OPEN ACCESS
IOP Publishing
Measurement Science and Technology
Meas. Sci. Technol. **36** (2025) 066102 (17pp)
https://doi.org/10.1088/1361-6501/add48e

Enhanced prognostic reliability for rotating machinery using neural networks with multi-scale vibration feature learning and uncertainty quantification

Kaicheng Zhao[1], Qi Wen[1], He Li[2,*], Wanfu Zhang[1], Zifei Xu[1,3,*] and Ke Feng[4,5]

[1] School of Energy and Power Engineering, University of Shanghai for Science and Technology, Shanghai, People's Republic of China
[2] School of Engineering, Liverpool John Moores University, L3 3AF Liverpool, United Kingdom
[3] School of Engineering, University of Liverpool, Liverpool, Liverpool, United Kingdom
[4] State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710054, People's Republic of China
[5] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, People's Republic of China

E-mail: h.li@ljmu.ac.uk and zxu@usst.edu.cn

Received 26 November 2024, revised 24 April 2025
Accepted for publication 6 May 2025
Published 19 May 2025

Abstract

Reliable remaining useful life (RUL) prediction contributes to fault analysis and preventive maintenance of rotating machinery. Existing artificial intelligence methodologies, however, are challenged by inaccurate feature extraction and uncertainty involved in the RUL prediction process. To this end, this paper proposes a reliable fault prognosis method for rotating machinery using neural networks with multi-scale vibration feature learning and uncertainty quantification. Specifically, the proposed fault prognosis framework starts with constructing a multi-scale semantic embedding module to identify the semantic information in mechanical vibrations. A neural network with local and global feature extraction capabilities is then created to capture information from each scale for RUL prediction. By quantifying the uncertainty of predictions, the framework provides a confidence level for each prediction, and therefore a confidence-based RUL decision fusion method is proposed to achieve the reliable RUL estimation. The feasibility, reliability, and superiority of the framework over state-of-the-art methods are validated by datasets from machinery. Overall, the proposed framework contributes to the safe operation and maintenance of rotating machinery systems.

* Authors to whom any correspondence should be addressed.

Keywords: fault prognostics, remaining useful life, uncertainty quantification, information fusion

1. Introduction

Rotating machinery is an essential category of equipment in the modern industry, faces challenges to the reliability and safety due to complex operating environments and stringent accuracy requirements [1, 2]. Reliability of bearings is the foundation of that of rotating machinery systems [3], which calls for the development of prognostics and health management methods to assist the state awareness and failure prediction of rotating machinery systems [4].

Generally, there are two main types of RUL prediction techniques: physics-based methods and data-driven methods. Physics-based methods use physical principles and mathematical functions to simulate degraddations of a system [5, 6]. However, these methods require extensive prior knowledge and struggle to capture the complex degradation processes of intricate systems. In contrast, data-driven methods do not require an accurate physical model and can predict RUL by modeling the relationship between monitored signals and degradation states, by: predicting RUL through health indicators (HI) derived from collected signals, or directly mining the useful information within signals to establish a nonlinear mapping with RUL [7, 8], the essence of both approaches lies in predicting the remaining life using time series data. Considering data availability, the RUL prediction can be transformed into predicting the remaining usage time based on a known failure threshold or into mining the nonlinear relationship between observations and RUL with full life cycle data [9]. Although using a known failure threshold for early warnings seems more feasible in engineering than relying on full life cycle data, the possibility of obtaining comprehensive life cycle data is increasing due to longer equipment service times, larger datasets from laboratory outputs, and advancements in generative AI [10]. Models based on full life cycle data are likely to outperform those using only failure thresholds for early warnings. In case full life cycle data is available, existing RUL prediction models focus on: (i) construction of reliable HIs to accurately reflect the equipment's performance and predict the initial degradation, and (ii) establishment of a nonlinear relationship between the monitoring data or HIs and the RUL.

To be specific, Xu *et al* [11] developed a multi-scale-multi-head attention with automatic encoder–decoder (MSMHA-AED) model, an unsupervised deep learning framework, to construct a reliable HI for rolling bearing RUL prediction. Experimental results demonstrated that the MSMHA-AED model outperforms traditional methods in RUL prediction, providing higher accuracy and stability, particularly in scenarios with variable operational conditions. Chen *et al* [12] introduced a lognormal-normal mixture model (LNMM), an unsupervised learning framework, to construct a robust HI for gear RUL prediction. This model leverages both normal and lognormal distributions to estimate data discrepancies in raw and exponentially transformed vibration signal domains. Results demonstrate that the approach outperforms traditional methods, providing enhanced predictive accuracy and robustness, especially in representing complex degradation processes under various operational scenarios. Zhou *et al* [13] employed LNMMs to construct unsupervised HIs, which employ a Gaussian mixture model (GMM) to estimate the distribution of raw vibration signals and compute the degradation process. Compared with traditional HI extraction methods, this approaches offer enhanced adaptability to nonlinear degradation trends and better robustness in handling sensor noise.

However, existing models are not able to quantify uncertainties, leading to overconfident RUL estimations. The integration of probabilistic models like GMM into HI construction provides a more robust way to capture degradation characteristics without requiring labeled failure data. However, such models primarily focus on degradation trend estimation while neglecting prediction reliability. On the other hand, Chen *et al* [14] introduced a quadratic function-based DCAE for HI construction, which enforces monotonicity and trend constraints to improve degradation representation. However, existing DCAE-based HIs operate in a single-scale feature space, limiting their generalization across diverse degradation patterns. The application of quadratic function constraints in DCAE ensures that extracted HIs exhibit a smoother degradation trajectory, making them more interpretable for RUL prediction. However, most deep learning-based HI construction methods primarily focus on single-scale representations, making them less effective when applied to machinery with complex and multi-scale degradation behaviors.

Additionally, Cheng *et al* [15] used a CNN and BiLSTM method to construct a degradation indicator (DI) model. This DI model automates feature extraction and simplifies transferability across different bearings without requiring parameter adjustments. Guo *et al* [16] proposed a hybrid method combining a novel HI and a nonlinear Wiener process to construct a fault prognosis model. It accurately reflects the health state of the bearing and is able to effectively captures individual variability in the degradation process. Wen *et al* [17] presented a hybrid CNN-Wiener method to construct a RUL prediction model as a basis of that to enhance feature-level fusion of multi-sensor data and improve noise reduction in RUL estimation. The proposed method provides higher accuracy and robustness in lifetime prediction of equipment. Feng *et al* [18] used a transmission error-based method to construct a fatigue monitoring indicator model to accurately track gear wear and fatigue progression. The proposed method improves the accuracy of RUL prediction for spur gears in intelligent manufacturing. Chen *et al* [19] used a transfer learning method to construct a gear life prediction model. This model enables

accurate RUL prediction under varying working conditions. Cao *et al* [20] utilized a parallel GRU with a dual-stage attention mechanism method to construct a probabilistic RUL prediction model as a basis to enhance the extraction of degradation information and quantifies prediction uncertainty more effectively. Gupta *et al* [21] used a real-time adaptive deep learning method to construct a model for bearing fault classification and RUL estimation. This model enables accurate and timely fault classification and improves the adaptability of RUL prediction by integrating change-point detection.

Similar studies also includes, multi-indicator fault prognostic method presented by Wu *et al* [22] to consider multiple failure modes, such as competition, redundancy, and fusion; the GARCH model proposed by Liu *et al* [23] to construct HIs for early fault detection and defect quantification; physics-informed data augmentation method constructed by Hervé de Beaulieu *et al* [24] to integrate prior physical knowledge to improve interpretability and reliability of RUL prediction; the risk assessment and degradation state coefficient method to construct a RUL prediction model proposed by Li *et al* [25] to deal with variable operational conditions; the RUL prediction model based on vibration signal-based method and a mechanism model established by Zhao *et al* [26] to improve adaptability across various bearing states; the state-space modeling method with a probabilistic entropy-based HI made by Kumar *et al* [27] to capture degradation trends effectively.

The above studies have analyzed the key aspects of RUL prediction, including the construction of HI with high predictability and monotonicity. In the prediction process, neural network models have also been developed to explore the nonlinear relationships between HI or observations and RULs. However, these models often overlook risks posed by overconfident estimates in downstream health maintenance processes, that is, the deterministic models can provide precise estimates of RUL, but are not able to reflect the confidence of predictions. Therefore, this study establishes a reliable RUL prediction framework to accurately predict equipment's RUL and provide confidence of predictions. To further enhance the predictive reliability, a decision fusion module based on uncertainty quantification is proposed. The proposed method consists of: (i) Local and global feature extraction: the vibration signal carries semantic information and is segmented into multiple sub-segments. Local features are extracted from individual sub-segments, while global features capture relationships between sub-segments. This enables precise feature extraction, ultimately improving the performance of RUL prediction; (ii) Reliable prediction: To further enhance predictive reliability, the RUL predictions obtained from semantic scales are fused based on uncertainty quantification. By integrating information across multiple levels of resolution, this fusion strategy ensures a robust and reliable RUL estimation. The main contributions of this paper are as follows:

(i) Develop a neural network to extract local and global degradation information for the RUL prediction.
(ii) Built a dynamic RUL prediction model to estimate the RULs and provide reliability metrics for the predictions.
(iii) Propose a decision fusion method based on uncertainty quantification to enhance the reliability of multi-sensor information in decision fusion.

The rest of the paper is as follows: section 2 introduces the proposed framework. Section 3 describes the experimental setup and dataset. Section 4 presents discussion, and section 5 offers the conclusions.

2. Proposed method

2.1. RUL prediction framework

A reliable RUL prediction framework is proposed for reliable prognosis. Unlike traditional methods that require a predefined end-of-life threshold, this approach predicts RUL through a data-driven deep learning model. Additionally, the model learns the degradation progression from labeled training data $\mathrm{RUL_t} = T_{\mathrm{failure}} - t, \mathrm{for}\ t \leqslant T_{\mathrm{failure}}$, where t is the current time index, $\mathrm{RUL_t}$ represents the failure time of the system. The trained model receives raw vibration signals as input and outputs a predicted RUL value based on the learned nonlinear degradation patterns. This approach allows model to generalize across different degradation scenarios without the need for manually defined failure criteria.

