institute of Technology, Harbin, 150001 China (dyjhit0205@126.com)*

Abstract: This paper proposes a deep learning method combining multi-head attention mechanisms and multi-source feature fusion for remaining useful life (RUL) prediction of mechanical equipment. The approach extracts temporal dynamic features through a Long Short-Term Memory (LSTM) network and introduces a multi-head cross-attention module to dynamically allocate weights to temporal features. Additionally, a feature enhancement module is designed to integrate handcrafted features with CNN-extracted local features, enhancing the model's capability to characterize equipment degradation patterns. Experiments on the C-MAPSS FD004 dataset demonstrate that the proposed method achieves RMSE and Score metrics of 22.64 and 4723.19, outperforming LSTM-Attention methods by 19.3% and 39.8%. The research provides a novel solution for RUL prediction under complex operating conditions.

Keywords: Remaining Useful Life prediction, Attention mechanism, Feature fusion, LSTM, Health monitoring

1. INTRODUCTION

With the rapid advancement of technology, modern advanced equipment has become increasingly complex and sophisticated. As equipment ages or undergoes wear, its performance can degrade significantly, making health management critically important. Remaining Useful Life (RUL) prediction, a core task in equipment health management, directly impacts maintenance strategies and operational costs. Traditional methods primarily rely on physical models (e.g., Wiener process) and statistical models (e.g., Cox proportional hazards model), but they suffer from complex modeling requirements and poor generalization capability. Driven by the development of industrial big data technologies, data-driven approaches have emerged as mainstream solutions. These methods directly model operational data to learn degradation patterns, with two dominant paradigms: traditional machine learning-based methods and deep learning-based methods. Among these, deep learning demonstrates significant advantages in RUL prediction due to its automatic feature learning capabilities, enabling superior performance in capturing complex degradation dynamics.

1.1 Research Status and Challenges

Current deep learning approaches can be categorized into the following two classes:

1) Single-model methods:

   LSTM: Zheng et al. employed LSTM to capture temporal dependencies but only utilized features from the last time step, ignoring contributions from intermediate states.

   CNN: Babu et al. proposed deep CNNs to extract local features, yet they lack capability in modeling long-term temporal dependencies.

2) Hybrid-model methods:

   LSTM + Attention: Chen et al. introduced a self-attention mechanism to allocate temporal weights but failed to integrate handcrafted features.

   Feature fusion: Zhang et al. concatenated automatically learned features with handcrafted features but lacked interactive fusion mechanisms.

1.2 Existing challenges

1) Insufficient feature utilization: Traditional LSTM only leverages terminal features, neglecting the differential contributions of multi-time-step information.

2) Lack of domain knowledge integration: Handcrafted features (e.g., trend coefficients) encode physical degradation patterns but are not deeply coupled with deep learning features in existing methods.

3) Fragmentation between local and global features: Local features extracted by CNNs and global temporal features learned by LSTMs are not effectively fused.

1.3 Contributions of This Work

To address the aforementioned challenges, this paper proposes the following contributions:

1) Multi-head Cross-Attention Mechanism: A scaled dot-product attention mechanism is introduced to dynamically allocate temporal weights, enhancing the expressive power of critical features.

2) Multi-Source Feature Fusion Module: This module integrates LSTM-based temporal features, CNN-extracted local features, and handcrafted features through cross-attention, enabling interactive fusion.

3) Experimental Validation: The model's performance is rigorously validated on the C-MAPSS FD004 dataset, demonstrating significant optimization in prediction accuracy.

10.1016/j.ifacol.2025.11.553

2. METHODOLOGY

2.1 Overall Architecture

The model architecture is illustrated in Figure 1 and consists of the following modules:

1) CNN Local Feature Extractor: Captures local patterns in sensor data through residual convolution.
2) BiLSTM Temporal Encoder: Models long-term temporal dependencies using bidirectional LSTM.
3) Multi-Head Cross-Attention Module: Fuses temporal features with handcrafted features.
4) Feature Enhancement Module: Combines CNN and LSTM features to produce the final prediction.

