Signal, Image and Video Processing (2025) 19:1072
https://doi.org/10.1007/s11760-025-04617-3

ORIGINAL PAPER

iTransformer-LSTM dual-stream architecture for rolling bearing remaining useful life prediction

Zhigang Chen[1] · Mengyao Shi[1] · Yanxue Wang[1] · Longqiao Chen[1]

Received: 24 April 2025 / Revised: 1 August 2025 / Accepted: 4 August 2025

Abstract
The full-lifecycle degradation data of rolling bearings exhibit large volume, strong heterogeneity, pronounced nonlinearity, and marked non-stationarity. Conventional deep learning approaches for remaining useful life prediction suffer from limited feature representation and poor modeling of temporal dependencies, leading to suboptimal accuracy and limited robustness. To overcome these limitations, this study proposes a dual-stream iTransformer–LSTM architecture for remaining useful life prediction of rolling bearings. First, intrinsic mode functions are extracted from raw vibration signals via empirical mode decomposition, followed by the construction of a multidimensional degradation feature set through combined time–frequency domain analysis. Subsequently, iTransformer and LSTM modules are employed to extract local temporal variations and long-term dependencies, respectively, and a cross-attention mechanism is introduced to facilitate feature fusion. Finally, the Kolmogorov–Arnold Network is employed for high-dimensional feature mapping, enhancing nonlinear representation and improving predictive performance. Experimental validation on the IEEE PHM2012 benchmark dataset demonstrates that, compared to several existing prediction methods, the proposed method reduces the mean absolute error by 5.8 to 50.5% and the root mean square error by 7.32 to 49.33% across different bearing samples, thereby confirming the effectiveness and feasibility of the approach.

Keywords Rolling bearing life prediction · iTransformer-LSTM · Cross-attention mechanism · Kolmogorov–Arnold networks

1 Introduction

With the rapid advancement of industrialization and manufacturing technologies, mechanical systems have been widely utilized in various engineering domains. Consequently, condition monitoring of critical components, particularly rolling bearings, has garnered increasing attention in recent research [1]. Rolling bearings are vulnerable to complex and variable operating conditions over prolonged service periods, resulting in progressive performance degradation and eventual failure. Remaining useful life prediction represents a key task in prognostics and health management (PHM), facilitating early fault detection to mitigate operational risks and reduce maintenance costs [2].

Existing RUL prediction approaches are typically classified into model-based and data-driven categories. Model-based approaches depend on explicit modeling of degradation mechanisms and operational conditions [3]. In contrast, data-driven approaches employ self-learning algorithms to extract degradation patterns from large-scale monitoring data, thereby reducing reliance on prior knowledge. With the rapid development of artificial intelligence, deep learning has emerged as a key paradigm in RUL prediction, owing to its powerful capacity for modeling complex nonlinear relationships [4]. Convolutional neural networks (CNNs), leveraging local receptive fields and weight-sharing, were initially adopted as mainstream architectures for extracting spatial features in RUL prediction tasks. Ren et al. [5] developed a CNN-based model for bearing RUL prediction and introduced a smoothing technique to mitigate discontinuities in prediction outputs. Wang et al. [6] proposed a bearing RUL prediction method using a deep convolutional autoencoder and 1D CNN, with a health indicator based on self-organizing maps to improve degradation characterization. However, the

✉ Zhigang Chen
zdketi@163.com

[1] School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

Published online: 10 September 2025

degradation process of mechanical components typically evolves as long-duration continuous signals, requiring models to effectively capture temporal dependencies across time.

Due to the fixed receptive field of CNNs, they are inherently limited in modeling such dependencies, which often leads to fragmented feature representations and reduced prediction stability when dealing with long-term degradation processes.

To address the limitations of CNNs in capturing long-range temporal dependencies, subsequent studies have explored the use of recurrent neural networks (RNNs), particularly long short-term memory (LSTM) architectures. LSTM, an advanced variant of RNNs [7], incorporates memory cells and gating mechanisms, enabling it to effectively learn temporal dependencies and extract latent features from sequential data. Compared with CNNs, LSTM exhibits superior performance in capturing localized temporal degradation patterns, making it more suitable for modeling the degradation processes of mechanical systems. Liu et al. [8] proposed an LSTM-based RUL prediction method incorporating regular interval sampling and locally weighted scatterplot smoothing for data preprocessing. Li et al. [9] developed a hybrid model combining LSTM and the Elman neural network, in which the input signal was first decomposed using empirical mode decomposition. However, these studies primarily focus on modeling local temporal dependencies, while often overlooking the global degradation trends that evolve across the entire input sequence. To simultaneously capture both local and global degradation information, some researchers have proposed LSTM models enhanced with attention mechanisms. Al-Dahidi et al. [10] proposed an RUL prediction method combining LSTM and multi-head self-attention, which effectively captures local temporal dependencies and global features, thereby enhancing prediction accuracy for high-reliability equipment.