Figure 1 concludes the training process and testing process. In the training process: Step 1 collects vibration signals from sensors. In Step 2, the collected signals undergo data normalization to standardize the input, followed by vibration semantic segmentation and embedding, which prepare the data for feature extraction. The next step proposes a deep learning model with multi-scale local feature extraction and residual connections to learn features from the vibration data. After model training, in step 4, vibration signals are monitored in real-time, and data is prepared for analysis similarly to the training phase and to be processed as the input of the intelligent model. MC dropout is used to address dynamic prediction in step 6. To ensure the reliability of the RUL prediction, step 7 quantifies the prediction uncertainty, which are used as trustworthy values for reliable fusion in step 8. Finally, step 9 consolidates the analysis to provide a reliable estimation of the RUL. This framework combines advanced feature extraction, uncertainty quantification, and information fusion techniques to deliver accurate and reliable RUL predictions, critical for effective condition monitoring and maintenance planning in industrial systems.

2.2. Multi-scale vibration semantics construction module

In the proposed module, the vibration signal x is hypothesized to encode semantic information that reflects the health status of the mechanical system. Assume that vibration $x \in \mathbb{R}^n$ comprises m characters or discrete units, each carrying relevant health information. To capture multi-scale semantic patterns, x is segmented into groups of m characters, where $m \in \{4, 8, 16\}$, resulting in a multi-scale structure that enables the model to analyze the signal across varying resolutions.

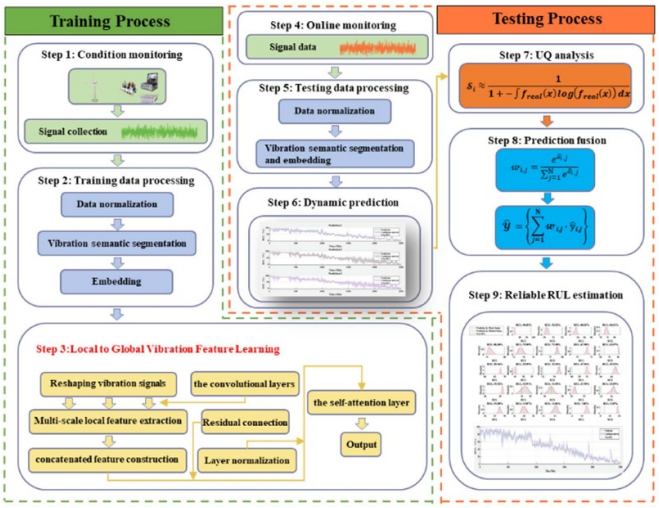


Figure 1. Intelligent reliable prognostic framework.

Mathematically, the signal x is reshaped into an matrix $\mathbb{x} \in \mathbb{R}^{m\times d}$ by reshape function: $\mathbb{x} = \text{reshape}(x, (m,d)$, where d represents the length of each segment at each scale, which also is original embedding dimension of the vibration,

By restructuring x in this way, the multi-scale semantics module leverages different 'character' groupings to provide a robust representation of the underlying vibration semantics, which contains crucial information on the system's condition across multiple levels of resolution. Due to differences in the resolution of captured vibration information, the usability of these scales for RUL prediction is inconsistent. In this study, this variability is reflected in the prediction uncertainty. By quantifying this uncertainty, we assess which scale provides more reliable predictions, forming the basis for a trustworthy decision-fusion module (discussed in section 2.4) that ultimately enhances prediction reliability. In the following section 2.3, we introduce a deep learning model for feature learning from local to global levels, designed to construct the nonlinear relationships between multi-scale vibration semantics and RUL.

2.3. Local to global vibration feature learning

The extracted local and global features play a fundamental role in mapping vibration signals to the implicit HI. Unlike traditional methods that rely on explicitly defined HI, the proposed deep learning model learns a latent representation of these indicators, ensuring a direct correlation between vibration signals and RUL.

The local features are extracted by applying multiple one-dimensional CNN with varying kernel sizes to the segmented vibration signals, aiming to capture fine-grained, short-term characteristics embedded in individual segments. Each convolutional operation independently processes segments at different semantic scales, ensuring comprehensive and diverse local feature representations. Unlike traditional recurrent neural networks, which capture temporal dependencies in a sequential

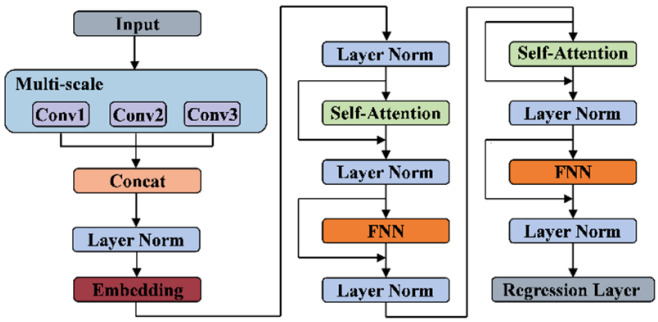


Figure 2. L2GNet architecture.

manner, the proposed method uses a self-attention mechanism to model global dependencies among local features. By computing attention weights across the concatenated local features, the self-attention mechanism captures long-range dependencies and relationships between segments, enabling the network to understand the overall degradation pattern more comprehensively.

By jointly leveraging local and global feature extraction, the approach is able to achieve robust and reliable RUL predictions. Local features enhance sensitivity to early-stage fault patterns and global features ensure long-term trend stability, addressing the limitations of existing deep learning-based RUL estimation approaches that lack customizations. This hybrid approach improves the model's ability to distinguish subtle damage-related patterns from normal operational variations, leading to a more accurate and trustworthy prognosis for rotating machinery.

According to figure 2, the mathematical description of the local-to-global network (L2GNet) and transformations are shown as follow:

The vibration signal will be reshaped by the vibration semantics construction module as $\mathbb{x} \in \mathbb{R}^{m\times d}$, where m is the number of channels, and d is the feature dimension of each channel. Three different separate 1–D convolutional operators $\text{Conv1d}(\cdot)$, with different kernel sizes but same number of filters are applied to extract local features, For each convolution operation, the extracted local features can be represented as:

$$z_i = \text{Conv1d}_i(x) = \mathbb{W}_i * x + \mathbb{b}_i \tag{1}$$

$$\text{Dropout}(x) = \begin{cases} 0 \\ \frac{\text{Conv1d}(x)}{1-\mathbb{p}} \end{cases} \tag{2}$$

$$\text{Batchnorm1d}(x) = \frac{\text{Conv1d}(x) - E(\text{Conv1d}(x))}{\sqrt{\text{Var}(\text{Conv1d}(x)) + \varepsilon}} \tag{3}$$

$$\text{ReLU}(x) = \max(0, x). \tag{4}$$

Where $\mathbb{W}_i \in \mathbb{R}^{c\times m\times k_i}$ is the weights of the convolution operator, c is the number of the channel of $\text{Conv1d}(\cdot)$, $E(\cdot)$ represents the sample mean operation and $\text{Var}(\cdot)$ is sample variance operation, and $\circledast$ is transpose convolutional operation. The local features are also subsequently processed by a

stack of layers including, a dropout layer Dropout $(\cdot)$ for regularization and uncertainty prediction based on Monte Carlo dropout, an activation layer ReLU $(\cdot)$, and batch normalization layer Batchnorm1d $(\cdot)$.

The output of each convolutional layer z_i equipped with a shape $\mathbb{R}^{c\times d'}$, d' depends on the padding and stride, chosen so that $d' = d$. The three local features extracted by the convolutional layers concatenated along the channel dimension resulting in a concatenated feature z:

$$z = \text{concat}\,(z_1, z_2, z_3) \in \mathbb{R}^{3\cdot c\times d}. \tag{5}$$

The concatenated feature z is the input of the self-attention layer, allowing global feature extraction over the combined local features. To account for the sequential nature of data, positional embeddings $p \in \mathbb{R}^{3\cdot c\times d}$ are added to z, resulting in:

$$z_{\text{pos}} = z + p \in \mathbb{R}^{3\cdot c\times d}. \tag{6}$$

The self-attention mechanism operates on z_{pos}, which computes three matrices: Query Q, Key K, and Value V, all of size $\mathbb{R}^{3c\times d_{\text{attn}}}$, where d_{attn} is the attention dimension.

The attention score matrix score is calculated as:

$$\text{score} = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_{\text{attn}}}}\right) \tag{7}$$

where, The self-attention output Attention (Q,K,V) is computed by multiplying score with value V.

To stabilize and enhance the learning process, a residual connection and layer normalization is applied, which is added to the original input z (which is with position embedded mark) and normalized by layer normalization, see equation (8).

$$z_{\text{norm}} = \text{LayerNorm}\,(\text{Attention}\,(Q,K,V) + z). \tag{8}$$

The normalized output z_{norm} is passed through a position-wise feed-forward network (FNN), which consists of two fully connected (dense) layers with a ReLU activation in between.

$$z_{\text{output}} = \text{LayerNorm}\,(\text{ReLU}\,(z_{\text{norm}}W_1 + b_1)\,W_2 + b_2 + z_{\text{norm}}) \tag{9}$$

where $W_1 \in \mathbb{R}^{d_{\text{attn}}\times d_{\text{FNN}}}$, $W_2 \in \mathbb{R}^{d_{\text{FNN}}\times d_{\text{attn}}}$, z_{output} is the final output after the FNN, residual connection, and layer normalization, which will be decoded by linear layer for RUL estimation. The estimation RUL $\hat{y}$ can be obtained by L2GNet $(x\,|\theta\,)$ in equation (10), where θ denotes the parameters of the network.

$$\hat{y} = \text{L2GNet}(x|\theta). \tag{10}$$

2.4. Decision-fusion module based on uncertainty quantification

In the decision fusion section, leveraging the proposed prognostic framework based on a probabilistic neural network, It obtains dynamic predictive RUL estimates at degradation time t through Monte Carlo sampling, resulting in an estimation set $\hat{\mathcal{Y}} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots \hat{y}_n)$, where n depends on sampling times. This approach quantifies the confidence of the predictions and provides a comprehensive set of information for decision-making, thereby enabling more informed and reasonable decisions under uncertainty.

