Engineering Applications of Artificial Intelligence 161 (2025) 112128

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

# Deep ensemble learning and error correction method for remaining useful life prediction of rolling bearings

Wenzhe Yin [a], Hong Xia [a,*], Enrico Zio [b,c], Xueying Huang [a]

[a] Fundamental Science on Nuclear Safety and Simulation Technology Laboratory, Harbin Engineering University, Harbin, 150001, China
[b] MINES Paris PSL University, CRC, Sophia Antipolis, France
[c] Energy Department, Politecnico di Milano, 20156, Milan, Italy

ARTICLE INFO

*Keywords:*
Rolling bearing
Remaining useful life
Deep learning
Ensemble learning
Error correction

ABSTRACT

Rolling bearings are critical components of rotating machinery, and the prediction of their remaining useful life (RUL) is important for system reliability and operating efficiency. A novel RUL prediction method based on deep ensemble learning and error correction is here proposed. Firstly, the moving average filter (MAF) is applied to preprocess vibration signals for removing outliers and noise. Then, time-domain features are extracted from the processed vibration signals, and optimized to obtain imperative feature signals. Subsequently, a deep ensemble learning model is built with gated recurrent unit (GRU), bidirectional GRU (BiGRU), long short-term memory (LSTM) and bidirectional LSTM (BiLSTM) as base learners, and the overall performance of the prediction model is enhanced by introducing an error correction strategy. The MAF method is also used to smooth the trend of the RUL prediction outcomes. Finally, the proposed method is applied to two full-lifecycle rolling bearing datasets: the Prognostics and Health Management 2012 (PHM2012) dataset and the Intelligent Maintenance System (IMS) dataset. It is evaluated using mean square error (MSE), mean absolute error (MAE), and the R-square coefficient of determination ($R^2$). The test results demonstrate that the method achieves highly accurate RUL predictions: on the PHM2012 dataset, the MSE, MAE, and $R^2$ reach 0.000380, 0.013695, and 0.994716, respectively; on the IMS bearing dataset, the corresponding values are 0.001056, 0.015403, and 0.978346. In addition, the method outperforms traditional single models (GRU, BiGRU, LSTM, BiLSTM) as well as the Transformer model in both cases, with $R^2$ improvements over the Transformer of 0.011065 and 0.008744, respectively. These results highlight the superior generalization capability and robustness of the proposed method, making it applicable to industrial environments requiring reliable RUL prediction.

## 1. Introduction

Rotating machinery is utilized in many industries, such as energy, aerospace, transportation and chemical. Among the key components in such machinery, rolling bearings are particularly prevalent, with their health status is critically linked to the reliability and safety of the entire rotating machinery system (Rai and Upadhyay, 2016). However, due to factors such as the operating environment, mechanical and structural characteristics, rolling bearings are also quite sustainable and prone to develop defection and faults (Nandi et al., 2005). Bearing failures can lead not only to critical equipment malfunctions and reduced system efficiency but also to severe accidents, resulting in significant financial losses and potential casualties (Mueller et al., 2023). Therefore, accurately assessing equipment health, avoiding unnecessary maintenance or replacements, and implementing precise and informed maintenance strategies are essential for ensuring operational safety and improving economic performance (Kumar et al., 2023). As an emerging maintenance strategy, Prognostics and Health Management (PHM) utilizes advanced monitoring and analysis techniques to assess equipment condition, predict potential failures, and provide maintenance decisions before failures occur, thereby optimizing the allocation of maintenance resources (Zio, 2022).

Remaining useful life (RUL) prediction is one of the core technologies for implementing PHM and remains a major research challenge (Peng et al., 2025). Effectively utilizing condition monitoring data, combined with historical information and appropriate modeling techniques, to accurately estimate the RUL of equipment or critical components has become a key research focus in this field. Recently,

* Corresponding author.
*E-mail address:* xiahong@hrbeu.edu.cn (H. Xia).

https://doi.org/10.1016/j.engappai.2025.112128
Received 23 October 2024; Received in revised form 7 August 2025; Accepted 26 August 2025
Available online 29 August 2025

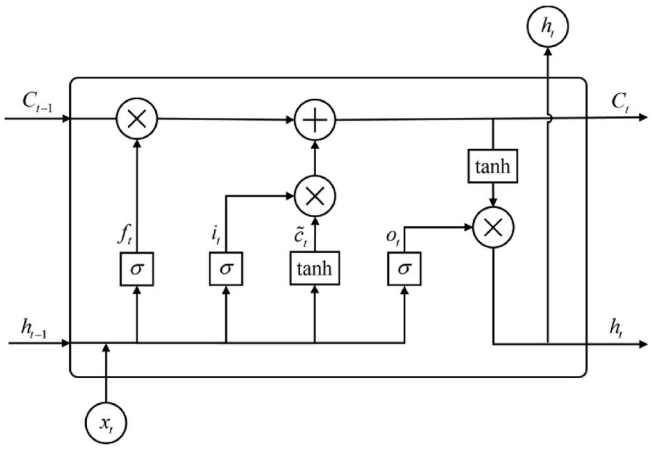



**Fig. 1.** The structure of LSTM.

considerable attention has been given to the advancement and implementation of RUL prediction methods for rolling bearings. These approaches can broadly be classified into two categories: physics-based methods (Bolander et al., 2009) and data-driven methods (Ochella et al., 2022). Physics-based methods heavily rely on prior knowledge of equipment degradation mechanisms, which limits their application in practice. With the swift progress in sensor and computing technologies, it has become feasible to gather and analyze operational data that reflects the health condition of the equipment. This has driven the broad adoption of data-driven methods for RUL prediction tasks (Zio et al., 2010) (Shaheen et al., 2023).

In particular, deep learning has achieved notable advancements in the field of industrial equipment health management (Fink et al., 2020). Recurrent Neural Networks (RNNs) are particularly effective for handling time-series data due to their ability to capture temporal dependencies (Amalou et al., 2022). As a result, RNNs and their variants often demonstrate strong performance in RUL prediction tasks. Guo et al. developed a bearing RUL method that utilizes sensitive features and RNN (Guo et al., 2017). However, RNNs are unable to handle long-term information and suffer from the vanishing gradients problem. To address these limitations, improved versions of RNNs, including long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and recurrent gated unit (GRU) (Cho, 2014), have been proposed. They effectively consider long-term dependencies and solve the vanishing gradients encountered during the training of traditional RNNs. Zhang et al. utilized convolutional neural network (CNN) as a local feature extractor, then, employed LSTM to analyze these features and predict the RUL (Zhang et al., 2020). Que et al. constructed a stacked GRU to predict bearing RUL (Que et al., 2021).

However, traditional LSTM and GRU networks are limited to extracting unidirectional features from time-series data, which restricts their ability to fully model temporal dependencies. To address this limitation, bidirectional architectures such as Bidirectional LSTM (BiLSTM) (Atef et al., 2022) and Bidirectional GRU (BiGRU) (Meng et al., 2024) have been introduced to capture both forward and backward temporal relationships. Several studies have applied BiLSTM and BiGRU networks to bearing RUL prediction tasks and have achieved good results. Shen et al. combined multi-head attention with BiLSTM and applied it to bearing RUL prediction (Shen et al., 2022). Shang et al. employed the CNN to extract features from original vibration signals, which were then fed into the BiGRU network for predicting the RUL of bearings (Shang et al., 2022).

On the other hand, compared to single models, ensemble models can leverage the strengths and complementary characteristics of multiple base learners (Yang et al., 2023). By integrating the predictions from multiple learners, ensemble models not only mitigate potential biases of individual models but also enhance overall prediction accuracy and robustness (Mienye and Sun, 2022). This makes them well-suited for handling complex tasks with greater adaptability. Usman et al. constructed a deep ensemble learning model based on support vector machine (SVM), CNN and LSTM, and applied it to disease prediction (Usman et al., 2021). Zhang et al. integrated LSTM, BiGRU, and Extreme Learning Machine (ELM) classifiers to perform facial micro-expression detection (Zhang et al., 2025).

Inspired by this success, this study develops and applies a deep ensemble learning model for rolling bearings RUL prediction. Given the good prediction performance of deep recurrent neural networks, such as LSTM, BiLSTM, GRU and BiGRU, they are here used as base learners to construct the deep ensemble learning model. Furthermore, in recent years, some studies have attempted to introduce error correction mechanisms to enhance the performance of prediction models. Liang et al. used SVM to predict wind speed series and established an error correction model using SVM and extreme learning machine to correct the wind speed predictions (Liang et al., 2016). Considering that artificial neural network (ANN) possesses good nonlinear fitting ability, which is conducive to learning the evolution patterns of errors, this study constructs an error correction model based on ANN further to improve the overall performance of the prediction method.

To enhance the accuracy of rolling bearing RUL prediction, this study proposes a novel hybrid method that combines deep ensemble learning with error correction. The proposed approach introduces several key contributions. First, a deep ensemble model is developed using recurrent neural networks (LSTM, BiLSTM, GRU, and BiGRU) as base learners. This ensemble leverages the strengths of each architecture, resulting in improved predictive performance through model diversity. Second, the ANN-based error correction module is introduced to refine the outputs of the base learners, further boosting accuracy. Third, the moving average filter (MAF) is employed both in the preprocessing of original signals and in the postprocessing of prediction outputs. The MAF effectively suppresses noise and outliers in the input data while also smoothing the prediction results to mitigate short-term fluctuations. Finally, bearing degradation stages are identified using root mean square values as health indicators, and a piecewise linear function is designed to construct the RUL labels. This strategy enables a more realistic and discriminative evaluation of model performance. Collectively, these contributions form a robust and adaptable framework that significantly enhances RUL prediction accuracy for rolling bearings.

The rest of the paper is described as follows. Section 2 briefly introduces the background knowledge of MAF, LSTM, BiLSTM, GRU, BiGRU and ANN. Section 3 outlines the proposed framework for bearing RUL prediction. Section 4 demonstrates the efficacy of the proposed method through validation using two full-lifecycle datasets of bearings, and comparative tests and result analysis are conducted. Finally, Section 5 summaries the conclusions.

## 2. Brief theoretical background

### 2.1. Moving average filter

Due to the complexity of the structure of rotating machinery and the variety of operating conditions, the collected rolling bearing vibration signals typically contain some noise and outliers (Chegini et al., 2019). The interfered signal data can negatively affect the accuracy of predicting the RUL of bearings. To address this issue and enable the prediction model to more effectively learn valuable information from the full-lifecycle bearing data, this study adopts MAF to preprocess the original signals, effectively eliminating the interference caused by noise and outliers.

MAF is essentially a low-pass filter. Due to its simple principle, easy implementation and high computational efficiency, it is widely used in signal processing (Golestan et al., 2013). The basic idea of MAF is that the value of a data point is replaced by the average value of the current data point and the data points within a corresponding time window. This

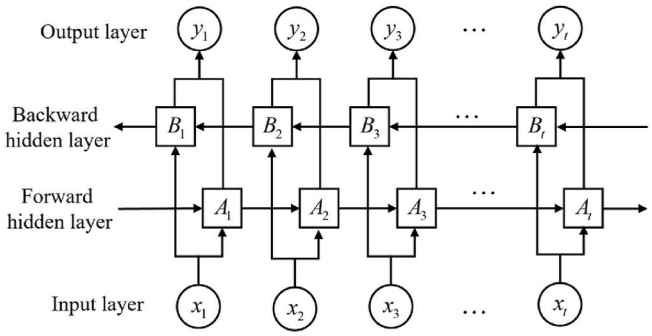



**Fig. 2.** The bidirectional network structure.

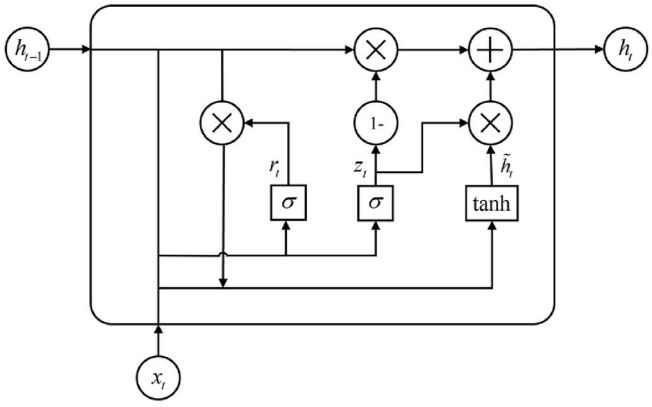



**Fig. 3.** The structure of GRU.

approach maintains the dynamic trend of the original signal while achieving the purpose of filtering. The calculation of MAF can be expressed as:

$$x_i = \frac{1}{m}\sum_{j=i}^{i+m-1} x_j \tag{1}$$

where $x_i$ represents the filtered data, $x_j$ represents the original data and $m$ represents the window length.

### 2.2. Bidirectional long short-term memory

RNNs are neural networks that handle time series data effectively, but they suffer from the issue of long-term dependencies. This means that as the length of the time series grows, the model becomes unable to make use of information from earlier parts of the sequence (Pascanu et al., 2013). To tackle this challenge, the LSTM network with a three-gate structure was proposed (Hochreiter and Schmidhuber, 1997), whose architecture is shown in Fig. 1. Through its three-gate structure consisting of input gate $i_t$, forget gate $f_t$, and output gate $o_t$, the LSTM effectively strengthens the long-term memory capabilities of RNNs, while addressing the issues of gradient vanishing and explosion. The calculation of the LSTM network can be expressed as:

$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i) \tag{2}$$

$$f_t = \sigma\left(W_f \cdot [x_t, h_{t-1}] + b_f\right) \tag{3}$$

$$o_t = \sigma(W_o \cdot [x_t, h_{t-1}] + b_o) \tag{4}$$

$$\widetilde{c}_t = \tanh(W_c \cdot [x_t, h_{t-1}] + b_c) \tag{5}$$

$$c_t = i_t \cdot \widetilde{c}_t + f_t \cdot c_{t-1} \tag{6}$$

$$h_t = o_t + \tanh(c_t) \tag{7}$$

where $x_t$ represents the input at time $t$, $h_{t-1}$ and $h_t$ represent the hidden state at the previous moment and the current moment, respectively, $c_{t-1}$ and $c_t$ represent the cell states at time $t-1$ and $t$, respectively, $\widetilde{c}_t$ represents the candidate value for the cell state, $\sigma$ represents the sigmoid activation function, $W$ and $b$ represent the corresponding weight matrices and bias vectors, respectively.