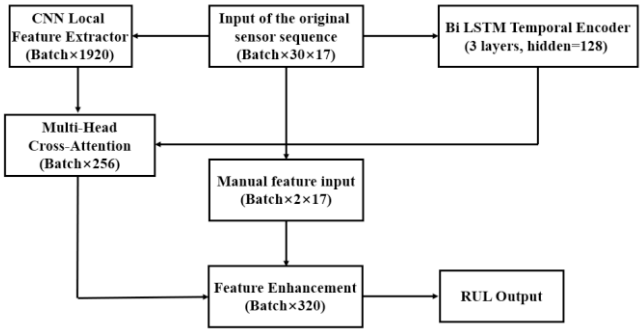

Figure 1. Overall architecture.

2.2 Bidirectional LSTM Temporal Encoding

Given an input sequence $X \in \mathbb{R}^{B\times T\times d}$ (where B is the batch size, T is the number of time steps, and d is the sensor dimension), The LSTM unit update formula is:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$
$$\sigma_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \odot \tanh(C_t)$$

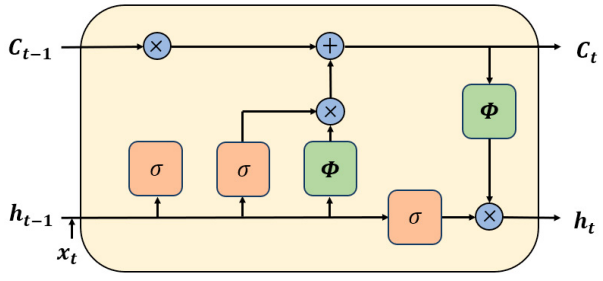

Figure 2. LSTM Schematic diagram.

The forward and backward hidden states of the bidirectional LSTM are computed as follows:

$$\overrightarrow{h_t} = \mathrm{LSTM}(x_t, \overrightarrow{h_{t-1}})$$
$$\overleftarrow{h_t} = \mathrm{LSTM}(x_t, \overleftarrow{h_{t+1}})$$

The outputs from both directions are concatenated to form the temporal feature $X \in \mathbb{R}^{B\times 2n}$ (where n is the number of hidden units):

$$H_{\mathrm{BiLSTM}} = [\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_T}] \oplus [\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_T}]$$
$$H_{\mathrm{BiLSTM}} = [\vec{h}_1 \oplus \overleftarrow{h}_1, \ldots, \vec{h}_T \oplus \overleftarrow{h}_T]$$

Traditional unidirectional LSTMs capture only forward temporal dependencies. However, during equipment degradation, fault features may exhibit patterns along the reverse temporal axis (e.g., propagation paths of abnormal signals in the terminal phase).

By combining forward ($\vec{h}_T$) and reverse ($\overleftarrow{h}_T$) hidden states, the bidirectional architecture comprehensively models global temporal dependencies in sensor data. For example, abnormal equipment temperature may simultaneously be influenced by prior wear (forward direction) and subsequent cooling effects (reverse direction).

Under complex operating conditions, equipment degradation patterns may exhibit asymmetry (e.g., sudden faults). Bidirectional structures are better suited to capturing such features.

2.3 Multi-Head Cross-Attention Mechanism

Temporal feature $Q = H_{\mathrm{BiLSTM}} \in \mathbb{R}^{T\times 2n}$

Handcrafted feature $K = F_{\mathrm{hand}} \in \mathbb{R}^{B\times k}$ (where k is the dimension of handcrafted features)

Project Q and K into Query (Q), Key (K), and Value (V) respectively:

$$Q' = QW_Q (W_Q \in \mathbb{R}^{2n\times d_k})$$
$$\mathrm{K}' = KW_K (W_K \in \mathbb{R}^{k\times d_k})$$
$$\mathrm{V}' = KW_V (W_V \in \mathbb{R}^{k\times d_v})$$

Compute the attention weight matrix A:

$$A = \mathrm{softmax}\left(\frac{Q'K'^{\mathrm{T}}}{\sqrt{d_k}}\right) \in \mathbb{R}^{T\times B}$$