A probability density distribution for each sample's prediction results is obtained using kernel density estimation (KDE). For ith sample that is at any degradation time, the estimated distribution function is denoted by:

$$f_{\text{real}}\,(x) \approx \text{KDE}\left(\hat{\mathcal{Y}}\right) \tag{11}$$

where x represents the range of predicted values.

The Shannon entropy is used to quantify the uncertainty of the distribution of prediction results. The entropy for the ith sample is calculated as:

$$H_i = -\int f_{\text{real}}\,(x)\log\left(f_{\text{real}}\,(x)\right)\mathrm{d}x \tag{12}$$

where the integral is approximated as a discrete sum in the implementation, with the integration step dx determined by the spacing of x in the kernel density estimate:

$$H_i \approx -\sum_k f_{\text{real}}\,(x_k)\log\left(f_{\text{real}}\,(x_k) + \varepsilon\right) \tag{13}$$

where ε is a small constant added for numerical stability. With the qualified uncertainty, the trustworthy value is estimated by transforming it into a trustworthy score:

$$\mathcal{S}_i \approx \frac{1}{1+H_i} \tag{14}$$

where a smaller entropy corresponds to a confidence score close to 1, indicating a concentrated prediction distribution and higher confidence of predictions. Any trustworthiness value less than zero will be set to zero, and any value greater than one will be capped at one.

For multi-scale L2GNet, let $\hat{y}_{i,j}$ denote the prediction for the ith time step by the jth L2GNet single network, where $j \in \{1, 2\ldots, N\}$, in this study as mentioned in section 2.1, $N = 3$. The confidence score $\mathcal{S}_{i,j}$ for each $\hat{y}_{i,j}$ is computed using the entropy-based formula in equations (11)–(14), so that the exponentiate the posterior scores can be obtained by normalizing the confidence scores for each time i across all models using the Softmax function:

$$w_{i,j} = \frac{\mathrm{e}^{\mathcal{S}_{i,j}}}{\sum_{j=1}^{N}\mathrm{e}^{\mathcal{S}_{i,j}}}. \tag{15}$$

The reliable prediction is:

$$\hat{y}_i = \sum_{j=1}^{N} w_{i,j}\cdot\hat{y}_{i,j}. \tag{16}$$

Based on the trustworthy value of the prediction, algorithm 1 proposed can obtain trustworthy RUL estimation.

Algorithm 1. Reliable fusion module.

Trustworthiness estimation based on uncertainty quantification for reliable decision fusion module:

Input: Testing Data $\mathcal{D}_{test} = \mathcal{X}_{test}$, L2GNet model: $\text{L2GNet}_j\left(\cdot|\hat{\theta}\right)$, sampling time K

Output: Prognostic result $\hat{\mathcal{Y}}$

for $\mathbb{x}_i$ in $\mathcal{X}_{test}$

Sampling K times from the branch network of MS-L2GNet model: $\text{L2GNet}_j\left(\mathbb{x}_i|\hat{\theta}\right)$

where k^{th} prediction is $\hat{\mathbb{y}}_{i,j}^{k} = \text{L2GNet}_j\left(\mathbb{x}_i|\hat{\theta}\right)$

Record $\hat{\mathbb{y}}_{i,j} = \left\{\hat{\mathbb{y}}_{i,j}^{k} | k \in N,\ 1 \leqslant k \leqslant \text{K}\right\}$

Obtain Trustworthy score $\mathcal{S}_{i,j}$ by **equations (11)–(14)**

Trustworthy prognostic $\hat{\mathbb{y}}_i$ for $\mathbb{x}$ **can be obtained by equations (15)** and **(16)**

end

Prognostic result for $\mathcal{X}_{test}$: $\hat{\mathcal{Y}} = \{\hat{\mathbb{y}}_i\}$

end

Output: Reliable Prognostic: $\hat{\mathcal{Y}}$;

end

This pseudocode presents the reliable Fusion module for uncertainty quantification-based prognostic fusion in multi-scale L2GNet models. For each test sample x_i in the test dataset $\mathcal{X}_{\text{test}}$, predictions are generated from multiple L2GNet models at different scales with K repeated samplings to capture predictive variability. The trustworthiness of each prediction is quantified using entropy-based scores, which are then normalized via a softmax function. A weighted fusion of predictions is conducted based on the trustworthiness scores, resulting in a final prognostic output $\hat{y}_i$ for each test sample. This approach enhances the reliability of predictions by accounting for model confidence at different scales, yielding a robust prognostic outcome.

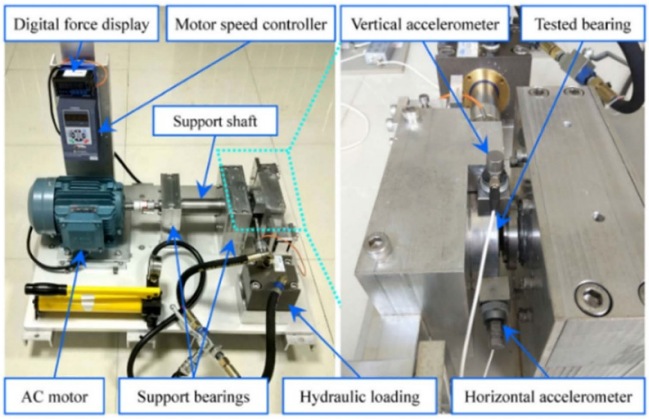


Figure 3. Bearing accelerated life test platform.

3. Experimental setup and dataset

3.1. Dataset description

The efficacy of the methodology is validated through the rolling bearing accelerated life test dataset from Xi'an Jiaotong University. The architecture of the bearing life test apparatus is illustrated comprehensively in figure 3.

Figure 3 indicates that the experimental platform is equipped with an accelerometer, an electric motor speed controller, a shaft, bearings, a hydraulic loading system, and an alternating current motor, all dedicated to the meticulous monitoring of the complete lifecycle data of rolling bearings. The sampling configuration is established at a frequency of 25.6 kHz, with intervals set at one minute and each sampling period lasting 1.28 s, generating 32 768 data points per sample. The dataset comprises data from 15 rolling bearings, each subjected to three distinct operational conditions. Rolling bearings of the LDK UER204 model are utilized, with their pertinent parameters detailed in table 1.

The experimental design encompassed three distinct operating conditions, see table 2, featuring five bearings per condition. This investigation conducts a detailed analysis of a bearing dataset from each specified operating condition, identified as 1_2, 2_1, and 3_4. The associated full-lifecycle time-domain representations are meticulously depicted in figure 4. Corresponding sample counts, actual service lifetimes, and failure locations are enumerated in table 3, where datasets highlighted in boldface are those subsequently employed for model performance validation.

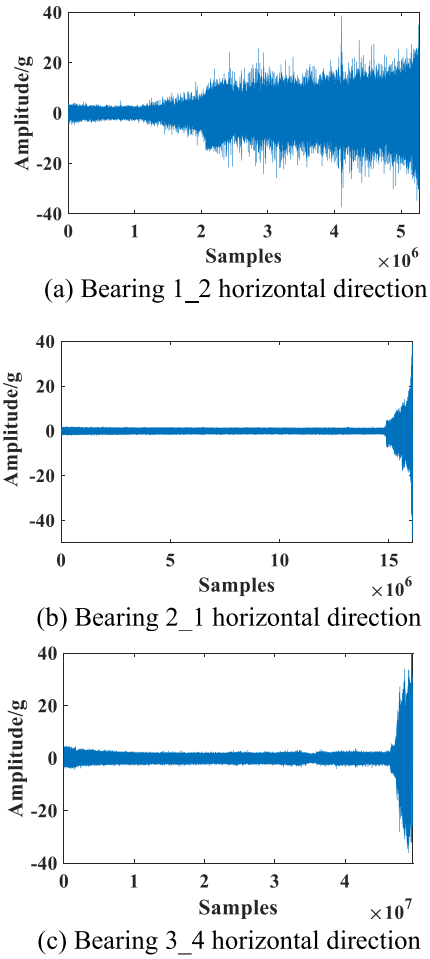
The three bearings reflect representative degradation patterns: (i) For Bearing 1–2, its amplitude shows that this bearing operated healthily for more than half of its lifespan before degradation began. This gradual degradation pattern is typical in scenarios where the component experiences consistent wear over time before any significant faults emerge; (ii) For Bearing 2–1, it operated healthily for most of its lifespan but then suddenly began to degrade and quickly failed. This abrupt failure pattern illustrates cases where a component may seem stable until an unexpected fault rapidly accelerates its deterioration; (iii) For Bearing 3–4, it started failing shortly after operation, indicating an early failure mode. This pattern is often associated with manufacturing defects or initial installation issues,

Table 1. LDK UER204 bearing parameters.

Parameter	Value	Parameter	Value
Inner ring raceway diameter (mm)	29.30	Ball Diameter/mm	7.92
Outer ring raceway diameter (mm)	39.80	Number of balls	8
Bearing diameter (mm)	34.55	Contact angle/(°)	0
Basic rated dynamic load (*N*)	12 820	Basic rated static load/*N*	6650

Table 2. Bearing accelerated life test conditions.

Condition number	Rotating speed (r min^{-1})	Radial force (kN)
1	2100	12
2	2250	11
3	2400	10



(a) Bearing 1_2 horizontal direction

(b) Bearing 2_1 horizontal direction

(c) Bearing 3_4 horizontal direction

Figure 4. Vibration signal of three kinds of bearings.

where faults emerge immediately after the component is put into use. These three bearings were chosen to represent distinct degradation modes, gradual wear, sudden failure, and early failure, providing insights into various scenarios in predictive maintenance.

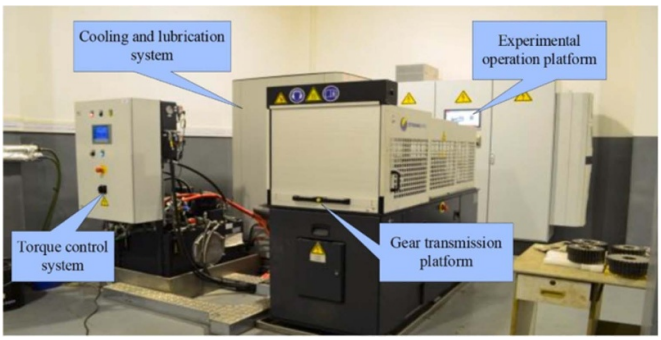


Figure 5. Gear test platform.