BiLSTM is an optimization and improvement of the traditional unidirectional LSTM. Using a bidirectional network structure allows for recursive feedback from both past and future hidden layer states. This enables it to further explore the intrinsic relationships between current data points and those from past and future time steps, thereby optimizing the utilization of feature data and enhancing the precision of the prediction model (Rathore and Harsha, 2022). The bidirectional network structure is depicted in Fig. 2. Among them, $x_1, x_2, \ldots, x_t, \ldots, x_n$ represent the corresponding input data at time $t_1, t_2, \ldots, t_t, \ldots, t_n$, $A_1, A_2, \ldots, A_t, \ldots, A_n$ and $B_1, B_2, \ldots, B_t, \ldots, B_n$ represent the forward and backward iterated LSTM hidden states, respectively, $y_1, y_2, \ldots, y_t, \ldots, y_n$ represent the corresponding output data. The calculation of the BiLSTM network can be represented as follows:

$$\overrightarrow{A_t} = LSTM_{forward}\left(x_t, \overrightarrow{A_{t-1}}\right) \tag{8}$$

$$\overleftarrow{B_t} = LSTM_{backward}\left(x_t, \overleftarrow{B_{t+1}}\right) \tag{9}$$

$$y_t = W_y\left[\overrightarrow{A_t}, \overleftarrow{B_t}\right] + b_y \tag{10}$$

where $\overrightarrow{A_t}$ and $\overleftarrow{B_t}$ are the hidden states of the forward and backward LSTM, respectively, $W_y$ and $b_y$ represent the weight matrix and bias vector for the output layer, respectively.

### 2.3. Bidirectional gated recurrent unit

As a variant of LSTM, GRU inherits the capability of LSTMs to handle gradient issues while simplifying the network structure, resulting in improved computational efficiency (Fu et al., 2016). The structure of GRU is illustrated in Fig. 3. Compared to the three-gate structure of LSTM, GRU comprises only an update gate and a reset gate. The reset gate $r_t$ governs the degree to which the current information is blended with the information from the previous time step, whereas the update gate $z_t$ manages the extent to which the state information from the previous time step is retained at the current time step. The computation of a GRU can be represented as follows:

$$r_t = \sigma(W_r \cdot [x_t, h_{t-1}] + b_r) \tag{11}$$

$$z_t = \sigma(W_z \cdot [x_t, h_{t-1}] + b_z) \tag{12}$$

$$\widetilde{h}_t = \tanh(W_h \cdot [x_t, r_t \cdot h_{t-1}] + b_h) \tag{13}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \widetilde{h}_t \tag{14}$$

where $x_t$ and $h_t$ represent the input and output at time t, respectively, $\widetilde{h}_t$ represents the candidate activation, $W$ and $b$ represent the weight matrices and bias vectors, respectively, tanh is the hyperbolic tangent function. Similarly, based on GRU, BiGRU can be constructed by introducing a bidirectional network structure is given in Fig. 2, to enable the network model to obtain stronger feature learning capabilities (Dai et al., 2022). The bidirectional computation process of BiGRU is similar to that described in Equations (8)–(10), and is implemented based on the standard GRU architecture.

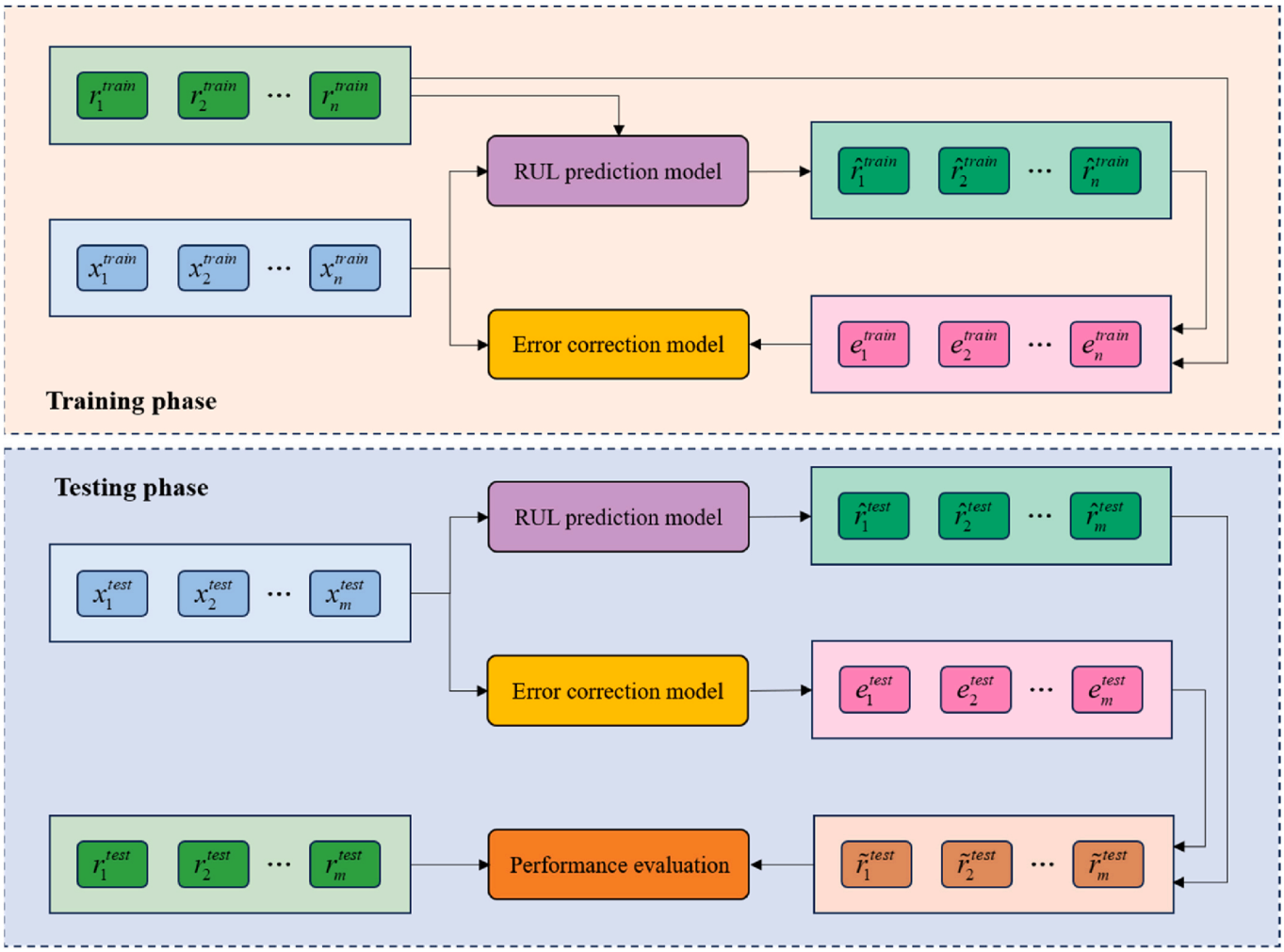



**Fig. 4.** Schematic diagram of error correction strategy.

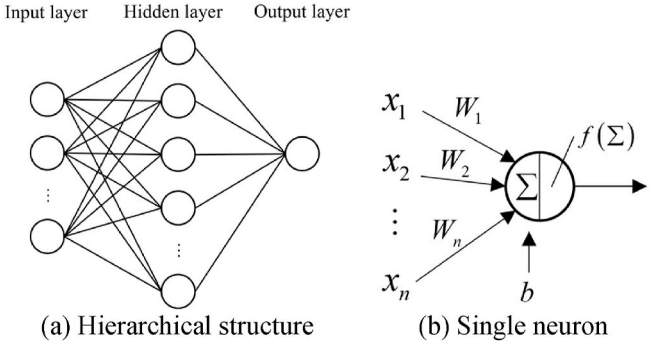



**Fig. 5.** General structure of ANN.

## 3. Error correction strategy

To enhance the overall prediction precision, this study innovatively introduces the error correction strategy into the task of rolling bearing RUL prediction. An error correction model is established based on the error results of the prediction model, and this error correction model is combined with the prediction model to achieve more accurate RUL predictions. Specifically, as illustrated in Fig. 4, let the training set be $\{x_1^{train}, x_2^{train}, \ldots, x_n^{train}\}$ and the test set be $\{x_1^{test}, x_2^{test}, \ldots, x_m^{test}\}$, with corresponding RUL labels $\{r_1^{train}, r_2^{train}, \ldots, r_n^{train}\}$ and $\{r_1^{test}, r_2^{test}, \ldots, r_m^{test}\}$, respectively. First, the prediction model $M_1$ is trained using $\{x_i^{train}\}$ and $\{r_i^{train}\}$.Once training is completed, the model $M_1$ can generate predictions on the training set, resulting in $\{\hat{r}_1^{train}, \hat{r}_2^{train}, \ldots, \hat{r}_n^{train}\}$.By comparing the predicted values with the true labels, the error series $\{e_i^{train}\}$ can be obtained as:

$$e_i^{train} = r_i^{train} - M_1\left(x_i^{train}\right) = r_i^{train} - \hat{r}_i^{train} \tag{15}$$

Next, the training set $\{x_i^{train}\}$ and the corresponding error series $\{e_i^{train}\}$ are used as inputs and targets, respectively, to train the error correction model $M_2$, aiming to learn the distribution and evolution patterns of the prediction errors.

In the testing phase, for the test set $\left\{x_j^{test}\right\}$, the prediction model $M_1$ produces the initial predictions $\left\{\hat{r}_j^{test}\right\}$. The error correction model $M_2$ then estimates the prediction errors $\left\{e_j^{test}\right\}$ based on the input features. The final RUL prediction results $\left\{\tilde{r}_j^{test}\right\}$ are obtained by combining the initial predictions with the estimated errors, which can be expressed as:

$$\tilde{r}_j^{test} = M_1\left(x_j^{test}\right) + M_2\left(x_j^{test}\right) = \hat{r}_j^{test} + e_j^{test} \tag{16}$$

The model performance can be evaluated based on the final prediction results $\left\{\tilde{r}_j^{test}\right\}$ and the test set RUL labels $\left\{r_j^{test}\right\}$.The proposed strategy, by introducing an error correction mechanism, helps reduce systematic bias and enhances the model's adaptability to complex

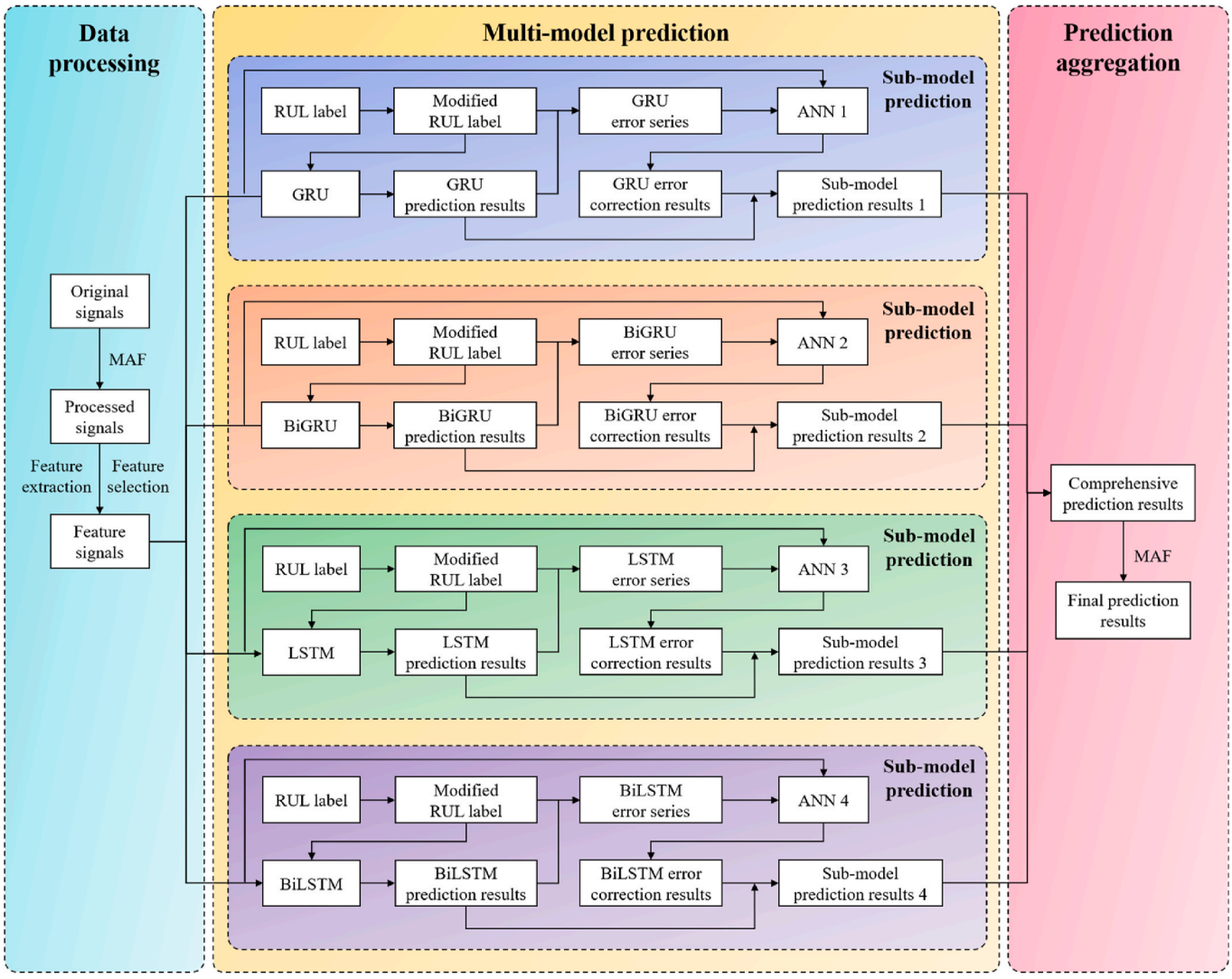



**Fig. 6.** The proposed RUL prediction framework.

degradation patterns, thereby improving the accuracy and robustness of bearing RUL prediction.