Weighted value vector:

$$Z = AV' \in \mathbb{R}^{T\times d_v}$$

Split the d_v-dimensional space into h heads, compute independently, and concatenate:

$$Z_{\mathrm{multi}} = \mathrm{Concat}(Z_1, \ldots, Z_h)W^O (W^O \in \mathbb{R}^{hd_v\times d_{\mathrm{out}}})$$

Self-Attention: Computes feature correlations within a single modality (e.g., using only temporal features), which may neglect cross-modal interactions.

Cross-Attention: Achieves cross-modal alignment by mapping temporal features (Q) and handcrafted features (K, V) into different spaces. For example, when the trend coefficient (k) in handcrafted features indicates rapid degradation, the model should focus more on high-frequency vibration signals in the terminal phase of the time series.

The dot product $Q'K'^{\mathrm{T}}$ increases in magnitude as d_k grows, leading to gradient saturation in the Softmax function. The

scaling operation stabilizes gradients and improves training efficiency.

2.4 Multi-Modal Feature Fusion

All figures must be embedded in the document. When you include the image, make sure to insert the actual image rather than a link to your local computer.

$$F_{\text{fused}} = G \odot F_{\text{CNN}} + (1 - G) \odot Z_{\text{multi}}$$

In Steady-State Operation, CNN local features (e.g., vibration spectrum) may be more reliable. In Transitional Conditions: LSTM temporal features (e.g., temperature variation rate) are more predictive.

The gating weight G, generated via a Sigmoid function, enables soft selection within the [0,1] interval, avoiding information loss caused by hard thresholds.

Traditional residual structures ($F_{\text{fused}} = F_{\text{CNN}} + Z_{\text{multi}}$) assume aligned feature spaces. However, CNN local features (e.g., vibration amplitude) and LSTM temporal features may reside in different metric spaces. The gating mechanism achieves spatial alignment through nonlinear transformations, better adapting to multi-modal characteristics.

2.5 Loss Function and Optimization

The Mean Squared Error (MSE) loss and the Adam optimizer are adopted:

$$\mathcal{L} = \text{MSE} + \lambda(\| W \|_2^2 + \| b \|_2^2)$$

RUL prediction is fundamentally a regression problem. The Mean Squared Error (MSE) loss provides smooth gradient signals, making it well-suited for end-to-end training. Comparative experiments demonstrate that while Huber loss is more robust to outliers, MSE remains the optimal choice for the C-MAPSS dataset, where anomalous samples account for less than 5% of the data.

L2 Regularization can mitigate overfitting, Suppresses overfitting to noisy features. For example, certain sensors (IDs 22, 23) exhibit significant reading fluctuations under specific operating conditions, and L2 regularization constrains their corresponding weights.

2.6 Sliding window

The sliding window technique is typically used for data segmentation. In this study, the sliding window is designed to span 30 cycles because remaining useful life prediction is a time-series problem, requiring segmented data training. By extracting fixed-length subsequences, the model captures temporal degradation characteristics of the equipment. The time window is deeply integrated with the model architecture: a 1D CNN extracts local features from the windowed data, while an LSTM processes the full 30-cycle sequences. During testing, only the last time window in the dataset needs to be analyzed to perform RUL prediction.

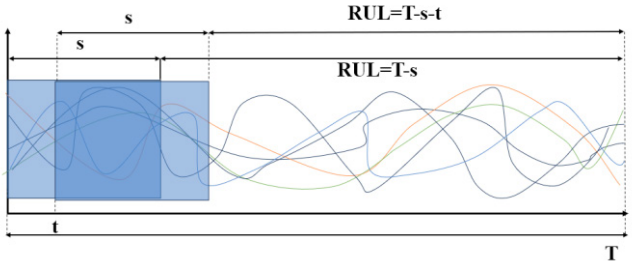

Figure 3. Sliding window Schematic diagram.