To ensure a fair evaluation of the proposed method and comparative models, the dataset was divided into a training set (80%) and a test set (20). The division was performed randomly but stratified to maintain a consistent distribution of degradation states across training and testing data. This ensures that the model can generalize well across different failure stages. The training set is used for model learning, including feature extraction and parameter optimization, while the test is used for performance evaluation. No overlap exists between the training and test sets to prevent data leakage.

3.2. Description of the gear dataset

To validate the predictive performance of the proposed model under conditions where true RUL labels are unavailable, an gear degradation dataset [28] is examined, see figure 5.

The experimental platform consists of four main subsystems: a torque control system, a gear transmission platform, a cooling and lubrication system, and an experimental operation platform. The torque control system ensures precise loading conditions during operation, the gear transmission platform simulates real-world mechanical degradation processes. The cooling and lubrication system helps maintain stable thermal and mechanical conditions throughout the experiment, and the experimental operation platform enables centralized control and data acquisition. Four run-to-failure datasets are collected under two distinct operating conditions: low speed–high load (500 rpm, 1.4 kN) and high speed–moderate load (1000 rpm, 1.3 kN). Each condition is designed to simulate different mechanical stress levels encountered in industrial gear systems. Vibration signals are monitored and sampled at a high frequency of 50 kHz to ensure accurate capture of degradation features. For each recording, 20 s of vibration data were collected at 40 second intervals, resulting in high-resolution time

Table 3. XJTU bearing data set information.

working condition	Bearing dataset	Sample number	Actual life	Failure position
1	1_1	123	2 h 3 min	Outer race
	1_2	**161**	**2 h 41 min**	**Outer race**
	1_3	158	2 h 38 min	Outer race
	1_4	122	2 h 2 min	Retainer
	1_5	52	52 min	Outer race, inner race
2	**2_1**	**491**	**8 h 11 min**	**Inner race**
	2_2	161	2 h 41 min	Outer race
	2_3	533	8 h 53 min	Retainer
	2_4	42	42 min	Outer race
	2_5	339	5 h 39 min	Outer race
3	3_1	2538	42 h 18 min	Outer race
	3_2	**2496**	**41 h 36 min**	**Inner race roller, retainer, outer race**
	3_3	371	6 h 11 min	Outer race
	3_4	**1515**	**25 h 15 min**	**Inner race**
	3_5	114	1 h 54 min	Outer race

Table 4. Hyper parameters setup of L2GNet model.

Hyperparameter	Value	Hyperparameter	Value
Convolutional kernels	[3, 5, 7]	Fully connected layers	$16 \rightarrow 4 \rightarrow 1$
Number of filters	32 (each branch)	Batch size	64
Dropout rate	0.25	Initial learning rate	0.001
Attention heads	16	Learning rate scheduler	Step size = 50, decay = 0.98
Attention hidden dim	512	Max epochs	200
Position embedding dim	96	Optimizer	Adam

series data that reflects the full progression of gear degradation from healthy to failure.

In the selected experiment, the gear experienced a typical pitting degradation process over its life cycle. This dataset provides a well-defined degradation trend, making it suitable for benchmarking the model's predictive performance under long-horizon closed-loop forecasting conditions.

3.3. Hyper parameters setup and evaluation

The program is developed in Python. The hardware configuration used for running the program includes an Intel i9-12900K processor and an NVIDIA RTX A5500 graphics card. The deep learning framework employed is PyTorch version 11.8. Three evaluation metrics are used to assess the performance of prognostic: mean absolute error (MAE), root mean square error (RMSE), and *R*-squared (R^2). These performance evaluation metrics (PEMs) are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{17}$$

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{18}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{19}$$

where, y_i specifically represents the true RUL, while $\hat{y}_i$ denotes the predicted RUL generated by the model, and $\bar{y}$ is the mean of the true RUL values.

The hyperparameter configuration for the proposed L2GNet model is summarized in table 4. The architecture is designed to integrate multi-scale convolution and attention-based mechanisms for robust degradation modeling. The convolutional kernel sizes are set to 3, 5, and 7, respectively, with causal padding to preserve temporal dependencies. The number of filters in each convolutional layer is fixed at 32. These kernels are applied in parallel to enable diverse local feature extraction across different receptive fields. The outputs are concatenated using a depth concatenation layer before entering the subsequent normalization and attention modules.

A dropout rate of 0.25 is employed after convolution to regularize training and prevent overfitting. The attention mechanism consists of two stacked multi-head self-attention layers, each configured with 16 attention heads and an internal hidden dimension of 512. Position embeddings are added prior to attention computation to encode temporal order information.

The fully connected layers used in the prediction head include two hidden layers with dimensions of 16 and 4, followed by a final regression layer outputting a single scalar. Training is performed using a mini-batch size of 64 and an initial learning rate of 0.001. The optimizer is Adam, and the learning rate decays by 2% every 50 iterations using a step scheduler. The model is trained for up to 200 epochs with early stopping enabled to improve training efficiency and avoid overfitting.

These hyperparameter settings are chosen to balance learning efficiency, prediction stability, and generalization capacity across different degradation scenarios.

4. Discussion and analysis

4.1. Ablation study

An ablation study is conducted to evaluate the functions with the greatest impact on the performance of the proposed model. Table 5 presents the prediction performances of various models, including TCN, Transformer, AttenNet, and the proposed method with different length of vibration semantics, for example, L2GNet-4 means each length of vibration sample is split as 4 vibration words.

the TCN model, which is designed to extract local features from vibration signals. However, this method shows the highest error metrics (RMSE and MAE) and the lowest R^2 value, suggesting that local feature extraction alone is insufficient to capture the underlying patterns in the. The Transformer model reduces both RMSE and MAE compared to TCN and achieves a higher R^2 value, indicating that global feature extraction enhances accuracy in predicting the RUL. AttenNet further improves the predictive performance by integrating global features. However, effectively capturing both global and local information is crucial. The proposed L2GNet model achieves the best overall performance, demonstrating that the combination of local and global features, as in AttenNet-4, improves prediction accuracy. This finding underscores our motivation: different vibration segments may carry unique information, much like multi-source sensor data, and combining them could yield richer insights.

4.2. Computational complexity analysis

To evaluate the computational feasibility of the proposed L2GNet-8, this section provides a complexity analysis of its key components: local feature extraction via convolutional layers, global feature learning via self-attention, and final regression prediction via fully connected layers. The computational complexity is analyzed in terms of floating-point operations considering the primary architectural components. Table 6 presents the results.

The computational complexity analysis demonstrates the efficient balance between local and global feature extraction, with self-attention serving as the key component for capturing long-range dependencies. Self-attention accounts for over 97.4% of the total FLOPs due to its quadratic dependency on sequence length LL, this mechanism enhances feature representation and model expressiveness. Meanwhile, the convolutional and fully connected layers contribute minimally to the overall complexity, ensuring efficient local feature extraction and lightweight prediction processing. This design enables L2GNet-8 to achieve high predictive accuracy while maintaining computational feasibility, making it well-suited for real-world applications that require both robust prognostics and scalable performance.

Table 5. Prediction performance of different techniques.

Method/PEMs	RMSE	MAE	R^2
Local feature extraction by TCN	10.425	8.214	0.871
Global feature extraction by transformer	5.193	3.986	0.968
Global feature fusion by -AttenNet	3.902	3.088	0.972
L2GNet-4	3.616	2.933	0.985
L2GNet-8	3.636	2.830	0.984
L2GNet-16	3.727	2.839	0.984

Table 6. Computational complexity of L2GNet-8.

Component	Computational complexity	Percentage contribution
CNN layers	2.62×10^5	0.003%
Self-attention mechanism	8.39×10^9	97.4%
Fully connected layers	2.2×10^9	2.56%

4.3. Prognostic performance comparison

Conventional TCN [29], CLSTM [30], CBi-LSTM [31], AttenNet [32], Atten-LSTM [33], AED-BNN [11] and CNN-GRU [34], methodologies, alongside the novel L2GNet approach introduced herein, are utilized to forecast the lifespan of the bearing dataset 1_2, 2_1, 3_2 and 3_4. The prognostic performance test under 95% confidence interval is shown in table 7.

Table 7 presents the prognostic performance of methods on four bearings, evaluated through three key metrics including RMSE, MAE, and R^2. The models listed include various state-of-the-art (SOTA) methods, as well as the proposed L2GNet framework: (i) Based on RMSE, L2GNet achieves the lowest RMSE across all tests, with values of 3.445, 10.382, 9.157, and 8.071, outperforming other methods. This suggests that L2GNet provides superior predictive accuracy for RUL estimation by effectively reducing prediction error. Comparatively, TCN and CLSTM yield higher RMSE values, indicating a less accurate prediction. This outcome highlights L2GNet's advantage in capturing relevant features for precise fault prognosis; (ii) Based on MAE, L2GNet demonstrates the lowest MAE values (2.720, 7.9793, 7.026, 5.714) across the bearings, further confirming its robust performance. AttenNet shows promising results but with a lower precision than L2GNet's; (iii) Based on R^2, L2GNet achieves the highest R^2 values (0.986, 0.871, 0.900, 0.923), indicating a strong correlation between predicted and actual RUL values. The high R^2 values

Table 7. Prognostic performance test on Bearing 1_2/Bearing 2–1/Bearing 3–2/Bearing 3–4.