Given that ANN has good nonlinear fitting capabilities (Aladag et al., 2009), it is used as the error correction model in this study. The general structure of ANN is shown in Fig. 5(a). It primarily consists of an input layer, one or more hidden layers, and an output layer. Each layer comprises several neurons, and there are corresponding weight relationships between the neurons of adjacent layers. For an individual neuron (Fig. 5(b)), its output can be expressed as

$$y = f\left(\sum_{i=1}^{n} W_i x_i + b\right) \tag{17}$$

where $x$ are the inputs to the neuron, $W$ are the weights associated with each input, $b$ is the bias term, $f$ is the nonlinear activation function. The primary purpose of introducing $f$ is to enable neural networks to establish nonlinear relationship between input and output data. Rectified linear unit (ReLU) is computationally fast and can impart a certain level of sparsity to the hidden layer nodes of the network model, effectively alleviating the issue of overfitting during the training process. Therefore, this study adopts the ReLU function as the activation function. It is defined as:

## 4. The proposed RUL prediction framework

To achieve accurate RUL prediction of the rolling bearings, we propose a hybrid approach named MAF-ECDELM, short for Moving Average Filtering and Error Corrected Deep Ensemble Learning Model. In this framework, MAF (Moving Average Filter) refers to the filtering strategy applied during both data preprocessing and postprocessing of prediction results; EC (Error Correction) denotes the error correction mechanism designed to enhance the stability and accuracy of predictions; and DELM (Deep Ensemble Learning Model) represents the adopted deep ensemble learning architecture. The overall framework of the proposed method is illustrated in Fig. 6, is mainly divided into the following two steps.

(1) Data processing. Firstly, the MAF method is utilized to preprocess the original signals to effectively eliminate the interference of noise and outliers. Then, feature engineering is performed on the processed signals to extract and select useful features, highlighting the relationship between the feature signals and the RUL of the rolling bearings.
(2) Multi-model prediction. In this step, RUL labels are first constructed and reasonably modified. Then, several base learners (GRU, BiGRU, LSTM, and BiLSTM) are trained using the training feature signals and the corresponding RUL labels, enabling them to produce initial predictions for the test input features. Based on

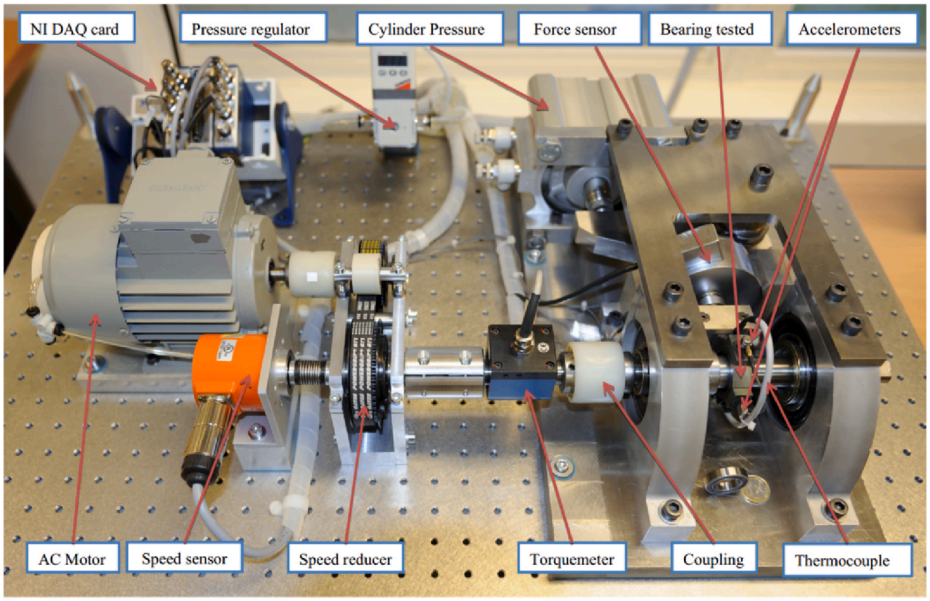



**Fig. 7.** PRONOSTIA test bench.

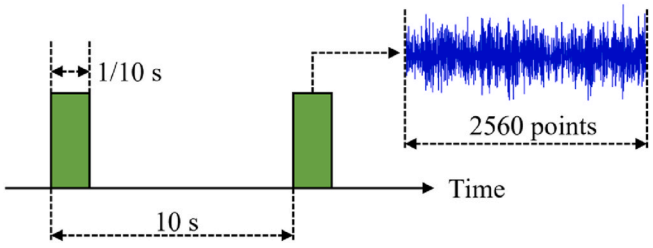



**Fig. 8.** The schematic diagram of the vibration signal collection.

the difference between the base learners' predictions on the training data and the actual RUL values, the train error series is obtained. This error series, along with the training feature signals, is used to train an error correction model (ANN), which learns to estimate the prediction error based on test input features. Finally, for test input features, the initial predictions from the base learner are combined with the corresponding output of the error correction model to obtain the final predictions of each sub-model.

(3) Prediction aggregation: To improve the accuracy of RUL prediction for bearings, this study performs a weighted aggregation of the prediction results from four sub-models to generate an overall prediction. Specifically, let the predictions of the four sub-models for the $i$-th sample be $\hat{y}_{i1}, \hat{y}_{i2}, \hat{y}_{i3}, \hat{y}_{i4}$. The aggregated prediction is computed as:

$$\hat{y}_i = \sum_{j=1}^{4} \omega_j \cdot \hat{y}_{ij} \tag{19}$$

where $\omega_j$ represents the weight assigned to the $j$-th sub-model, subject to the constraint $\sum_{j=1}^{4} \omega_j = 1$. In this study, a simple average weighting is adopted, i.e., $\omega_j = 0.25$, which is a widely used and easy-to-implement ensemble method. To further reduce short-term prediction fluctuations and enhance the stability of the overall prediction, MAF is applied to the aggregated results to produce the final smoothed prediction results.

To assess the performance of the proposed MAF-ECDELM method, three commonly used evaluation metrics in prediction tasks are considered: mean square error (*MSE*), mean absolute error (*MAE*) and R-square coefficient of determination ($R^2$). *MSE* is more sensitive to large deviations and highlights the impact of significant errors on model performance. *MAE*, with lower sensitivity to outliers, reflects the model's overall stability across the entire sample space. $R^2$ provides a dimensionless measure of goodness-of-fit, indicating how well the predicted values explain the variance in the actual values. These three metrics are complementary and jointly provide a comprehensive evaluation of the model's predictive performance. Therefore, they are collectively used to assess the effectiveness of the proposed RUL prediction model. Their expressions are as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{20}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{21}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (\bar{y} - y_i)^2} \tag{22}$$

where $\hat{y}_i$ and $y_i$ represent the predicted RUL value and true RUL value of the $i$-th sample, respectively, $\bar{y}$ represents the mean of the true values, and $N$ represents the total number of test samples. The smaller the *MSE* and *MAE*, the closer the $R^2$ is to 1, indicating that the performance of the prediction model is better.

## 5. Case study

### 5.1. Case 1: IEEE PHM2012 bearing

In this case study, the full-lifecycle data from the IEEE PHM2012 bearing prognostics challenge (Nectoux et al., 2012) is utilized to test the MAF-ECDELM method. The data comes from the PRONOSTIA test bench, shown in Fig. 7. The test bench consists of multiple components, including a motor, bearings, a coupling, accelerometers, data acquisition cards, etc. During the degradation simulation experiment, the vibration signals of the bearing are acquired by accelerometers in both horizontal and vertical directions. The sensor sampling frequency is set

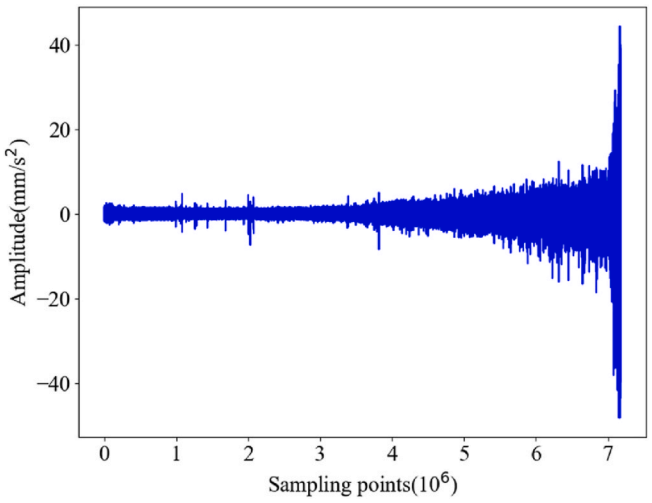



**Fig. 9.** The time domain diagram of bearing full-life cycle vibration data.

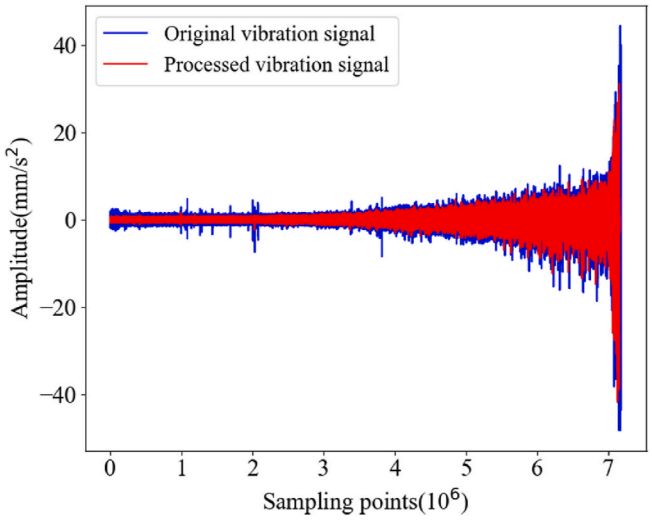



**Fig. 10.** Comparison of original vibration signals and processed vibration signals.

to 25.6 kHz to effectively cover the frequency range associated with mechanical faults and ensure that high-frequency degradation signals are not lost. Data are collected every 10 s, with each sampling lasting 0.1 s (i.e., 2560 data points), which balances data volume and analysis complexity while capturing the evolving trends of bearing degradation. The schematic diagram illustrating the vibration signal collection process is presented in Fig. 8.

Since the vibration signals in the horizontal direction contain more state information for bearings (Soualhi et al., 2014), this case study focuses on the horizontal vibration signals of Bearing 1 under Condition 1 (with a rotational speed of 1800 rpm and a load of 4000 N). All vibration samples in the bearing dataset are concatenated in the order of acquisition by linking the end of one sample's signal to the beginning of the next. This results in the time-domain diagram of the bearing full

**Table 1**
Expressions of different time domain features.

| Time domain feature | Expression | $\rho_s$ |
|---|---|---|
| Maximum value | $X_{\max} = \max\{x_i\}$ | 0.9742 |
| Minimum value | $X_{\min} = \min\{x_i\}$ | −0.9697 |
| Mean | $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ | 0.0034 |
| Variance | $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})^2$ | 0.9996 |
| Standard deviation | $\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})^2}$ | 0.9996 |
| Peak-to-peak value | $X_{pp} = \max\{x_i\} - \min\{x_i\}$ | 0.9828 |
| Root mean square | $X_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$ | 1.0000 |
| Kurtosis | $\beta = \frac{1}{N}\sum_{i=1}^{N} x_i^4$ | 0.9984 |
| Square root amplitude | $X_r = \left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert x_i\rvert}\right)^2$ | 0.9987 |
| Waveform indicator | $S_f = \frac{X_{rms}}{\lvert\overline{X}\rvert}$ | 0.4953 |
| Kurtosis indicator | $K_f = \frac{\beta}{X_{rms}^4}$ | 0.8287 |
| Absolute average amplitude | $\lvert\overline{X}\rvert = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert$ | 0.9995 |
| Skewness indicator | $P = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i^3}{X_{rms}^3}$ | −0.2897 |
| Peak indicator | $C_f = \frac{X_{\max}}{X_{rms}}$ | 0.4802 |
| Pulse indicator | $I_f = \frac{X_{\max}}{\lvert\overline{X}\rvert}$ | 0.5374 |
| Margin indicator | $CL_f = \frac{X_{\max}}{X_r}$ | 0.6235 |