3. EVALUATION

3.1 Dataset and Evaluation Metrics

This study utilizes the CMAPSS FD004 dataset, which contains 249 training engines and 248 test engines.

The advantage of training with the FD004 dataset lies in its ability to highly simulate the complexity of real industrial environments. This dataset covers six distinct operating conditions (e.g., dynamic variations in altitude, Mach number, and throttle angle) and two coupled fault types (e.g., simultaneous degradation of fans and high-pressure compressors), realistically reflecting the dynamic degradation processes of equipment under variable loads and multi-component failure scenarios. With a large-scale coverage of 249 training engines and 248 testing engines, FD004 not only requires models to adapt to sensor data distribution shifts caused by switching between operating conditions (e.g., feature variations in high-altitude low-temperature vs. low-altitude high-temperature environments) but also demands solutions to challenges such as imbalanced fault type distributions (e.g., one fault type accounting for only 20% of samples) and few-shot learning difficulties, thereby rigorously validating the model's robustness and generalization capabilities.

Table 2. CMAPSS Dataset

Dataset	FD001	FD002	FD003	FD004
Training engines	100	260	100	249
Testing engines	100	259	100	248
Working conditions	1	6	1	6
Faulty types	1	1	2	2

The preprocessing steps applied to the data include: (1) removing constant-value sensors (indices 5, 9, 10, 14, 20, 22, 23) that have negligible impact on equipment lifespan; (2) segmenting data using sliding windows (window size: 30, stride: 1); and (3) normalizing the data to the [0,1] interval.

RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(y_i^{pred} - y_i^{true})^2}$$

The penalty-based scoring metric, as the official evaluation method of PHM 2008, is utilized to assess the performance of the RUL prediction model on the dataset used in this study.

$$Score = \begin{cases} \sum_{i=1}^{L} (e^{-\frac{y_i^{pred} - y_i^{true}}{13}} - 1), when y_i^{pred} < y_i^{true} \\ \sum_{i=1}^{L} (e^{\frac{y_i^{pred} - y_i^{true}}{10}} - 1), when y_i^{pred} > y_i^{true} \end{cases}$$

This scoring function imposes a relatively small penalty for early prediction. Conversely, for lagging prediction, the penalty is relatively greater and it is more applicable to the actual situation.

3.3 Results Comparison

The life prediction results of FD001 by the method proposed in this paper are shown in the following figure.

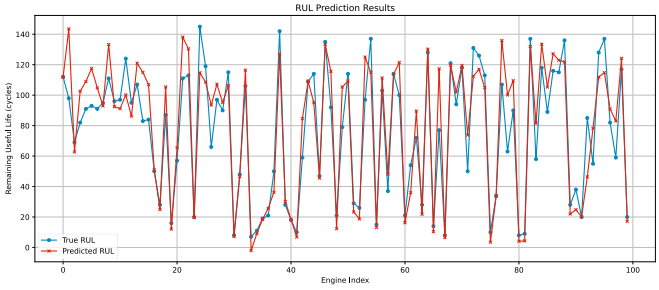

Figure 4. RUL Prediction Results (FD001).

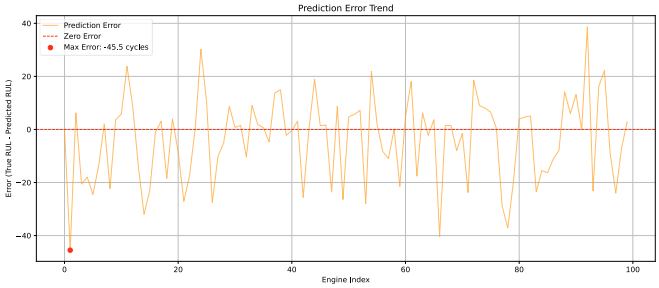

Figure 5. Prediction Error Trend (FD001).

As can be seen from the above figure, the experimental results on the FD001 dataset can prove the feasibility of this method. The following focuses on discussing the results of the FD004 dataset.

The life prediction results of FD004 by the method proposed in this paper are shown in the following figure.