Method/PEMs	RMSE	MAE	R^2	Training time
TCN	10.110/20.231/12.163/13.619	8.149/ 16.125/9.228/10.339	0.879/ 0.507/0.823/0.777	23 mins
CLSTM	9.168/ 18.627/13.284/13.324	7.372/ 14.856/9.765/10.312	0.900/ 0.585/0.788/0.783	42 mins
CBi-LSTM	5.138/ 10.833/9.225/9.254	3.896/ 8.2458/6.964/7.093	0.969/ 0.859/0.898/0.897	56 mins
AttenNet	3.664/ 11.258/9.206/9.460	2.897/ 8.7289/7.079/7.385	0.981/ 0.848/0.894/0.892	78 mins
Atten-LSTM	8.575/ 18.525/13.344/13.189	7.045/ 14.640/9.817/9.946	0.912/ 0.589/0.786/0.791	122 mins
AED-BNN	4.393/ 15.550/8.702/9.513	3.467/ 12.059/6.357/7.307	0.977/ 0.711/0.902/0.891	63 mins
CNN-GRU	4.352/ 14.623/11.321/12.239	4.021/ 17.323/17.317/11.342	0.944/ 0.629/0.824/0.821	67 mins
L2GNet	3.445/ 10.382/8.071/9.157	2.720/ 7.9793/5.714/7.026	0.986/ 0.871/0.923/0.900	23 mins

signify that L2GNet models the underlying degradation trend across different bearings. In comparison, models like CLSTM and AED-BNN, which integrate convolutional and probabilistic elements, achieve relatively lower R^2 values, suggesting a lack of the capacity to capture the complete degradation trend. Overall, L2GNet's consistent superiority across RMSE, MAE, and R^2 metrics suggests that the framework's combination of global and local feature extraction provides an essential advantage in RUL prediction. The comparison with SOTA methods, including TCN, AttenNet, and CLSTM, underscores the impact of L2GNet's architecture in improving prediction accuracy and model reliability. To visually illustrate the results of RUL prediction, figure 6 presents the predicted RUL curves.

Figure 6 presents the RUL predictions for four bearings, illustrating different failure modes: inner race failure, outer race failure, and hybrid failure. The performance of various methods is shown, including TCN, CLSTM, CBi-LSTM, AttenNet, Atten-LSTM, and AED-BNN, alongside the proposed method L2GNet in comparison to the true RUL. From overall prediction accuracy point of view, the proposed method L2GNet demonstrates a closer alignment with the true RUL across all failure modes and bearings indicating the degradation trend of which is more accurate than others, particularly in the later stages of degradation, where accurate RUL estimation is crucial for timely maintenance decisions. For bearings with inner race failure (subplots (a) and (c)), L2GNet shows a better performance in tracking the decreasing RUL compared to others. It highlights L2GNet's advantage in capturing complex degradation patterns specific to inner race failures. In the case of outer race failure (subplot (b)), L2GNet's curve remains closely aligned with the true RUL, showing minimal divergence even as degradation accelerates. Other methods, such as TCN and CLSTM, show more variability and fail to follow the RUL trend, especially at the end-of-life approaches. L2GNet's ability to handle outer race failure with higher accuracy reflects the robustness of its global-local feature extraction, which enables it to model distinctive failure modes more precisely. For hybrid failure (subplot (d)), L2GNet demonstrates superior performance. The model accurately follows the RUL decline, managing the mixed degradation characteristics better than competing methods. The ability to handle hybrid failure effectively underscores L2GNet's flexibility and adaptability in modeling complex degradation modes that involve multiple failure mechanisms. Across all cases, competing methods (TCN, CLSTM, CBi-LSTM, AttenNet, Atten-LSTM, and AED-BNN) exhibit higher variability and deviation from the true RUL, especially in the later stages of degradation. This tendency towards divergence highlights their limitations in generalizing across different failure types. In contrast, L2GNet consistently achieves high accuracy and minimal deviation from the true RUL curve, suggesting that the combined global-local feature extraction approach is highly effective in capturing the nuances of bearing degradation.

4.4. Multiple timestep prognosis validation

To evaluate the robustness of the proposed L2GNet model under realistic constraints, a multi-timestep prediction in a closed-loop manner is proposed. Specifically, the model predicts the root mean square (RMS, see equation (20)) of vibration signals over future steps, using initial input segments without access to the full degradation trajectory.

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \tag{20}$$

where x_i is the signal at the ith sample and n reflects the total number of samples.

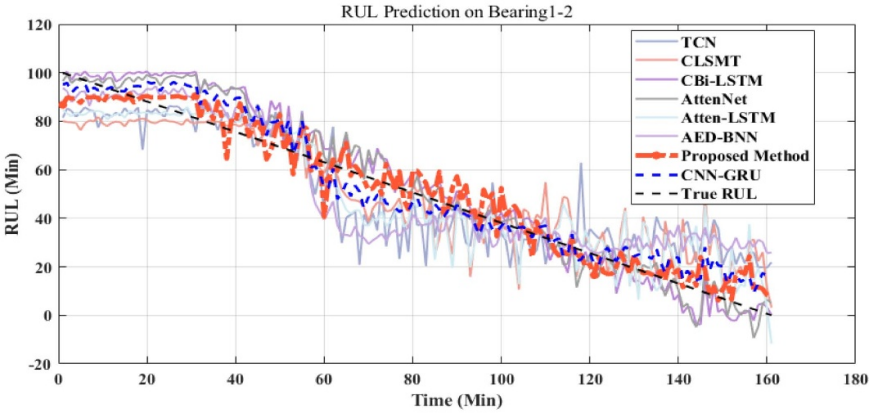
In the closed-loop prediction setting, the model is initialized with a real sequence from historical data and performs recursive forecasting to simulate the realistic scenario where the future trajectory is unknown, by:

$$\hat{y}_{t+1} = f(x_{t-p+1}, \ldots, x_t) \tag{21}$$

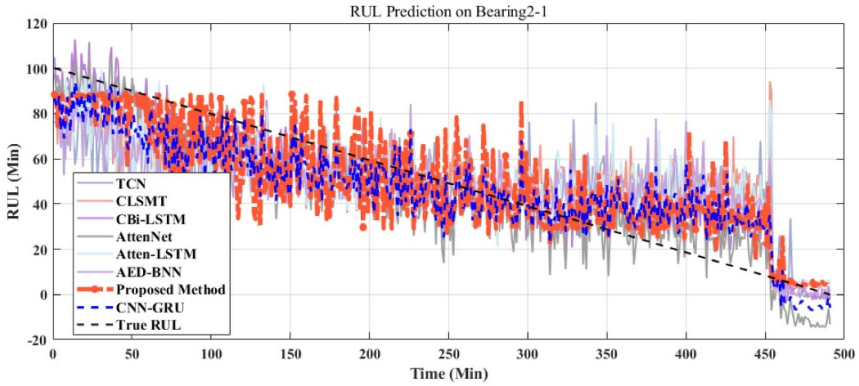
$$\hat{y}_{t+k} = f\left(\hat{y}_{t+k-p}, \ldots, \hat{y}_{t+k-1}\right). \tag{22}$$

Where $f(\cdot)$ denotes the trained model, x_i represents the real input value at time step i, $\hat{y}_{t+k}$ is the model's predicted value at time step $t+k$, p is the input window size, K is the total number of predictions.

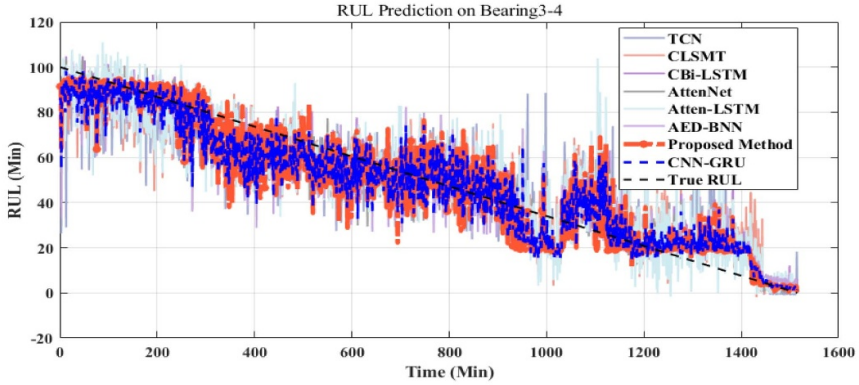
This setup reflects a practical prognostic scenario where the model must iteratively forecast degradation signals without knowing the endpoint of failure. The prediction performance under this setting allows to assess the model's ability of generalization of long-horizon forecasting and its resilience to cumulative errors inherent in recursive prediction.



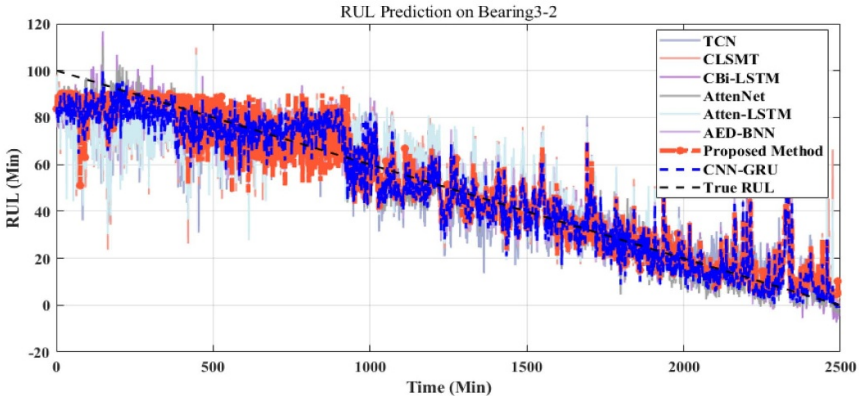
(a) RUL prediction of bearing 1-2 (Inner race failure)



(b) RUL prediction of bearing 2-1 (outer race failure)



(c) RUL prediction of bearing 3-4 (Inner race failure)



(d) RUL prediction of bearing 3-2 (Hybrid failure)

Figure 6. RUL prediction comparison.