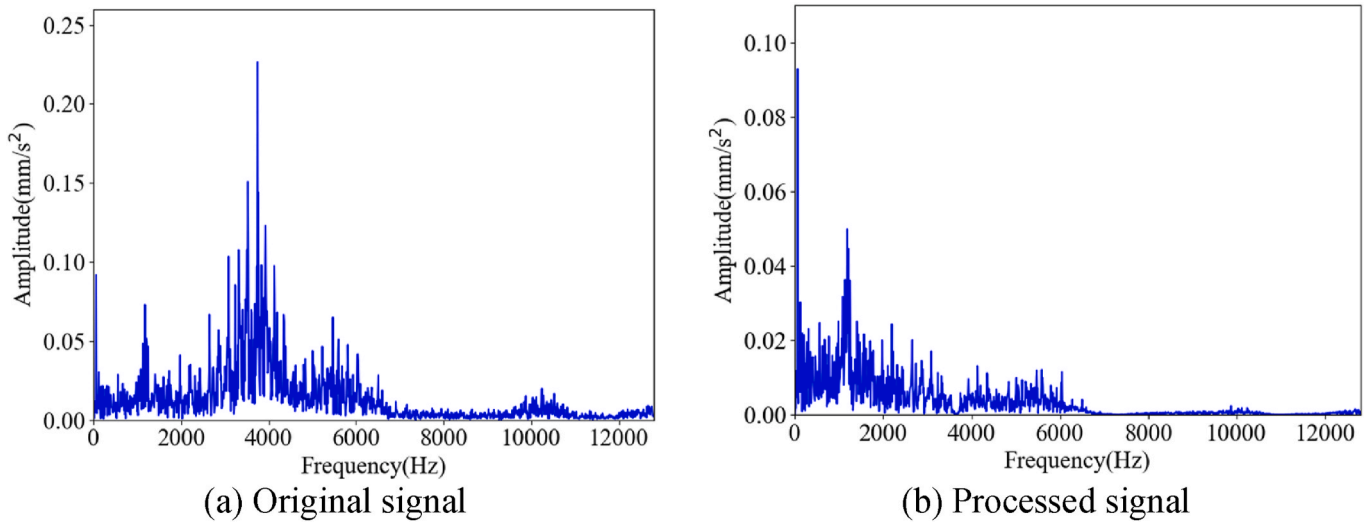



**Fig. 11.** Frequency domain diagrams of the original and processed vibration signals.

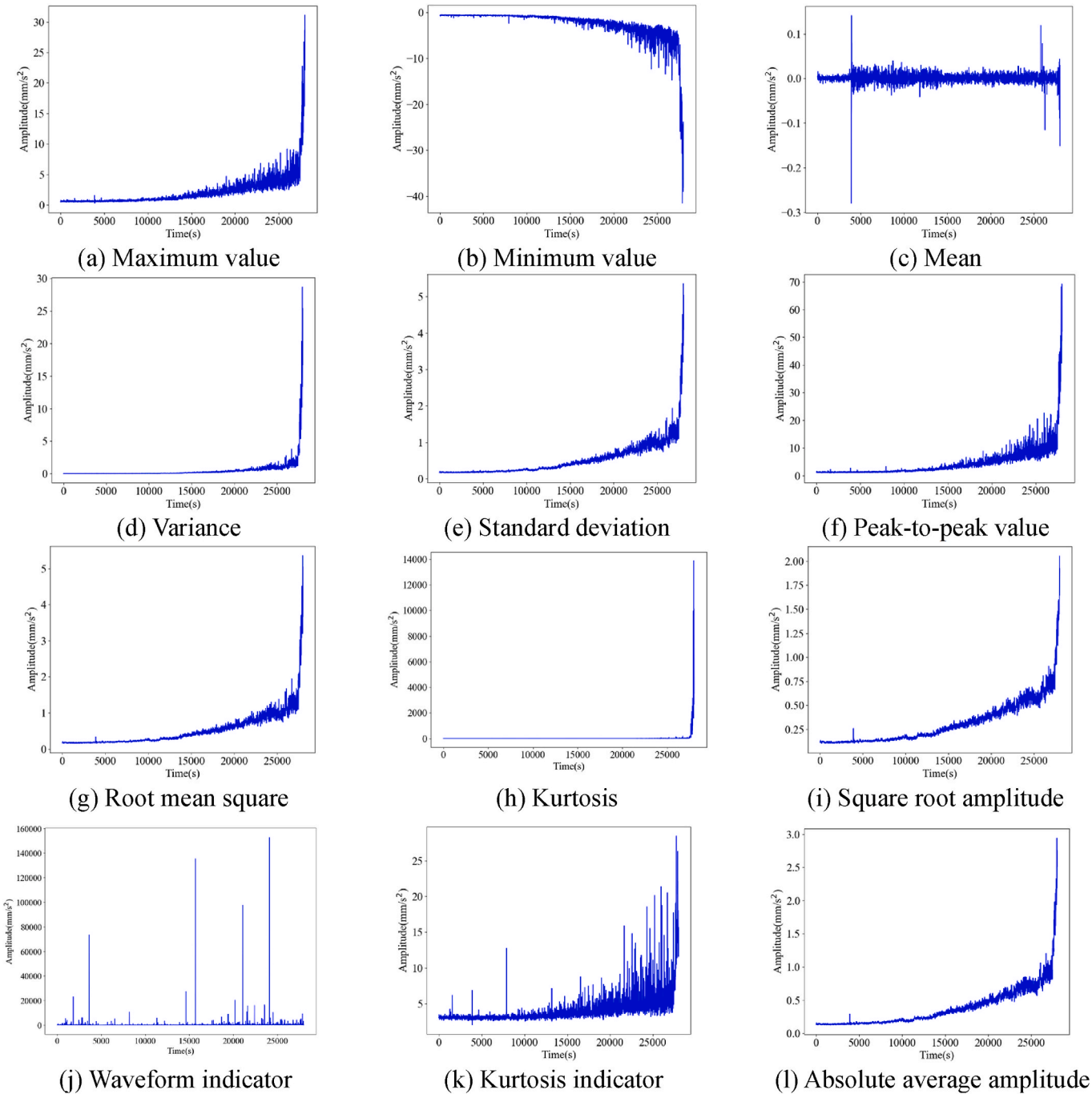



**Fig. 12.** Performance degradation curves formed by different time domain features.

life-cycle vibration data, as shown in Fig. 9. The figure clearly illustrates the trend in vibration amplitude from the initial healthy state to the final failure stage. In the early stage of degradation, the bearing vibration signals exhibit a stable trend with no significant changes in amplitude range. As the performance deteriorates, the amplitude range of the vibration signals gradually increases, with the most noticeable change occurring during the failure phase.

*5.1.1. Data processing analysis and feature selection*

First, the MAF method is applied to process the original vibration signal. During the processing, the window length of MAF is determined to be 8 time steps after repeated trials. The original vibration signals and the processed vibration signals are illustrated in Fig. 10. The processed signals are smoother than the original signals. The MAF method effectively removes a large amount of outlier data from the original vibration signals while retaining the valid information in the original vibration signals. This positive effect is particularly evident for the vibration signals in the early stage of degradation.

To highlight the denoising effect of the MAF method, a specific sample of bearing signals is selected for analysis. During the analysis, the original vibration signals and the processed vibration signals are converted from the time domain to the frequency domain, to facilitate the intuitive analysis of the noise reduction effect. Fig. 11 shows the comparison of the frequency domain diagrams before and after processing with the MAF method. From the Figure, it can be seen that the MAF method can filter out the noise in the high-frequency part to a certain

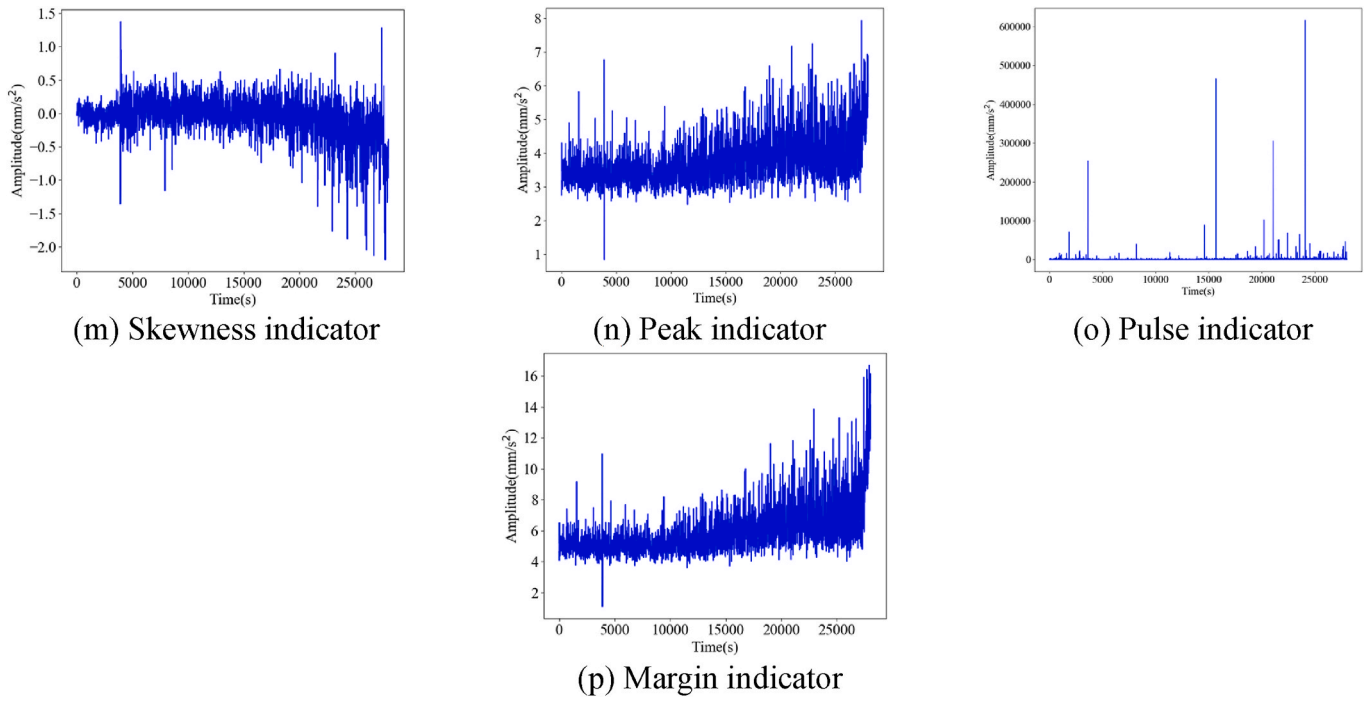



**Fig. 12.** (*continued*).

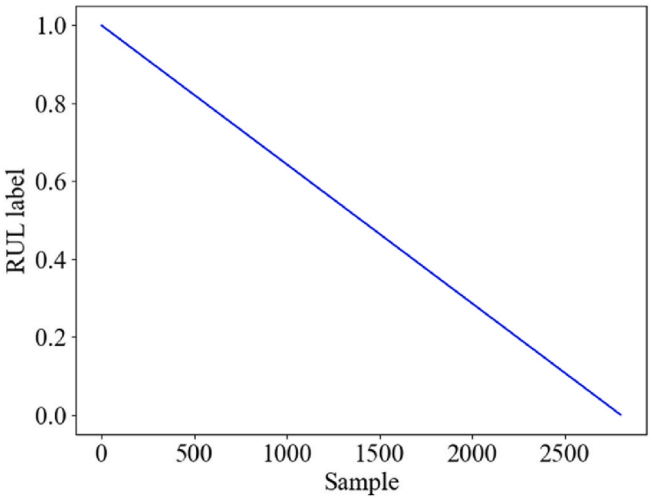



**Fig. 13.** RUL labels of the bearing dataset for Case 1.

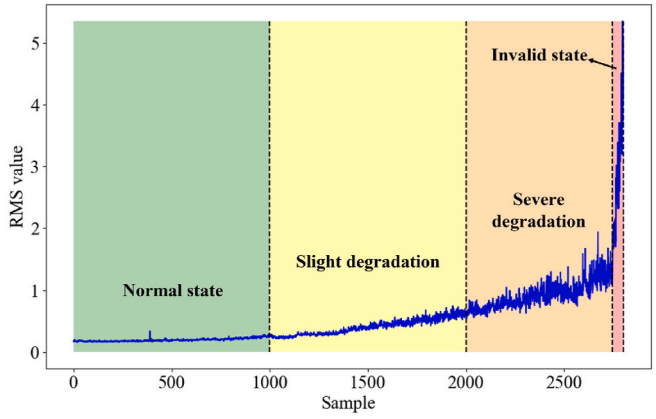



**Fig. 14.** Different degradation stages of bearings in Case 1.

extent, while preserving the valid information in the low-frequency part. This makes it more conducive to extracting sensitive features that characterize the degradation state of the bearing.

In RUL prediction tasks, the selection and optimization of health indicators (HIs) are critical to ensuring model performance. In this study, 16 time-domain features commonly used in bearing condition monitoring and fault diagnosis were extracted as candidate HIs, as shown in Table 1. The various HIs were connected in chronological order to obtain the performance degradation curves at different operating times, as shown in Fig. 12. Among them, the root mean square (RMS) value of the time-domain signal characterizes signal intensity and is highly sensitive to the progression of mechanical faults. It is widely used to describe the degradation process of mechanical systems and is one of the most common HIs for bearing RUL prediction (Li et al., 2015) (Huang et al., 2015). Therefore, this study calculated the Spearman Rank Correlation Coefficient between each time-domain feature and RMS to identify features that effectively capture the degradation trend. The Spearman coefficient $\rho_s$ is suitable for assessing nonlinear monotonic relationships between variables and is defined as follows:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{23}$$

where $d_i$ denotes the rank difference. The coefficient ranges from $-1$ to 1, with larger absolute values indicating stronger monotonic associations. The correlation threshold was set to 0.95 in this study to select features with statistically significant correlations. The $\rho_s$ values corresponding to different parameters are presented in Table 1. Based on this criterion, nine features were ultimately selected as inputs for the RUL prediction model: maximum value, minimum value, variance, standard deviation, peak-to-peak value, root mean square, kurtosis, square root amplitude, and absolute average amplitude.