To achieve unified modeling of global temporal dependencies, recent studies have turned to attention-based architectures such as the Transformer, which offer greater flexibility in capturing global contextual information across sequences. As a novel architecture [11], the Transformer employs an encoder–decoder framework driven by stacked self-attention mechanisms, enabling more effective modeling of global dependencies and efficient parallel computation. Structurally, it offers clear advantages over conventional approaches that merely append attention modules to existing models. Chen et al. [12] proposed a Transformer-based method for bearing RUL prediction by integrating a spatial attention–enhanced CNN with a Transformer, while Mu et al. [13] developed a Random Forest–Transformer–LSTM hybrid model for aircraft engine RUL prediction. Although these methods demonstrate the effectiveness of combining global and local features, the standard Transformer still exhibits limitations in handling multivariate time series. It typically embeds multiple variables into a single token at each time step, which restricts its ability to model inter-variable dependencies. To address these limitations, the iTransformer architecture [14] was introduced. The iTransformer architecture improves multivariate dependency modeling by treating each variable sequence as an independent token, enabling more effective attention-based feature learning, and further enhancing the extraction of global dependencies across complex time-series data.

The inability to simultaneously capture global degradation trends and localized temporal dynamics may lead to incomplete degradation representation, reduced prediction accuracy, and degraded robustness in RUL forecasting. To address this challenge, this paper proposes a novel dual-stream architecture that integrates the iTransformer and LSTM networks. The iTransformer, with its inverted architecture that treats variable dimensions as sequence tokens, employs a standard self-attention mechanism to capture global inter-variable dependencies. In parallel, the LSTM network is designed to learn localized temporal degradation patterns. A cross-attention mechanism is introduced to enable dynamic fusion of global and local features, while a KAN is incorporated at the output layer to enhance nonlinear representation and improve predictive accuracy. Experimental results indicate that the proposed method outperforms several existing prediction models across multiple performance metrics.

The main contributions of this study are as follows:

(1) An innovative iTransformer–LSTM dual-stream architecture is proposed, which decouples and simultaneously models global inter-variable dependencies and local temporal degradation dynamics. This design enables comprehensive representation of complex multivariate time-series characteristics in rolling bearing degradation processes.
(2) A cross-attention fusion mechanism is incorporated to dynamically align and integrate global and local feature streams, enhancing complementary information extraction. Furthermore, the Kolmogorov–Arnold Network (KAN) is introduced at the output layer to perform flexible nonlinear mapping, improving prediction accuracy and interpretability.
(3) Experimental results on the IEEE PHM2012 dataset show that the proposed method achieves better performance than multiple baseline models in terms of MAE, RMSE, and R^2, and the effectiveness of the dual-stream architecture and each module is further validated through ablation experiments and feature visualization.

The theoretical background is provided in Sect. 2. Section 3 details the overall framework design and dual-stream processing strategy. Experimental setups, dataset