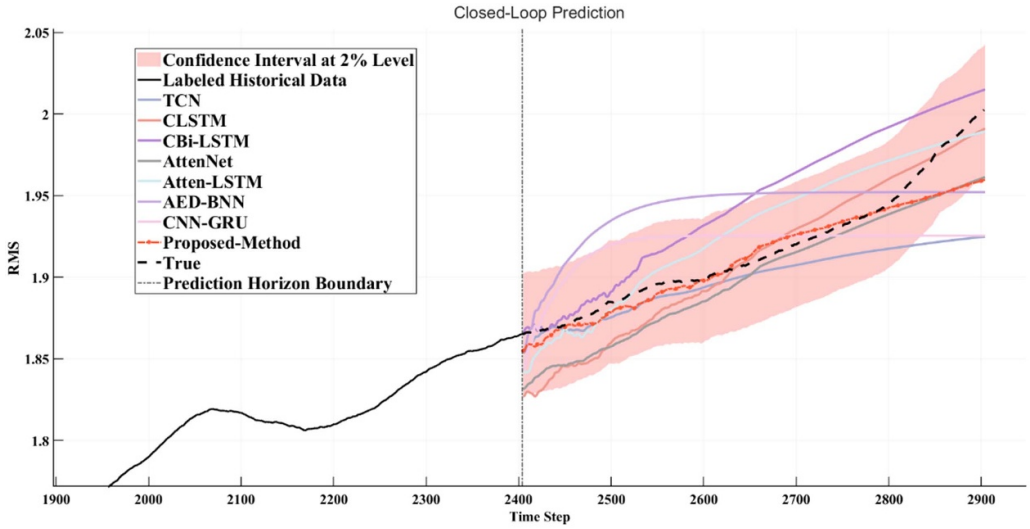


Figure 7. Closed-loop prediction comparation.

Table 8. Prediction performance of gear.

Method/PEMs	RMSE	MAE	R^2
TCN	0.0272	0.0188	0.4401
CLSTM	0.0172	0.0144	0.7758
CBi-LSTM	0.0326	0.0284	0.1966
AttenNet	0.0218	0.0188	0.7041
Atten-LSTM	0.0197	0.0169	0.7072
AED-BNN	0.0374	0.0340	−0.0585
CNN-GRU	0.0312	0.0258	0.2642
L2GNet	0.0129	0.0083	0.8738

To further assess the generalization ability of the proposed model on different mechanical components and degradation scenarios, a gear dataset was selected, in which the gear was made of 20CrMnMo and subjected to a contact fatigue test. The experiment was conducted on a dedicated gear contact fatigue test rig, with a rotational speed of 1000 r min^{-1} and a torque of 1300 N*m. Vibration signals were continuously recorded with a high sampling frequency of 50 KHz. The gear experienced a typical pitting degradation process throughout its life cycle. To ensure the relevance of the prediction task, the final stage data containing clear degradation trends are used for model validation.

Figure 7 shows the closed-loop prediction performance of the proposed model applied to the gear dataset. It illustrates the closed-loop prediction performance of models on the gear degradation dataset, where the vertical dashed line represents the prediction horizon boundary, the point beyond which all future values must be recursively predicted by the model without access to ground truth. The proposed method demonstrates a close alignment with the true degradation trend, especially within the early and mid-stages of prediction. Compared to others, L2GNet achieves the lowest prediction error and the highest R^2 value, see table 8.

Overall, the proposed method provides accurate short-term predictions, and which maintains robust across extended time steps. The reason can be trace back to the L2GNet adopts a sliding-window-based reconstruction strategy to generate multiple supervised samples from a single life-cycle sequence, effectively expanding the training dataset. The multi-scale convolutional architecture equips the proposed model an efficient extraction of degradation patterns at different temporal resolutions, reducing the reliance on long input sequences and mitigating the risk of overfitting associated with small sample sizes. The local-to-global feature fusion mechanism further enhances the model's ability to generalize across different degradation stages within a limited dataset, making it particularly suitable for small-sample RUL prediction tasks.

4.5. Trustworthy analysis of the prognosis

Predictive uncertainty is critical for RUL predictions. This section focuses on the quantification of the predictive uncertainty. In figure 8, bearing 3–2 is used to examine the dynamic RUL prediction of the proposed method because hybrid failures commonly exist in real industrial scenarios. It presents the predicted RUL distributions for three multi-scale models (L2Gnet-4, L2Gnet-8, L2Gnet-16) across three degradation stages: early (RUL: 98.84%), middle (RUL: 51.94%), and late (RUL: 11.86%). During the early degradation stage, the true RUL is approximately 98, the peak predicted distributions from L2Gnet-4, L2Gnet-8, and L2Gnet-16 are between 86–88, failing to fully capture the true RUL. Although L2Gnet-4's prediction slightly closer to the true RUL, the overall distribution still deviates significantly from the true value. This explains why the confidence score for L2Gnet-16 is higher (75.26%) compared to L2Gnet-8 (62.07%) and L2Gnet-4 (50.93%). The higher confidence in L2Gnet-16 is due to its more concentrated distribution, indicating greater confidence in its predictions. At the middle degradation stage, the true RUL is around 52, with the peak distribution of L2Gnet-4 between 50–51 (close to the true RUL. The distributions for L2Gnet-8 and L2Gnet-16 are close to the true RUL as well, but the overall range is wider, indicating some level of

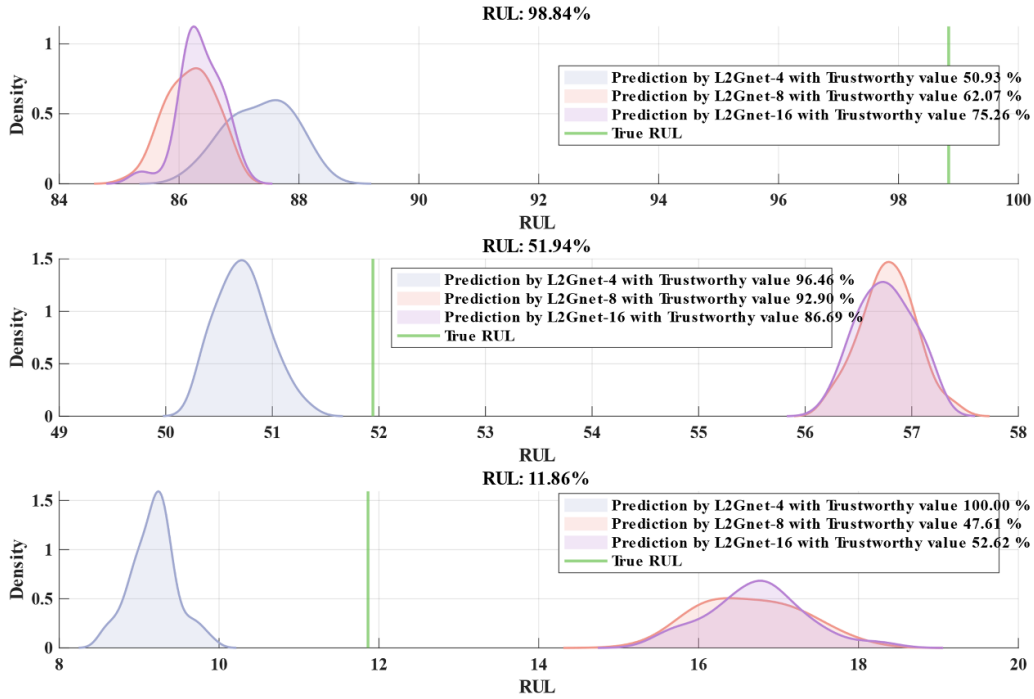


Figure 8. Uncertainty quantification and trustworthy analysis.

uncertainty in their predictions. As a result, L2Gnet-4 achieves the highest confidence score of 96.46%, primarily due to its highly concentrated prediction distribution and its proximity to the true RUL, demonstrating higher reliability in the middle degradation phase. In the late degradation stage, the true RUL is approximately 12. The prediction distribution of L2Gnet-4 almost overlaps with the true RUL and is highly concentrated, indicating a precise prediction of RUL. On the other hand, L2Gnet-8 and L2Gnet-16 exhibit a significant deviation, with predicted values concentrated between 14–18 and with broader distributions. But, L2Gnet-4 has a confidence score of 100.

In summary, the confidence score reflects the model's level of trust in its prediction, which is based on the alignment between the predicted results and the true value, as well as the uncertainty in the prediction distribution. These analyses provide a deeper understanding of each model's predictive capabilities and limitations at different degradation stages, emphasizing the importance and value of incorporating multi-scale information for RUL prediction.

4.6. Validation of the decision fusion based on uncertainty quantification

The analysis of prognosis with uncertain quantification of the RUL prediction by the proposed method indicates that the quantified predicting uncertainty can provide trustworthiness of the corresponding prediction. Figure 9 gives the uncertainty quantification distributions of RUL predictions using different vibration semantic scales, that are respectively are L2GNet-4, L2GNet-8, and L2GNet-16. L2GNet-4 generally show a narrower distribution, indicating lower uncertainty, especially in the early stages (top rows). This suggests that the vibration semantic scale of 4 provide more stable predictions when the system is relatively healthy, which is often close to the true RUL (green line), indicating that L2GNet-4 provides accurate predictions, particularly in the middle and early degradation stages. L2GNet-8 has broader distributions with higher uncertainty in predictions. In some cases, As RUL decreases, L2GNet-8 tends to provide uncertain prediction. L2GNet-16 are generally characterized by a wider spread in the later RUL stages (lower right portion of the image). This shows a high level of uncertainty, particularly at the end-of-life predictions. L2GNet-16 shows narrower uncertainty compared to L2GNet-8 in the early stages, but not as consistent as L2GNet-4. In summary, L2GNet-4 holds the most reliable and concentrated prediction distributions, indicating it captures the early failure dynamics well. As the system degrades (RUL decreases), all models exhibit increased uncertainty, as seen by the broader prediction distributions. Thus, it is necessary to combine all scales information together to improve the reliability of RUL predictions.