### 5.1.2. RUL labels construction and dataset division

In this case study, the full-lifecycle bearing dataset consists of a total of 2803 samples, with each sample collected at a 10-s interval. Conse-

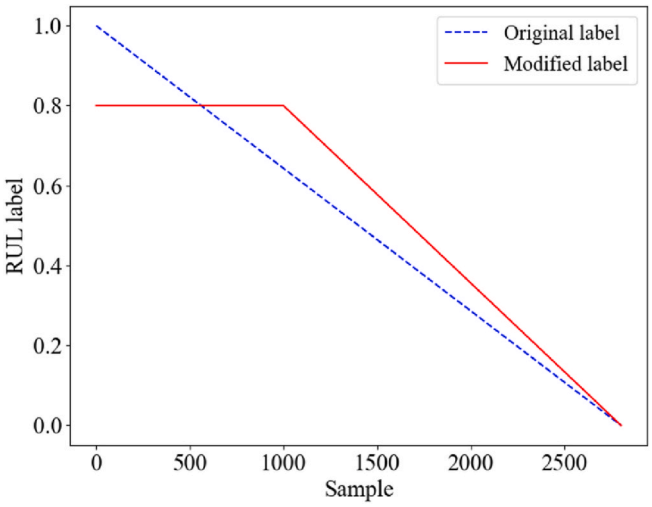



**Fig. 15.** RUL labels before and after modification in Case1.

quently, the total lifespan of the bearing is 28030 s. To facilitate the fitting of the RUL labels by the neural network prediction model, this study assigns a value of 1 to represent the bearing in a fully healthy state and a value of 0 to represent the bearing reaching complete failure. For instance, for the 1000th sample, its corresponding RUL is 18030s, and according to Equation (24), the RUL label of this sample can be calculated to be 18030/28030 = 0.643239. Following this logic, the RUL labels for the entire dataset can be constructed as shown in Fig. 13.

$$\text{RUL label} = \frac{\text{Remaining life}}{\text{Total life}} \tag{24}$$

Typically, the degradation of a system does not become apparent until the equipment has been in operation for some time. Setting the RUL labels to monotonically decrease from the beginning of operation may not be reasonable. Some studies have modified the monotonically decreasing RUL labels into piecewise linear functions, which have been proven to have good predictive performance (Heimes, 2008) (Sateesh et al., 2016). Therefore, this study modifies the RUL labels in Fig. 13, which show an overall monotonically decreasing trend, into a piecewise linear function. This means that the RUL labels remain unchanged before a certain critical point and exhibit a monotonically decreasing trend after the critical point.

This study uses RMS as a degradation indicator to divide the different degradation stages of bearings. The performance degradation curve formed by the RMS indicator of the full-lifecycle bearing data after processing by the MAF method is depicted in Fig. 14. RMS effectively characterizes the degradation state of the bearing. Based on the growth rate and range of the RMS values, the degradation process of the bearing can be primarily divided into four stages: from the 1st sample to the

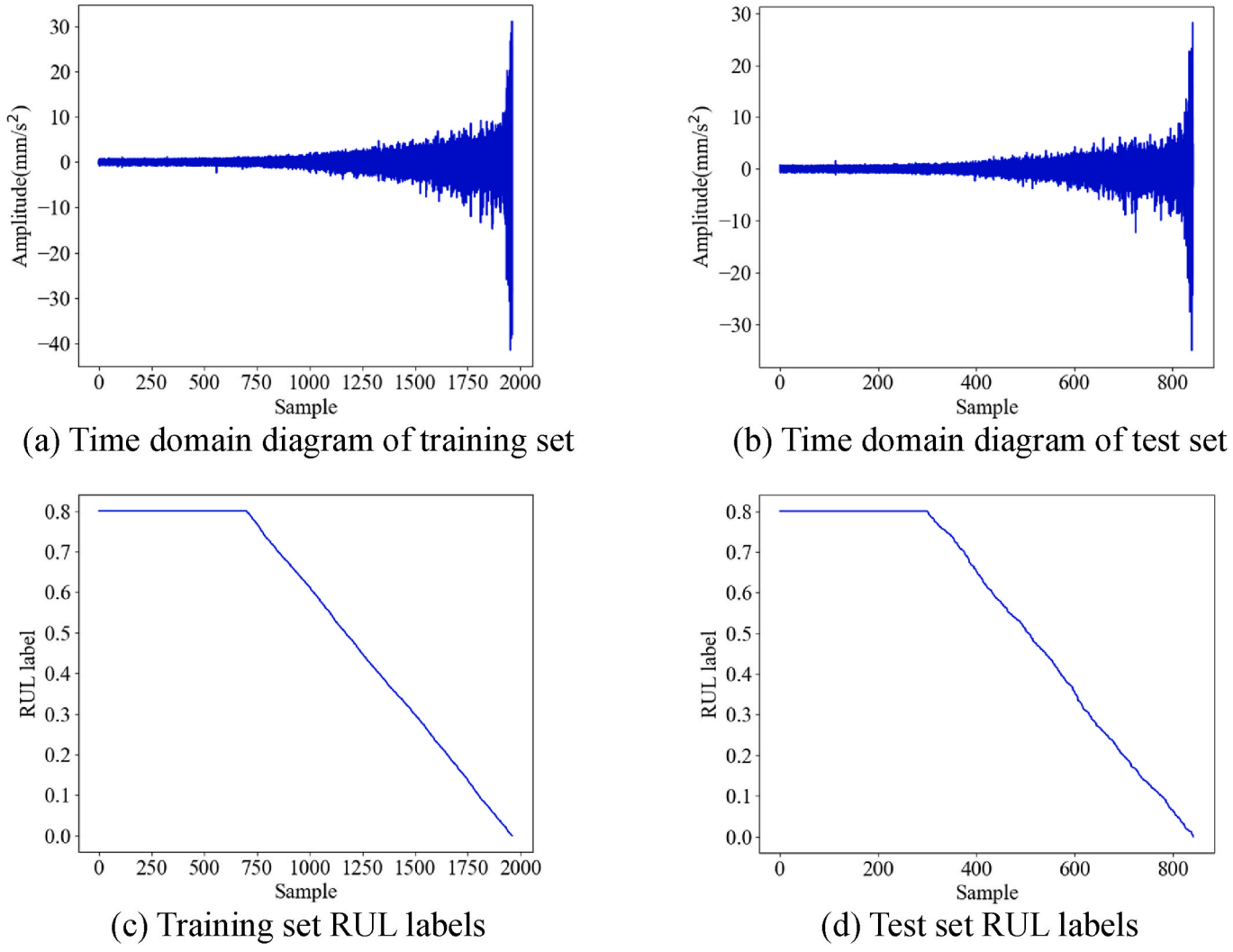
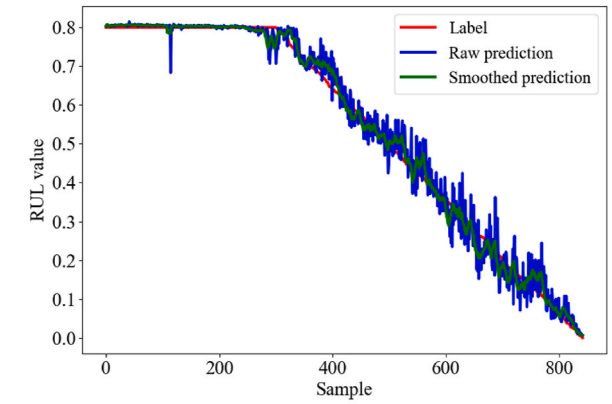
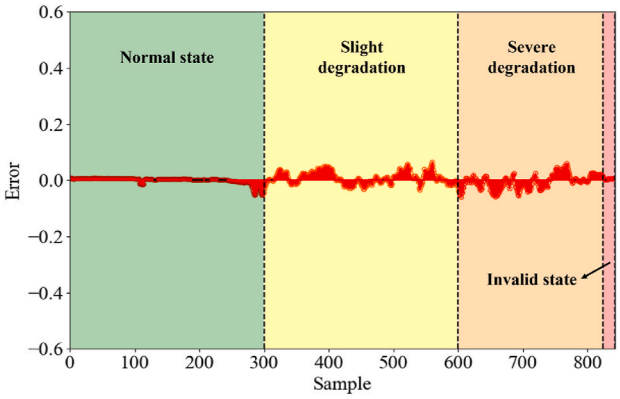



**Fig. 16.** Vibration data and RUL labels of the training and test sets in Case 1.




(a) True labels and predicted values of the proposed method




(b) The errors between the true labels and the predicted values

**Fig. 17.** Prediction performance of the proposed method on the test set in Case1.

**Table 2**
Comparison results of different methods.

| Method | *MSE* | *MAE* | $R^2$ |
|---|---|---|---|
| SVR | 0.043507 | 0.178450 | 0.394695 |
| ANN | 0.002527 | 0.035659 | 0.964837 |
| GRU | 0.002188 | 0.031259 | 0.969561 |
| BiGRU | 0.002122 | 0.031183 | 0.971090 |
| LSTM | 0.002181 | 0.031249 | 0.969650 |
| BiLSTM | 0.002126 | 0.031204 | 0.970423 |
| MAF-GRU | 0.001133 | 0.023497 | 0.984237 |
| MAF-ECGRU | 0.001050 | 0.021806 | 0.985385 |
| MAF-DELM | 0.001023 | 0.020924 | 0.985762 |
| MAF-Transformer | 0.001175 | 0.022713 | 0.983651 |
| WPD-ECDELM | 0.001857 | 0.027674 | 0.974159 |
| MAF-ECDELM1 | 0.001278 | 0.024842 | 0.982221 |
| MAF-ECDELM2 | 0.001099 | 0.021854 | 0.984709 |
| MAF-ECDELM3 | 0.000990 | 0.020861 | 0.986231 |
| Proposed method | 0.000380 | 0.013695 | 0.994716 |

1000th sample, the bearing is in a normal state; from the 1001th sample to the 2000th sample, it is in a slight degradation stage; from the 2001th sample to the 2745th sample, it is in a severe degradation stage; from the 2746th sample to the 2803rd sample, it is in an invalid state.

Based on the above analysis, this study sets the critical point for bearing degradation at 1000th sample, and the RUL label for the healthy state is set to 0.8. This means that the value of the RUL label remains at 0.8 until the 1000th sample, after which it exhibits a monotonically decreasing trend. When the bearing reaches complete failure, the value of the RUL label is 0. Fig. 15 illustrates the RUL labels before and after the piecewise modification.

To preserve the integrity of time correlation, the time series data is divided based on the staged characteristics of bearing degradation. Specifically, for the four degradation stages (normal state, slight degradation stage, severe degradation stage and invalid state), 70 % of the samples are independently selected as the training set, while the remaining 30 % are assigned to the test set, ensuring that there is no temporal overlap between training and testing data. After the division, the training set and the test set contain 1961 and 842 samples, respectively. Fig. 16(a) and (b) present the time domain diagrams of the vibration data for the training set and the test set, respectively. Fig. 16(c) and (d) show the RUL labels for the training set and the test set, respectively.

### 5.1.3. Analysis of prediction results

The training set is utilized to optimize the parameters of the prediction model, whereas the test set is employed to assess the performance of the prediction model. The hyperparameters are set through repeated trials: the number of hidden layer units for the base learners GRU, BiGRU, LSTM and BiLSTM is set to 32. Each error correction model ANN contains two hidden layers, with 64 and 32 neurons, respectively. The optimization algorithm used during training is the Adam algorithm, with a learning rate set to 0.01, and the number of training iterations is set to 500. When MAF smooths the prediction results, it takes the average value of the predictions over a certain period as the prediction value at the end of the period. To eliminate the effects of randomness, the final test results are the mean value of five repeated trials. The experiments are conducted on a Windows 10 operating system with hardware comprising an Intel Core i9-10900 (10-core) processor and 48 GB of RAM. The code is implemented using the TensorFlow framework in Python 3.9. Under this setup, the average prediction time per test sample in this case study is approximately 0.0132 s. The prediction results of the MAF-ECDELM method on the test set is given in Fig. 17.

Fig. 17(a) illustrates the true labels of the test set, the raw prediction values and the smoothed prediction values. The MAF method effectively reduces the abnormal fluctuations in the prediction results, making the overall trend of the prediction results more consistent with the overall trend of the true labels. Fig. 17(b) shows the error results between the true labels and the prediction values: the proposed MAF-ECDELM method achieves the smallest prediction errors and best prediction performance during the normal and invalid state stages of bearing degradation. This is primarily because the sensitive features within the normal and invalid state stages exhibit distinct differences. In contrast, the sensitive features during the slight and severe degradation stages change slowly in a nonlinear manner, making it more challenging to establish a mapping relationship with the linearly changing true labels. Therefore, the prediction performance of the MAF-ECDELM method is slightly poorer during these two stages. Overall, the MAF-ECDELM method achieves excellent prediction performance on the test set, with the *MSE*, *MAE* and $R^2$ metrics reaching 0.000380, 0.013695 and 0.994716, respectively.