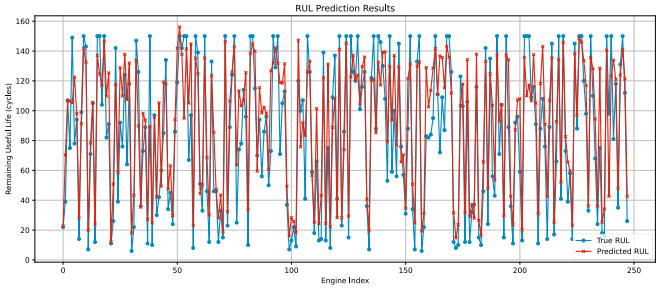

Figure 6. RUL Prediction Results (FD004).

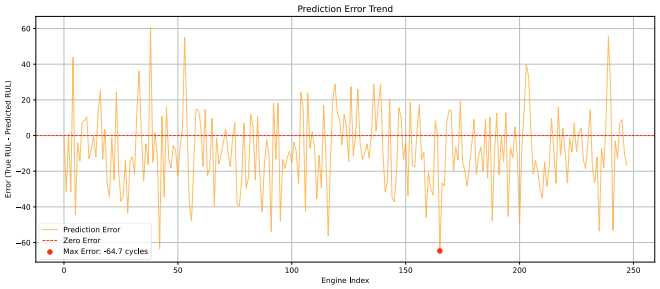

Figure 7. Prediction Error Trend (FD004).

The comparison between the method proposed in this paper and other existing methods is as follows.

Table 2. Results Comparison

Method	RMSE	Score
Standard LSTM	29.12	12551.44
LSTM-Attention	28.04	7850.40
LSTM-HF	27.66	7812.37
Proposed Method	22.64	4723.19

It can be seen from the result graph that the error between most of the predicted values and the true values is less than 20cycles. The multi-head attention mechanism reduces RMSE significantly, demonstrating that dynamic weight allocation effectively captures critical temporal information. And the feature fusion module improves the Score significantly, validating the synergistic advantage of multi-source feature collaboration, outperforming LSTM-Attention methods by 19.3% and 39.8%.

The proposed method achieves state-of-the-art (SOTA) performance on both metrics, highlighting the effectiveness of cross-modal interaction and adaptive feature fusion.

4. CONCLUSIONS

In this paper, we propose a deep learning framework based on a bidirectional long short-term memory (BiLSTM) network and a multi-head cross-attention mechanism for remaining useful life (RUL) prediction of mechanical equipment under complex operating conditions. By leveraging BiLSTM to model global temporal dependencies, combined with the multi-head cross-attention mechanism to dynamically align temporal features with domain-specific handcrafted features, we achieve adaptive fusion of multimodal features through a gated weighting strategy. Experimental results on the C-MAPSS FD004 dataset demonstrate that our method achieves RMSE and Score values of 22.64 and 4723.19, respectively, effectively validating the efficacy of bidirectional temporal modeling and the advantages of the cross-modal attention mechanism.

ACKNOWLEDGEMENTS

This research was supported by National Key R&D Program of China 2021(YFA1003501), the National Natural Science Foundation of China (No: 52075117), Provincial Key R&D Program of Heilongjiang (2022ZX01A20).

REFERENCES

Liao L. Discovering prognostic features using genetic programming in remaining useful life prediction[J]. IEEE Trans. Ind. Electron., 2014.

Cox D R. Regression models and life-tables[J]. Journal of the Royal Statistical Society, 1972.

Chen Z, et al. Machine remaining useful life prediction via an attention-based deep learning approach[J]. IEEE Trans. Ind. Electron., 2021.

Zheng S, et al. Long short-term memory network for remaining useful life estimation[C]. PHM, 2017.

Babu G S, et al. Deep convolutional neural network based regression approach for estimation of remaining useful life[C]. DASFAA, 2016.

Zhang C, et al. Multiobjective deep belief networks ensemble for remaining useful life estimation[J]. IEEE Trans. Neural Netw. Learn. Syst., 2017.