To validate the effectiveness of the proposed decision-fusion method. RUL predictions using two fusion methods are listed in figure 9: Direct Fusion and Reliable Fusion. Direct Fusion combines the results of L2GNet-4, L2GNet-8, and L2GNet-16 without considering the prediction uncertainty. The proposed method fuses predictions by weighting each component's reliability, to produce a more accurate and trustworthy RUL estimation. The Reliable Fusion predictions (red curves) tend to be sharper and more concentrated around the true RUL value (green line) compared to the Direct Fusion predictions (blue curves). The Direct Fusion predictions with higher uncertainty. In several early-stage plots,

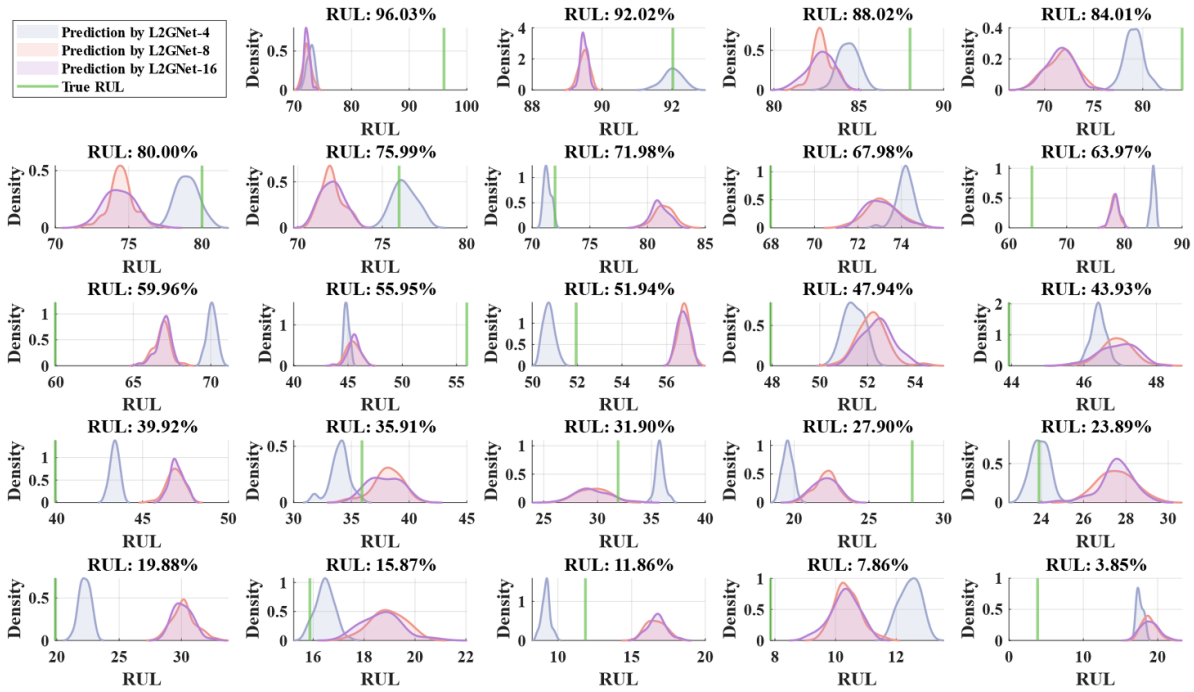


Figure 9. Uncertainty analysis of three single L2GNet.

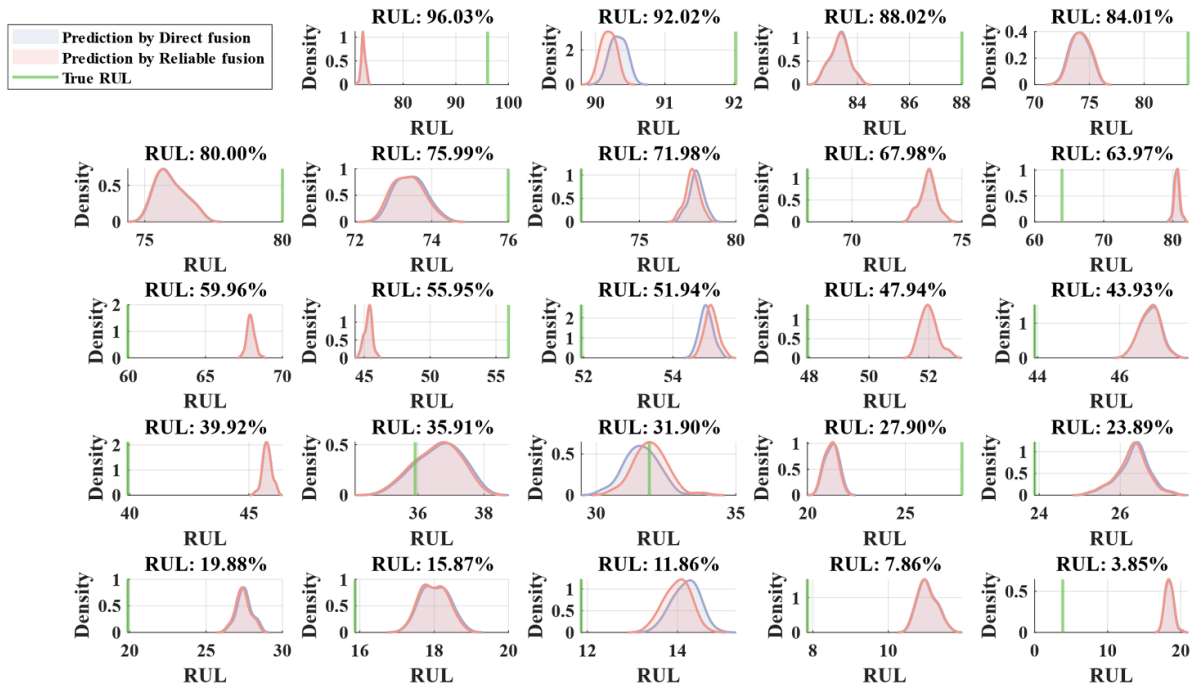


Figure 10. Uncertainty analysis of different fusion methods.

Direct Fusion struggles to capture the early degradation characteristics. Reliable Fusion produces a distribution that aligns well with the true RUL, highlighting its advantage in capturing early degradation signals by weighting contributions from each semantic scale. As the RUL decreases, Direct Fusion continues to exhibit more variability in its predictions, with a broader distribution that is often skewed or offset from the true RUL. This suggests that the simple combination of L2GNet-4, −8, and −16 does not adapt to changing system conditions. Reliable Fusion, on the other hand, maintains a narrower and more centered distribution, showing that it can adaptively emphasize the most reliable components, thereby reducing the prediction uncertainty.

Figure 10 gives the Uncertainty analysis of different fusion methods. In RUL regions like RUL: 59.96% and RUL: 55.95%, Reliable Fusion centers the prediction around the

Table 9. Prediction performance of feature extraction techniques.

Method/PEMs	RMSE	MAE	R^2
L2GNet-4	9.082	6.592	0.901
L2GNet-8	8.351	5.911	0.916
L2GNet-16	8.354	5.913	0.916
Direct Fusion	8.206	5.808	0.918
Reliable Fusion	8.189	5.791	0.920

Table 10. Predictive performance of real industrial.

Method/PEMs	RMSE	MAE	R^2
L2GNet-4	9.522	6.846	0.882
L2GNet-8	8.568	6.213	0.896
L2GNet-16	8.516	6.413	0.895
Direct Fusion	8.500	6.308	0.897
Reliable Fusion	8.484	6.120	0.901

true RUL, while Direct Fusion is either more spread out or misaligned, indicating reduced predictive accuracy. As the system approaches failure (low RUL values), both methods show increased uncertainty, but Reliable Fusion remains more consistent and concentrated compared to Direct Fusion. For example, in plots like RUL: 19.88% and RUL: 15.87%, Direct Fusion shows a considerable spread in predictions, often failing to align with the true RUL. This reflects the challenges in capturing the system's behavior close to failure when using a straightforward fusion approach. Conversely, Reliable Fusion produces a more concentrated prediction, indicating that the method successfully identifies the most trustworthy predictions from the underlying models and focuses on them during fusion. This leads to improved accuracy even when the system is highly degraded.

The proposed reliable fusion method demonstrates a clear reduction in the spread of prediction distributions across all stages of degradation. The reduced uncertainty is critical for maintenance decision-making, as it allows operators to make more informed judgments regarding intervention times. Table 9 gives estimate the performance of each L2GNet

An industrial dataset [35] is introduced for model validation. The raw dataset was collected from the high-speed shaft of a 2 MW wind turbine, driven by a 20-tooth pinion gear. Vibration signals were continuously recorded for 50 d, with 6-second samples collected per day. An inner ring fault occurred and led to a bearing failure. Table 10 presents the Reliable Fusion performance of the model.

Single-scale models perform RUL prediction within a specific time window, which may yield favorable results at certain temporal scales (e.g. L2Gnet-4 may achieve higher accuracy in short-term predictions, while L2Gnet-16 may better capture long-term trends). However, these models are limited by their inability to leverage information across multiple scales, often resulting in higher RMSE and MAE and potentially lower R^2. As a baseline multi-scale model, Direct Fusion improves the prediction accuracy by simply combining the outputs from different scales. Although this approach benefits from multi-scale information, it fails to account for the varying confidence levels associated with each scale's prediction, which can lead to the direct accumulation of uncertainties from different scales and prevents full utilization of multi-scale fusion advantages. Consequently, Direct Fusion may yield lower RMSE and MAE than single-scale models but still exhibits considerable room for improvement.

To achieve a more reliable and accurate information fusion, the proposed Reliable Fusion method integrates predictions from each scale by weighting them according to confidence levels. This approach avoids the error accumulation, resulted from direct weighting and enables a more effective selection of critical information during multi-scale fusion, thereby further reducing prediction errors and enhancing reliability. The high R^2 values achieved by Reliable Fusion demonstrate its superior fitting performance, indicating an improved capacity to capture the effects of equipment wear and faults on RUL and thereby enhancing the model's robustness for practical applications. Reliable RUL prediction in industrial applications requires models capable of effectively capturing degradation patterns across multiple temporal scales. Table 10 presents the performance of the proposed L2GNet variants and comparative fusion methods on a real industrial dataset. The results in table 10 exhibit slightly lower performance compared to table 9, they still demonstrate strong predictive capabilities in real industrial scenarios of the model. The decrease in accuracy is expected due to the increased complexity and variability of real-world operating conditions compared to controlled experimental datasets. Despite these challenges, the proposed Reliable Fusion approach continues to deliver robust and reliable RUL predictions, achieving the lowest RMSE (8.484) and MAE (6.120), along with the highest R^2 (0.901). These results suggest that although real industrial data introduces additional uncertainties, the multi-scale fusion strategy remains effective in mitigating prediction errors and improving model generalization.