### 5.1.4. Comparative analysis of different methods

To value the performance of the MAF-ECDELM method, extensive comparisons are made with several methods, which include the following: (1) Support vector regression (SVR) (Platt, 1999): the kernel function of SVR is the radial basis function; (2) ANN; (3) GRU; (4) BiGRU; (5) LSTM; (6) BiLSTM; (7) MAF-GRU: this method processes the original data using MAF and predicts using GRU; (8) MAF-ECGRU: this method uses GRU with error correction as the prediction model, and the error correction model is the same as that of the proposed method; (9) MAF-DELM: this method uses a deep ensemble learning model composed of GRU, BiGRU, LSTM and BiLSTM as the prediction model; (10) MAF-Transformer: This method uses Transformer as the prediction model. The structure of Transformer refers to references (Kim et al., 2024) and (Chen et al., 2022), that is, the encoder part is the main structure, the number of layers is 1, the number of multi-head attention

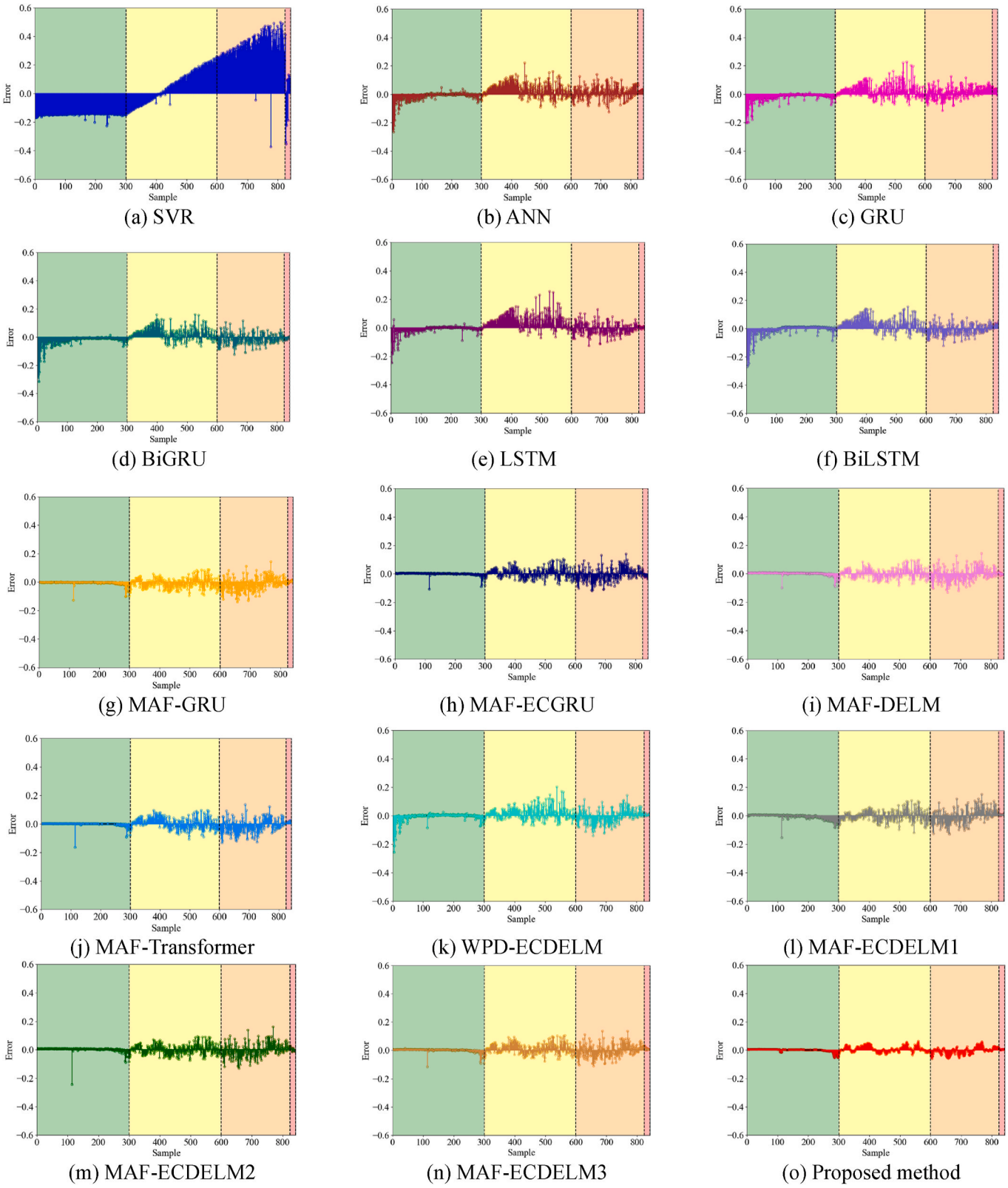



**Fig. 18.** Prediction error results of different methods.

heads is 5, and the number of hidden layer units is 64. (11) WPD-ECDELM: this method uses the wavelet packet decomposition (WPD) for denoising the original data. WPD employs the soft threshold method with the threshold set to the median of the wavelet coefficients; the prediction model is DELM with error correction; (12) MAF-ECDELM1: this method has the same prediction model as the proposed method but only uses the RMS value as the input feature; (13) MAF-ECDELM2: this method uses all the time-domain features listed in Table 1 as input features; (14) MAF-ECDELM3: this method does not perform smoothing on the prediction results, and the other parts are

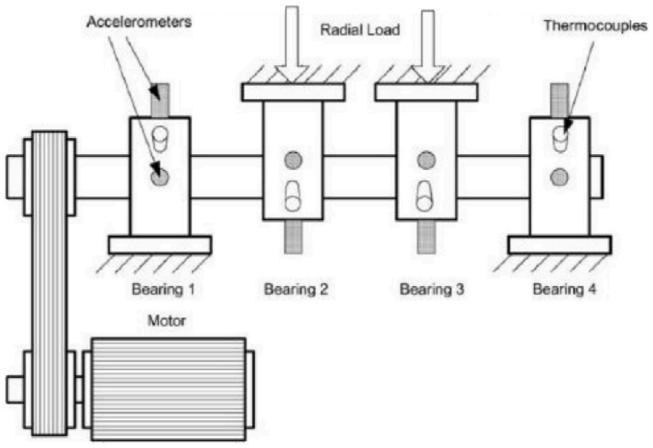



**Fig. 19.** IMS bearing test bench.

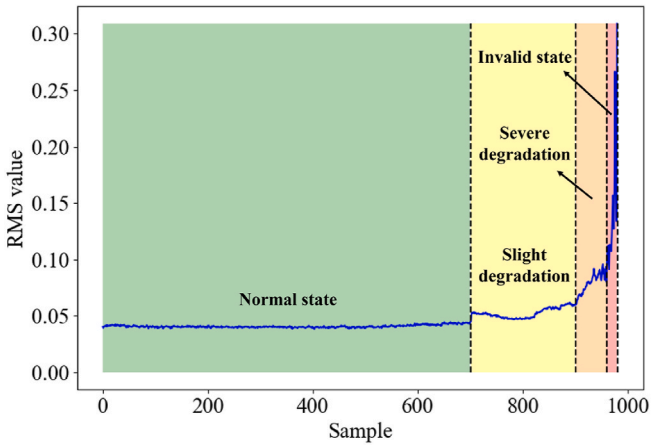



**Fig. 22.** Different degradation stages of bearings in Case 2.

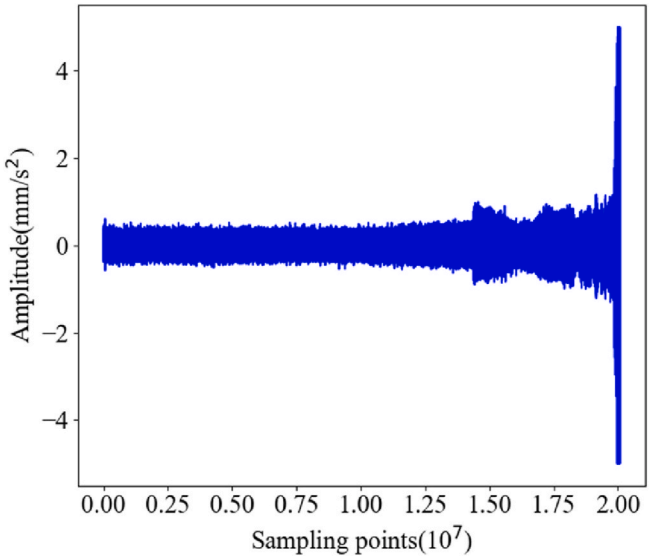



**Fig. 20.** Time domain diagram of the full lifecycle vibration data of bearing in Case 2.

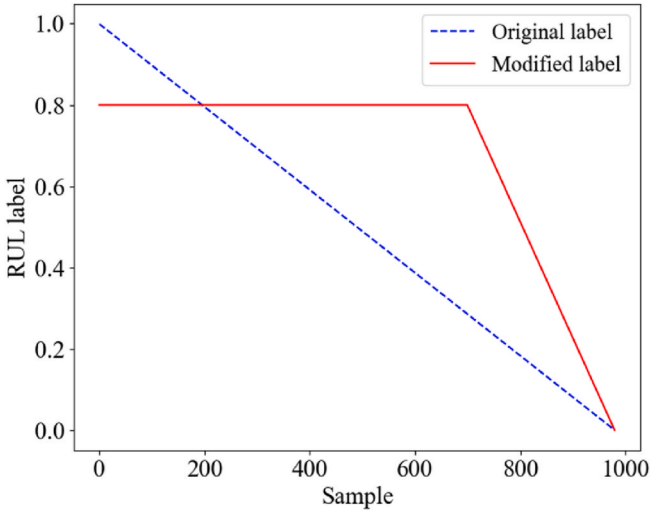



**Fig. 23.** RUL labels before and after piecewise modification in Case2.

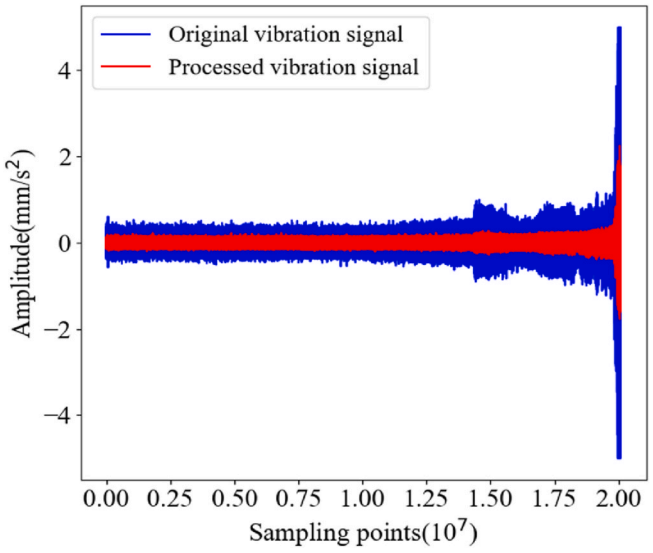



**Fig. 21.** Comparison of original vibration signal and processed vibration signal.

consistent with the proposed method. Except for MAF-ECDELM1 and MAF-ECDELM2, the input features of the other methods are consistent with those of the proposed method. The structure and setting for training of the ANN, GRU, BiGRU, LSTM and BiLSTM models are the same as those of the proposed method. Table 2 and Fig. 18 provide the prediction performance of different methods on the test set.

Compared to the methods that use MAF for processing, those that do not employ MAF generally perform worse. In particular, they exhibit larger errors in predicting the RUL of healthy state data. This is mainly because the original signals in this degradation stage contain a significant amount of outliers and noise, which interfere with the model's prediction performance, leading to a noticeable decrease in accuracy. The MAF method effectively eliminates these interferences and improves the performance of the prediction model, which further shows the effectiveness of the proposed data processing method. The comparison results between WPD-ECDELM and MAF-ECDELM method indicate that for the RUL prediction task, the MAF method outperforms WPD in data processing. This is primarily because the denoising effect of WPD relies on the effective selection of thresholds, which in turn depends on extensive expert prior knowledge. The comparison results between MAF-GRU, MAF-DELM, MAF-ECGRU and the proposed method indicate that relative to single models, the proposed deep ensemble learning model exhibits stronger generalization performance than the other methods. Furthermore, the error correction strategy can enhance the performance of the prediction model even further. The comparison

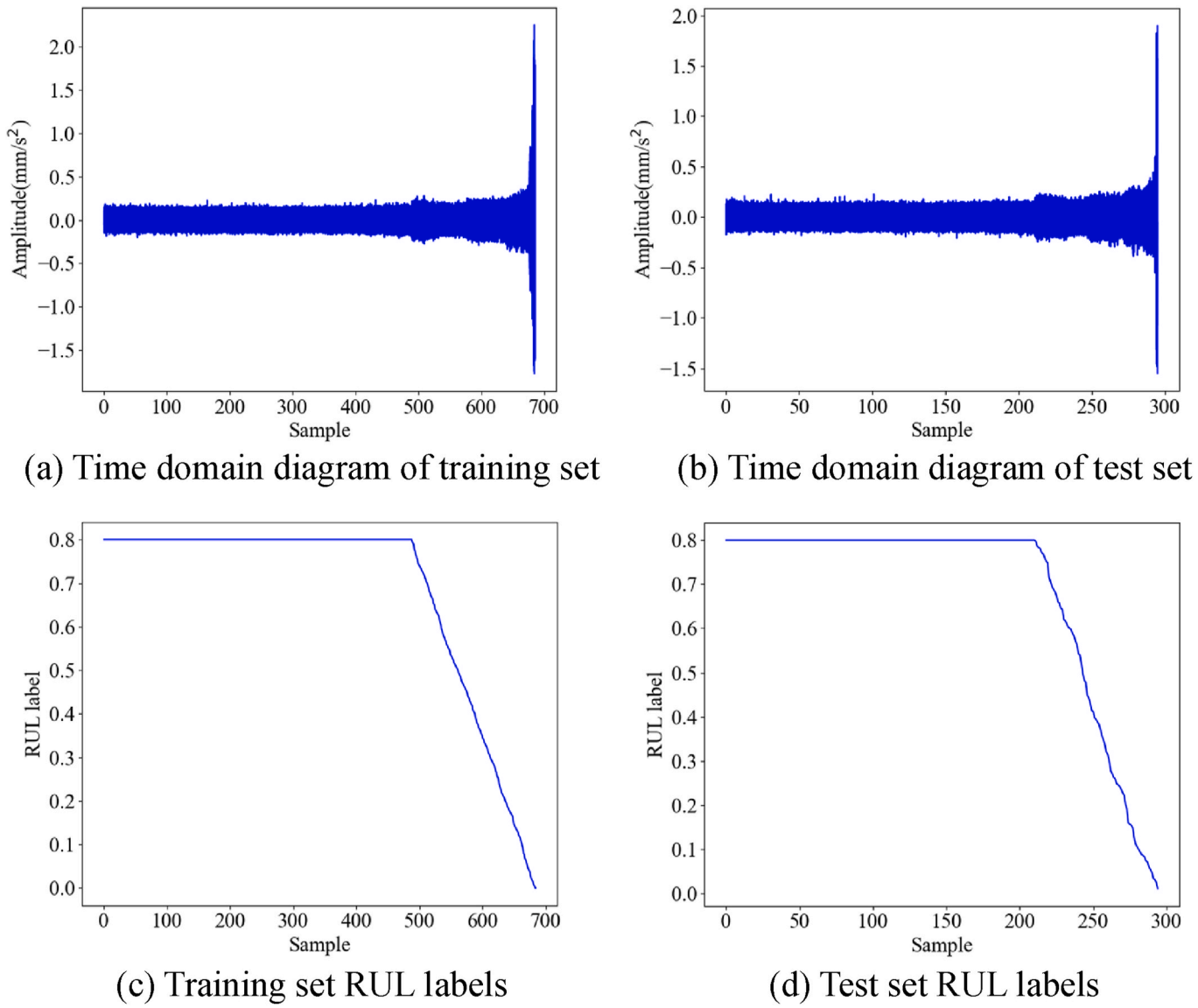



**Fig. 24.** Vibration data and RUL labels of the training and test sets in Case 2.

among MAF-GRU, MAF-DELM, and MAF-Transformer shows that in this case, the prediction performance of the Transformer is inferior to that of the ensemble learning model and even lower than that of GRU. This may be attributed to the complexity of the Transformer architecture and its reliance on large-scale data to fully leverage its global modeling capabilities. When faced with limited samples, Transformers are more prone to convergence difficulties and overfitting, resulting in weaker generalization compared to the more streamlined and temporally robust GRU model.

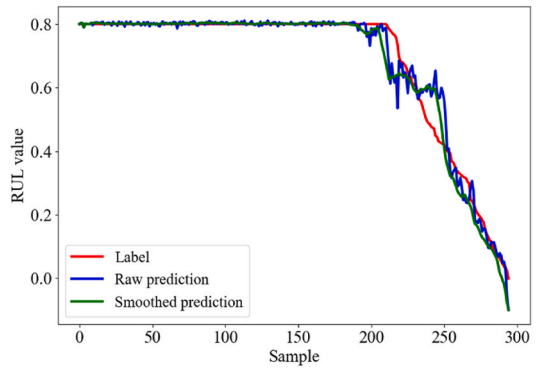
The superiority of MAF-ECDELM2 over MAF-ECDELM1 shows that enhancing the richness of feature information can help improve the performance of the prediction model. However, it is necessary to effectively screen for sensitive features, as irrelevant features can hinder the model's ability to fit the data. This is also the reason why MAF-ECDELM2 performs worse than MAF-ECDELM3 and the proposed method. The comparison results between MAF-ECDELM3 and the proposed method indicate that the MAF method can effectively reduce the abnormal fluctuations in the prediction results, making the predictions follow more closely the trend of the true labels. In summary, the MAF-ECDELM method achieves the lowest *MSE*, *MAE* and the highest $R^2$ in the comparative tests, achieving ting the best performance among the different methods considered.

### 5.2. Case 2: IMS bearing

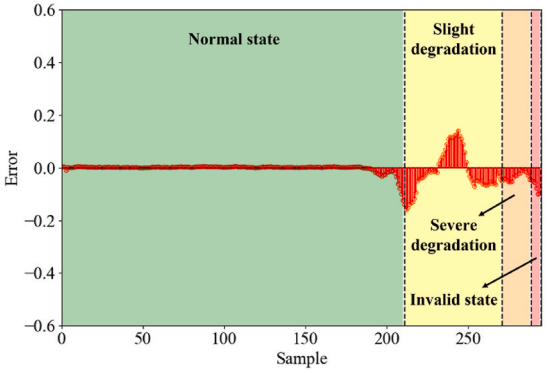
To further validate the effectiveness of the MAF-ECDELM method, it is applied to the full lifecycle data of rolling bearings from the Intelligent Maintenance System (IMS) Center at the University of Cincinnati (Qiu et al., 2006). The bearing data come from the test bench depicted in Fig. 19, where the vibration signals of the bearings are acquired by accelerometers. The sampling frequency of the sensors is 20 kHz and data are collected every 10 min, with each collection comprising 20480 data points. At the end of the experiment, three datasets were obtained. For this case study, the focus is on the full life cycle vibration data of bearing 1 from dataset 2, which contains a total of 984 samples. Since the amplitude of the last few samples has been to zero, the first 980 samples are taken to construct the research dataset. The time domain diagram of the full lifecycle vibration data of the bearing is shown in Fig. 20.

#### 5.2.1. Data processing analysis and feature selection

As the Case 1, first the MAF method is applied to process the original signals, resulting in the processed vibration signals illustrated in Fig. 21. The MAF method effectively removes outlier data. Notably, the original vibration signals exhibit a clear fluctuating trend during a certain degradation phase, which makes it difficult to establish a relationship




(a) True labels and predicted values of the proposed method




(b) The error results between the true labels and the predicted values

**Fig. 25.** Prediction performance of the proposed method on the test set in Case2.

**Table 3**
Comparison results of different methods.

| Method | *MSE* | *MAE* | $R^2$ |
|---|---|---|---|
| SVR | 0.009854 | 0.059450 | 0.797887 |
| ANN | 0.006082 | 0.045785 | 0.875267 |
| GRU | 0.005308 | 0.042573 | 0.891139 |
| BiGRU | 0.004674 | 0.037232 | 0.904132 |
| LSTM | 0.005281 | 0.040968 | 0.891682 |
| BiLSTM | 0.004790 | 0.038501 | 0.901763 |
| MAF-GRU | 0.001458 | 0.018255 | 0.970098 |
| MAF-ECGRU | 0.001269 | 0.017288 | 0.973982 |
| MAF-DELM | 0.001298 | 0.017642 | 0.973386 |
| MAF-Transformer | 0.001482 | 0.022851 | 0.969602 |
| WPD-ECDELM | 0.004968 | 0.037391 | 0.898110 |
| MAF-ECDELM1 | 0.003267 | 0.033069 | 0.932986 |
| MAF-ECDELM2 | 0.001464 | 0.019526 | 0.969967 |
| MAF-ECDELM3 | 0.001182 | 0.015949 | 0.975750 |
| Proposed method | 0.001056 | 0.015403 | 0.978346 |

between this segment of data and the monotonically trending RUL labels. However, the MAF method significantly reduces the amplitude of the fluctuations during this degradation phase, which facilitates the prediction model fitting the mapping relationship between vibration data and RUL labels. Similarly to the previous Case 1, after data preprocessing, the following nine-time domain features are extracted as the input of the RUL prediction model: maximum value, minimum value, variance, standard deviation, peak-to-peak value, root mean square, kurtosis, square root amplitude and absolute average amplitude.

#### 5.2.2. RUL labels construction and dataset division

Similarly to Case 1, the degradation stages of the bearing are delineated using the RMS value as the degradation indicator. The performance degradation curve formed by the RMS metric is shown in Fig. 22. Based on the rate of increase and the range of change in the RMS values, the degradation process of the bearing can be primarily divided into four stages: from the 1st sample to the 700th sample, the bearing is in the normal state; from the 701th sample to the 900th sample, it is in the slight degradation stage; from the 901th sample to the 960th sample, it is in the severe degradation stage; from the 961th sample to the 980th sample, it is in the invalid state. Therefore, the critical point of bearing degradation in this case is set to 700. Fig. 23 illustrates the RUL labels before and after being modified.

For each degradation stage, 70 % of the samples are independently selected as the training set, while the remaining 30 % are used as the test set. After the division, the training set and test set contain 685 and 295 samples, respectively. Fig. 24 (a) and 24(b) present the time domain diagrams of the vibration data for the training set and the test set, respectively. Fig. 24 (c) and 24(d) show the RUL labels for the training set and the test set, respectively.

#### 5.2.3. Analysis of prediction results

Under the previously mentioned hardware and software setup, the proposed method achieves an average prediction time of approximately 0.1026 s per test sample in this case study. Fig. 25 presents the prediction outcomes of the MAF-ECDELM method on the test set. As illustrated in Fig. 25(a), the smoothed prediction results exhibit fewer abnormal fluctuations, whose overall trend is similar to that of the true labels. As illustrated Fig. 25(b), the MAF-ECDELM method demonstrates good prediction performance for the three degradation stages: normal state, severe degradation and invalid state. The prediction error of the MAF-ECDELM method is larger during the slight degradation stage. This is mainly because the vibration data during the slight degradation stage exhibit fluctuating trends, making it more challenging for the model to establish a mapping relationship between them and the monotonically changing RUL labels. The *MSE*, *MAE* and $R^2$ values achieved by the MAF-ECDELM method on the test set reach 0.001056, 0.015403, and 0.978346, respectively.

#### 5.2.4. Comparative analysis of different methods

As the Case 1, the thirteen methods SVR, ANN, GRU, BiGRU, LSTM, BiLSTM, MAF-GRU, MAF-ECGRU, MAF-DELM, MAF-Transformer, WPD-ECDELM, MAF-ECDELM1, MAF-ECDELM2 and MAF-ECDELM3 are compared with the proposed method. The test results of each method are illustrated in Table 3 and Fig. 26. The methods not employing MAF generally perform worse than those that do use MAF, and the MAF method still outperforms WPD also in this case. The comparison results of MAF-GRU, MAF-DELM, MAF-ECGRU and the proposed method indicate that the proposed deep ensemble learning model exhibits stronger generalization performance than single models, and the error correction strategy can further enhance the performance of the prediction model. The comparison between MAF-DELM and MAF-Transformer indicates that, in this case, the ensemble learning model outperforms the Transformer in prediction performance. The comparison results of MAF-ECDELM1, MAF-ECDELM2, MAF-ECDELM3 and the proposed method demonstrate that enhancing the richness of effective feature information can enhance the performance of the prediction model, and the smoothing of prediction results by the MAF method also can enhance prediction performance. The MAF-ECDELM method still achieves the best prediction performance in the comparative test, confirming the superiority of the MAF-ECDELM method also in the Case 2.

### 5.3. Analysis of key parameters

This section analyzes several key parameters of the MAF-ECDELM method using the IEEE PHM2012 bearing dataset as a case study, aiming to investigate how different parameter combinations affect prediction performance. In this study, the MAF is employed for both data preprocessing and smoothing of prediction results, with the goal of eliminating noise and outliers, reducing short-term prediction

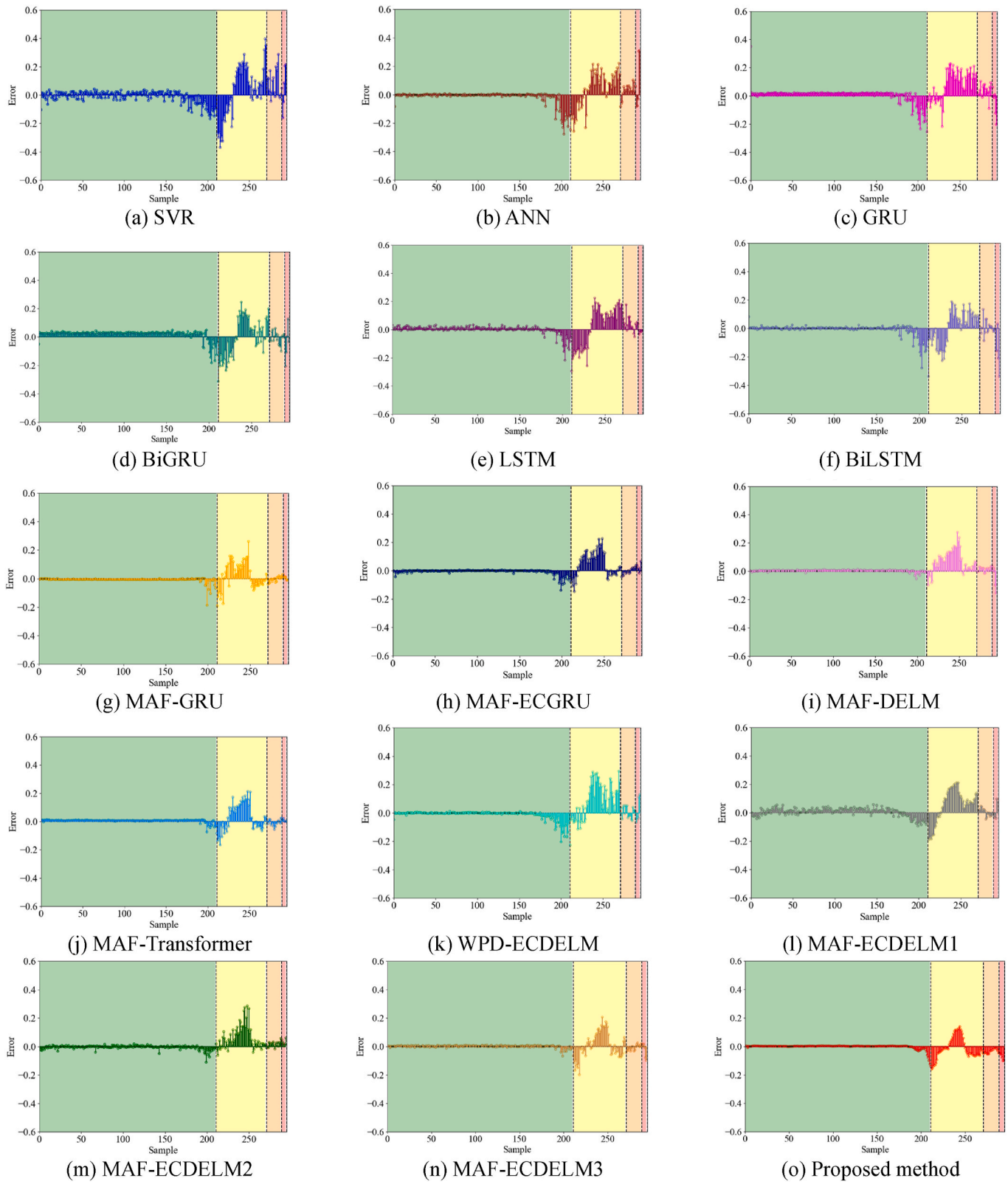



**Fig. 26.** Prediction error results of different methods.

fluctuations, and enhancing the overall stability of the predictions. Table 4 presents the prediction performance of the proposed method under different MAF window lengths, ranging from 4 to 12.