5. Conclusion

This paper developed a reliable intelligent prognostic framework, consisting of L2GNet and reliable fusion module, to balance RUL prediction accuracy and the reliability of results. The framework integrates an understanding of mechanical vibration semantics, a multi-scale semantic embedding module, and a neural network capable of both local and global feature extraction, to make RUL predictions. In addition, by quantifying prediction uncertainty, the framework provides a confidence level, and a trustworthy RUL decision fusion approach enables reliable for RUL predictions. The proposed framework's feasibility and reliability have been validated using degradation datasets from mechanical equipment, validating its predictive accuracy and its ability to quantify and manage uncertainty, leading to improved decision-making reliability, thereby demonstrating significant improvements across multiple evaluation metrics, with specific results affirming its effectiveness. Through ablation experiments, the study validates the reliability and superiority of the local-to-global feature extraction of vibration signal

semantics in RUL prediction. The framework achieves not only the best prediction results for individual mechanical failure modes but also performs well in scenarios involving mixed faults. Additionally, by using Monte Carlo dropout, the framework quantifies the uncertainty of predictions, revealing varying levels of reliability across different semantic scales. Based on this insight, a fusion method leveraging quantified uncertainty is introduced, which combines predictions from different scales through weighted fusion, resulting in more reliable RUL estimations. These findings highlight the robustness of the proposed framework and its advantage over SOTA methods in providing reliable RUL predictions. The computational efficiency of the proposed framework, the generalizability of the model, incorporating reinforcement learning or adaptive thresholding mechanisms in decision fusion are the future works of this paper.

Data availability statement

No new data were created or analysed in this study.

Acknowledgment

This paper is funded by the Horizon Europe Marie Skłodowska-Curie Postdoctoral Fellowship (ULTIMATE and DROMS-FOWT–101146961), UKRI (EPSRC EP/Y014235/1 and EPSRC EP/Z001501/1), the Natural Science Fund of China (72301299), the State Key Laboratory of Mechanical System and Vibration (Grant, No. MSV202411) and Natural Science Fund of Shanghai 'Science and Technology Innovation Action Plan' (No. 24ZR1454800).

ORCID iDs

He Li https://orcid.org/0000-0001-6429-9097
Zifei Xu https://orcid.org/0000-0003-2661-517X

References

[1] Li H, Ding Y, Sun Y, Xie M and Soares C G 2025 An intelligent failure feature learning method for failure and maintenance data management of wind turbines *Reliab. Eng. Syst. Saf.* **261** 111113
[2] Lei Y, Lin J, He Z and Zuo M J 2013 A review on empirical mode decomposition in fault diagnosis of rotating machinery *Mech. Syst. Signal Process.* **35** 108–26
[3] Matania O, Dattner I, Bortman J, Kenett R S and Parmet Y 2024 A systematic literature review of deep learning for vibration-based fault diagnosis of critical rotating machinery: limitations and challenges *J. Sound Vib.* **590** 118562
[4] Huang C, Bu S, Lee H H, Chan C H, Kong S W and Yung W K C 2024 Prognostics and health management for predictive maintenance: a review *J. Manuf. Syst.* **75** 78–101
[5] Xu X, Zhou J, Weng X, Zhang Z, He H, Steyskal F and Brunauer G 2024 A novel evidence reasoning-based RUL prediction method integrating uncertainty information *Reliab. Eng. Syst. Saf.* **250** 110250
[6] Deng W, Nguyen K T P, Medjaher K, Gogu C and Morio J 2023 Physics-informed machine learning in prognostics and health management: state of the art and challenges *Appl. Math. Model.* **124** 325–52
[7] Rezaeianjouybari B and Shang Y 2020 Deep learning for prognostics and health management: state of the art, challenges, and opportunities *Measurement* **163** 107929
[8] Vrignat P, Kratz F and Avila M 2022 Sustainable manufacturing, maintenance policies, prognostics and health management: a literature review *Reliab. Eng. Syst. Saf.* **218** 108140
[9] Li H, Zhang Z, Li T and Si X 2024 A review on physics-informed data-driven remaining useful life prediction: challenges and opportunities *Mech. Syst. Signal Process.* **209** 111120
[10] Xu Z, Zhao K, Wang J and Bashir M 2024 Physics-informed probabilistic deep network with interpretable mechanism for trustworthy mechanical fault diagnosis *Adv. Eng. Inform.* **62** 102806
[11] Xu Z, Bashir M, Liu Q, Miao Z, Wang X, Wang J and Ekere N 2023 A novel health indicator for intelligent prediction of rolling bearing remaining useful life based on unsupervised learning model *Comput. Ind. Eng.* **176** 108999
[12] Chen D, Wu F, Wang Y and Qin Y 2024 A lognormal-normal mixture model for unsupervised health indicator construction and its application into gear remaining useful life prediction *Mech. Syst. Signal Process.* **220** 111699
[13] Zhou J, Qin Y, Luo J and Zhu T 2023 Remaining useful life prediction by distribution contact ratio health indicator and consolidated memory GRU *IEEE Trans. Ind. Inform.* **19** 8472–83
[14] Chen D, Qin Y, Wang Y and Zhou J 2020 Health indicator construction by quadratic function-based deep convolutional auto-encoder and its application into bearing RUL prediction *ISA Trans.* **114** 44–56
[15] Cheng Y, Hu K, Wu J, Zhu H and Shao X 2021 A convolutional neural network based degradation indicator construction and health prognosis using bidirectional long short-term memory network for rolling bearings *Adv. Eng. Inform.* **48** 101247
[16] Guo J, Wang Z, Li H, Yang Y, Huang C G, Yazdi M and Kang H S 2024 A hybrid prognosis scheme for rolling bearings based on a novel health indicator and nonlinear Wiener process *Reliab. Eng. Syst. Saf.* **245** 110014
[17] Wen L, Su S, Wang B, Ge J, Gao L and Lin K 2023 A new multi-sensor fusion with hybrid convolutional neural network with Wiener model for remaining useful life estimation *Eng. Appl. Artif. Intell.* **126** 106934
[18] Feng K, Ji J C and Ni Q 2023 A novel gear fatigue monitoring indicator and its application to remaining useful life prediction for spur gear in intelligent manufacturing systems *Int. J. Fatigue* **168** 107459
[19] Chen D, Cai W, Yu H, Wu F and Qin Y 2023 A novel transfer gear life prediction method by the cross-condition health indicator and nested hierarchical binary-valued network *Reliab. Eng. Syst. Saf.* **237** 109390
[20] Cao L, Zhang H, Meng Z and Wang X 2023 A parallel GRU with dual-stage attention mechanism model integrating uncertainty quantification for probabilistic RUL prediction of wind turbine bearings *Reliab. Eng. Syst. Saf.* **235** 109197
[21] Gupta M, Wadhvani R and Rasool A 2023 A real-time adaptive model for bearing fault classification and remaining useful life estimation using deep neural network *Knowl.-Based Syst.* **259** 110070
[22] Wu B, Zhang X, Shi H and Zeng J 2024 Failure mode division and remaining useful life prognostics of multi-indicator systems with multi-fault *Reliab. Eng. Syst. Saf.* **244** 109961
[23] Liu Z, Li H, Lin J, Jiao J, Zhang B, Liu H and Li W 2024 GARCH family models oriented health indicators for bearing degradation monitoring *Meas. J. Int. Meas. Confed.* **231** 114604

[24] Hervé de Beaulieu M, Jha M S, Garnier H and Cerbah F 2024 Remaining useful life prediction based on physics-informed data augmentation *Reliab. Eng. Syst. Saf.* **252** 110451
[25] Li Q, Yan C, Chen G, Wang H, Li H and Wu L 2022 Remaining useful life prediction of rolling bearings based on risk assessment and degradation state coefficient *ISA Trans.* **129** 413–28
[26] Zhao X, Yang Y, Huang Q, Fu Q, Wang R and Wang L 2025 Rolling bearing remaining useful life prediction method based on vibration signal and mechanism model *Appl. Acoust.* **228** 110334
[27] Kumar A, Parkash C, Vashishtha G, Tang H, Kundu P and Xiang J 2022 State-space modeling and novel entropy-based health indicator for dynamic degradation monitoring of rolling element bearing *Reliab. Eng. Syst. Saf.* **221** 108356
[28] Qin Y, Yang J, Zhou J, Pu H, Zhang X and Mao Y 2023 Dynamic weighted federated remaining useful life prediction approach for rotating machinery *Mech. Syst. Signal Process.* **202** 110688
[29] Cao Y, Ding Y, Jia M and Tian R 2021 A novel temporal convolutional network with residual self-attention mechanism for remaining useful life prediction of rolling bearings *Reliab. Eng. Syst. Saf.* **215** 107813
[30] de Pater I and Mitici M 2023 Developing health indicators and RUL prognostics for systems with few failure instances and varying operating conditions using a LSTM autoencoder *Eng. Appl. Artif. Intell.* **117** 105582
[31] Balamurugan R, Takale D G, Parvez M M and Gnanamurugan S 2024 A novel prediction of remaining useful life time of rolling bearings using convolutional neural network with bidirectional long short term memory *J. Eng. Res.* (https://doi.org/10.1016/j.jer.2024.05.005)
[32] Yang Q, Tang B, Deng L, Zhu P and Ming Z 2024 WTFormer: RUL prediction method guided by trainable wavelet transform embedding and lagged penalty loss *Adv. Eng. Inform.* **62** 102710
[33] Tian H, Yang L and Ju B 2023 Spatial correlation and temporal attention-based LSTM for remaining useful life prediction of turbofan engine *Measurement* **214** 112816
[34] Qu G, Qiu T, Si Y, Yuan Q, Ma Q and Wang C 2022 Remaining useful life prediction for aero-engine based on hybrid CNN-GRU model *2022 IEEE Int. Conf. on Unmanned Systems* pp 1523–8
[35] MathWorks 2024 WindTurbine HighSpeedBearingPrognosis-Data GitHub repository (available at: https://github.com/mathworks/WindTurbineHighSpeedBearingPrognosis-Data) (Accessed 20 February 2024)