The results show that as the window length increases, model performance improves significantly at first and then stabilizes. Specifically, when the window length increases from 4 to 6, *MSE* decreases from 0.000868 to 0.000411, *MAE* decreases from 0.019939 to 0.014344, and $R^2$ increases from 0.987923 to 0.994278. This indicates that smaller windows are insufficient to smooth short-term prediction fluctuations, and predictions remain vulnerable to noise and outliers. When the

**Table 4**
Prediction performance of MAF-ECDELM method under different MAF window lengths.

| Window length | *MSE* | *MAE* | $R^2$ |
|---|---|---|---|
| 4 | 0.000868 | 0.019939 | 0.987923 |
| 5 | 0.000563 | 0.016302 | 0.992166 |
| 6 | 0.000411 | 0.014344 | 0.994278 |
| 7 | 0.000391 | 0.013834 | 0.994558 |
| 8 | 0.000380 | 0.013695 | 0.994716 |
| 9 | 0.000387 | 0.014039 | 0.994610 |
| 10 | 0.000394 | 0.014767 | 0.994521 |
| 11 | 0.000387 | 0.014328 | 0.994621 |
| 12 | 0.000395 | 0.014586 | 0.994506 |

**Table 5**
Prediction performance of MAF-ECDELM method under different hyperparameter combinations.

| Optimizer | Learning rate | *MSE* | *MAE* | $R^2$ |
|---|---|---|---|---|
| Adam | 0.1 | 0.000576 | 0.018293 | 0.991980 |
| Adam | 0.01 | 0.000380 | 0.013695 | 0.994716 |
| Adam | 0.001 | 0.000403 | 0.014586 | 0.994388 |
| SGD | 0.1 | 0.000635 | 0.019435 | 0.991163 |
| SGD | 0.01 | 0.000758 | 0.021180 | 0.989458 |
| SGD | 0.001 | 0.001349 | 0.029849 | 0.981232 |
| RMSprop | 0.1 | 0.001411 | 0.030576 | 0.980372 |
| RMSprop | 0.01 | 0.000430 | 0.015058 | 0.994017 |
| RMSprop | 0.001 | 0.000404 | 0.014563 | 0.994377 |

window length further increases to 7–12, the three metrics exhibit minor variations, and performance stabilizes. Notably, when the window length is 8 (the value adopted in this study), *MSE* reaches its lowest value (0.000380) and $R^2$ its highest (0.994716), suggesting that this window length strikes a good balance between suppressing local anomalies and preserving meaningful trends.

In addition, given that the optimizer and learning rate are two critical hyperparameters influencing machine learning model performance, this study evaluated the proposed method under combinations of three commonly used optimizers (Adam, SGD, RMSprop) and three typical learning rates (0.1, 0.01, 0.001). The results are shown in Table 5. Overall, the Adam optimizer outperformed the other two and achieved the best performance when the learning rate was set to 0.01 (the hyperparameter combination adopted in this study). This suggests that Adam offers an effective weight update strategy at this learning rate, balancing convergence speed and training stability. When the learning rate was either too high (0.1) or too low (0.001), model performance declined slightly, indicating that the learning rate has a certain influence on the convergence path of Adam.

In contrast, SGD showed the weakest overall performance, especially when the learning rate is 0.001, its *MSE*, *MAE*, and $R^2$ were 0.001349, 0.029849, and 0.981232, respectively. This may be attributed to SGD's lack of adaptive adjustment mechanisms, which makes it less effective at navigating complex loss surfaces. RMSprop demonstrated stable performance at learning rates of 0.01 and 0.001, with $R^2$ values of 0.994017 and 0.994377, respectively, which are comparable to those of Adam. However, its performance significantly dropped when the learning rate was set to 0.1 ($R^2 = 0.980372$), indicating that a large learning rate may destabilize the training process.

## 6. Conclusion

This study proposed a novel method, MAF-ECDELM, for improving the accuracy of rolling bearings RUL prediction by integrating deep ensemble learning with error correction. The method was validated on two publicly available full-lifecycle bearing datasets, where it achieved accurate and reliable RUL predictions. Through the case study and comparative experiments, several key findings were established: (1) The deep ensemble model outperformed individual recurrent neural network models by leveraging the complementary strengths of LSTM, BiLSTM, GRU, and BiGRU base learners, demonstrating superior predictive capability; (2) The error correction mechanism further enhanced prediction accuracy by compensating for systematic biases in the ensemble outputs; (3) The MAF effectively suppressed noise and fluctuations in the input vibration signals and prediction results, improving overall stability; (4) Enriching the input with diverse time-domain features improved model performance, but selecting relevant features proved essential, as irrelevant inputs negatively impacted model generalization. These findings highlighted the robustness and effectiveness of the MAF-ECDELM method, offering a valuable advancement in the field of data-driven bearing RUL prediction.

Future work would consider incorporating multi-scale composite signal processing strategies to extract features from degradation signals better. In addition, lightweight or improved attention mechanisms would be explored as base learners to enhance performance in complex temporal modeling. The ensemble strategy would also be optimized to enable dynamic weight allocation, thereby improving the overall generalization capability of the model.

## CRediT authorship contribution statement

**Wenzhe Yin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Hong Xia:** Writing – review & editing, Supervision, Funding acquisition. **Enrico Zio:** Writing – review & editing, Formal analysis, Conceptualization. **Xueying Huang:** Validation, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. U21B2083) and the China Scholarship Council Program.

## Data availability

Data will be made available on request.

## References

Aladag, C.H., Egrioglu, E., Kadilar, C., 2009. Forecasting nonlinear time series with a hybrid methodology. Appl. Math. Lett. 22 (9), 1467–1470.
Amalou, I., Mouhni, N., Abdali, A., 2022. Multivariate time series prediction by RNN architectures for energy consumption forecasting. Energy Rep. 8, 1084–1091.
Atef, S., Nakata, K., Eltawil, A.B., 2022. A deep bi-directional long-short term memory neural network-based methodology to enhance short-term electricity load forecasting for residential applications. Comput. Ind. Eng. 170, 108364.
Bolander, N., Qiu, H., Eklund, N., Hindle, E., Rosenfeld, T., 2009. Physics-based remaining useful life prediction for aircraft engine bearing prognosis. Annual Conference of the PHM Society 1, 1.
Chegini, S.N., Bagheri, A., Najafi, F., 2019. Application of a new EWT-based denoising technique in bearing fault diagnosis. Measurement 144, 275–297.
Chen, D., Hong, W., Zhou, X., 2022. Transformer network for remaining useful life prediction of lithium-ion batteries. IEEE Access 10, 19621–19628.
Cho, K., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. arXiv preprint arXiv.1406.1078.
Dai, L., Guo, J., Wan, J.L., Wang, J., Zan, X., 2022. A reliability evaluation model of rolling bearings based on WKN-BiGRU and wiener process. Reliab. Eng. Syst. Saf. 225, 108646.
Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.J., Ducoffe, M., 2020. Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng. Appl. Artif. Intell. 92, 103678.

Fu, R., Zhang, Z., Li, L., 2016. Using LSTM and GRU neural network methods for traffic flow prediction. In: 2016 31st Youth Academic Annual Conference of Chinese Association of Automation, pp. 324–328.
Golestan, S., Ramezani, M., Guerrero, J.M., Freijedo, F.D., Monfared, M., 2013. Moving average filter based phase-locked loops: performance analysis and design guidelines. IEEE Trans. Power Electron. 29 (6), 2750–2763.
Guo, L., Li, N., Jia, F., Lei, Y., Lin, J., 2017. A recurrent neural network based health indicator for remaining useful life prediction of bearings. Neurocomputing 240, 98–109.
Heimes, F.O., 2008. Recurrent neural networks for remaining useful life estimation. In: 2008 International Conference on Prognostics and Health Management, pp. 1–6.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Huang, Z., Xu, Z., Wang, W., Sun, Y., 2015. Remaining useful life prediction for a nonlinear heterogeneous wiener process model with an adaptive drift. IEEE Trans. Reliab. 64 (2), 687–700.
Kim, S., Seo, Y.H., Park, J., 2024. Transformer-based novel framework for remaining useful life prediction of lubricant in operational rolling bearings. Reliab. Eng. Syst. Saf. 251, 110377.
Kumar, P., Raouf, I., Kim, H.S., 2023. Review on prognostics and health management in smart factory: from conventional to deep learning perspectives. Eng. Appl. Artif. Intell. 126, 107126.
Li, N., Lei, Y., Lin, J., Ding, S.X., 2015. An improved exponential model for predicting remaining useful life of rolling element bearings. IEEE Trans. Ind. Electron. 62 (12), 7762–7773.
Liang, Z., Liang, J., Wang, C., Dong, X., Miao, X., 2016. Short-term wind power combined forecasting based on error forecast correction. Energy Convers. Manag. 119, 215–226.
Meng, S., Shi, Z., Peng, M., Li, G., Zheng, H., Liu, L., Zhang, L., 2024. Landslide displacement prediction with step-like curve based on convolutional neural network coupled with bi-directional gated recurrent unit optimized by attention mechanism. Eng. Appl. Artif. Intell. 133, 108078.
Mienye, I.D., Sun, Y., 2022. A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access 10, 99129–99149.
Mueller, P.N., Woelfl, L., Can, S., 2023. Bridging the gap between AI and the industry—A study on bearing fault detection in PMSM-Driven systems using CNN and inverter measurement. Eng. Appl. Artif. Intell. 126, 106834.
Nandi, S., Toliyat, H.A., Li, X., 2005. Condition monitoring and fault diagnosis of electrical motors—A review. IEEE Trans. Energy Convers. 20 (4), 719–729.
Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., Varnier, C., 2012. PRONOSTIA: an experimental platform for bearings accelerated degradation tests. In: IEEE International Conference on Prognostics and Health Management. IEEE Catalog Number: CPF12PHM-CDR, 1-8.
Ochella, S., Shafiee, M., Dinmohammadi, F., 2022. Artificial intelligence in prognostics and health management of engineering systems. Eng. Appl. Artif. Intell. 108, 104552.
Pascanu, R., Mikolov, T., Bengio, Y., 2013. On the difficulty of training recurrent neural networks. In: International Conference on Machine Learning, pp. 1310–1318.
Peng, P., Li, Y., Guo, Z., 2025. High-performance remaining useful life prediction for aeroengine based on combining health states and trajectory similarity. Eng. Appl. Artif. Intell. 141, 109799.
Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10 (3), 61–74.
Qiu, H., Lee, J., Lin, J., Yu, G., 2006. Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. J. Sound Vib. 289 (4–5), 1066–1090.
Que, Z., Jin, X., Xu, Z., 2021. Remaining useful life prediction for bearings based on a gated recurrent unit. IEEE Trans. Instrum. Meas. 70, 1–11.
Rai, A., Upadhyay, S.H., 2016. A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings. Tribol. Int. 96, 289–306.
Rathore, M.S., Harsha, S.P., 2022. An attention-based stacked BiLSTM framework for predicting remaining useful life of rolling bearings. Appl. Soft Comput. 131, 109765.
Sateesh Babu, G., Zhao, P., Li, X.L., 2016. Deep convolutional neural network based regression approach for estimation of remaining useful life. In: Database Systems for Advanced Applications: 21St International Conference. 2016, pp. 214–228.
Shaheen, B., Kocsis, Á., Németh, I., 2023. Data-driven failure prediction and RUL estimation of mechanical components using accumulative artificial neural networks. Eng. Appl. Artif. Intell. 119, 105749.
Shang, Y., Tang, X., Zhao, G., Jiang, P., Lin, T.R., 2022. A remaining life prediction of rolling element bearings based on a bidirectional gate recurrent unit and convolution neural network. Measurement 202, 111893.
Shen, Y., Tang, B., Li, B., Tan, Q., Wu, Y., 2022. Remaining useful life prediction of rolling bearing based on multi-head attention embedded Bi-LSTM network. Measurement 202, 111803.
Soualhi, A., Medjaher, K., Zerhouni, N., 2014. Bearing health monitoring based on Hilbert–Huang transform, support vector machine, and regression. IEEE Trans. Instrum. Meas. 64 (1), 52–62.
Usman, S.M., Khalid, S., Bashir, S., 2021. A deep learning based ensemble learning method for epileptic seizure prediction. Comput. Biol. Med. 136, 104710.
Yang, Y., Lv, H., Chen, N., 2023. A survey on ensemble learning under the era of deep learning. Artif. Intell. Rev. 56 (6), 5545–5589.
Zhang, H., Zhang, Q., Shao, S., Niu, T., Yang, X., 2020. Attention-based LSTM network for rotatory machine remaining useful life prediction. IEEE Access 8, 132188–132199.
Zhang, F., Liu, Y., Yu, X., Wang, Z., Zhang, Q., Wang, J., Zhang, Q., 2025. Towards facial micro-expression detection and classification using modified multimodal ensemble learning approach. Inf. Fusion 115, 102735.
Zio, E., 2022. Prognostics and Health Management (PHM): where are we and where do we (need to) go in theory and practice. Reliab. Eng. Syst. Saf. 218, 108119.
Zio, E., Di Maio, F., 2010. A data-driven fuzzy approach for predicting the remaining useful life in dynamic failure scenarios of a nuclear system. Reliab. Eng. Syst. Saf. 95 (1), 49–57.