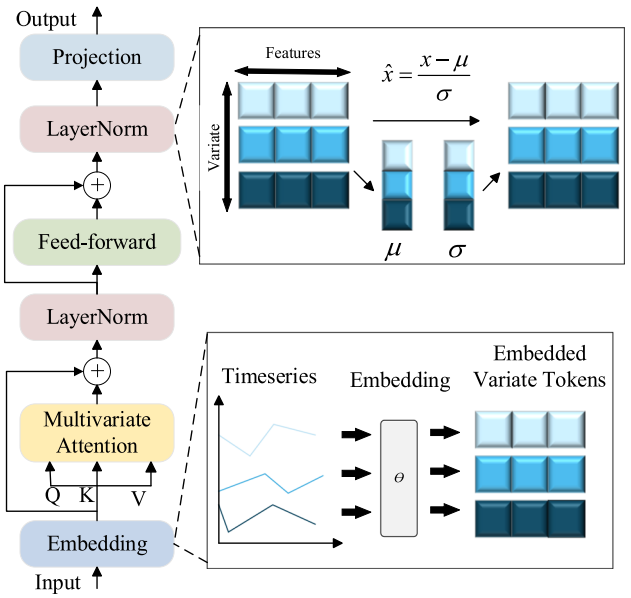

Fig. 1 iTransformer architecture

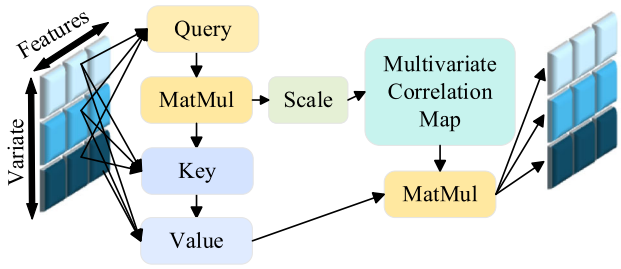

Fig. 2 Multivariate attention mechanism

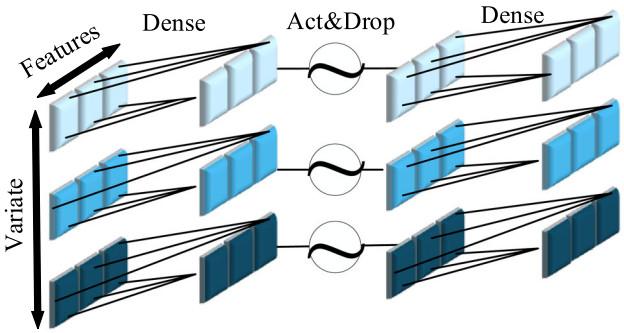

Fig. 3 FNN

descriptions, and comparative analyses are discussed in Sect. 4. The conclusions and future research directions are provided in Sect. 5.

2 Theoretical fundamentals

2.1 iTransformer model

The iTransformer serves as the global modeling branch of the proposed dual-stream architecture, focusing on capturing long-range temporal dependencies and complex inter-variable relationships. It adopts an encoder-only architecture, consisting of an embedding layer, a projection layer, and a Transformer module. The module incorporates a multivariate attention mechanism, a feedforward neural network (FFN), and layer normalization (LayerNorm), thereby enhancing the ability to model inter-variable correlations. The module structure is illustrated in Fig. 1.

Unlike traditional Transformers that model temporal dependencies via query–key correlations, the iTransformer treats each time series as an independent token and captures temporal representations through a self-attention mechanism. Q, K, and V are obtained through linear projections, and attention weights are computed prior to the Softmax operation to reveal inter-variable dependencies, as illustrated in Fig. 2.

The feedforward neural network (FNN) component of the module includes an activation layer and two 1D convolutional layers: the first encodes historical data, and the second decodes it to form global representations. This tokenization strategy enables effective modeling of global temporal features across variables, as shown in Fig. 3.

To ensure training stability and reduce distortions from heterogeneous data, normalization is applied. In the conventional Transformer architecture, multivariate data at the same timestamp are normalized to integrate information; however, when the collected data do not correspond to the same event, spurious interactions may be introduced due to non-causal or delayed correlations. In contrast, the inverted model normalizes univariate series by treating each sequence as a token, standardizing to a Gaussian distribution. This not only mitigates inconsistencies from asynchronous measurements but also enhances global dependency extraction across time and variables.

The LayerNorm formula is as follows:

$$LayerNorm(H) = \left\{ \frac{h_n - Mean(h_n)}{\sqrt{Var(h_n)}} | n = 1, \cdots, N \right\} \tag{1}$$

where H denotes the input sequence with h_t representing the observation at timestep t, Mean(h_n) and Var(h_n) are the mean and variance computed across all N timesteps or features $\{h_n\}_{n=1}^{N}$, and the normalization ensures zero-mean and unit-variance outputs while preserving temporal dependencies.

The computational process of the entire model is described as follows:

$$\begin{aligned} &h_n^0 = Embedding(X_i),\ i = 1, 2, \cdots, m, \\ &\hat{h}_i^l = LN\left(h_i^l + SA\left(h_i^l\right)\right),\ l = 0, 1, \cdots, L-1, \\ &\hat{h}_i^{l+1} = LN\left(FNN\left(\hat{h}_i^l\right) + \hat{h}_i^l\right),\ l = 0, 1, \cdots, L-1, \\ &\hat{Y} = \Pr ojection\left(h_i^l\right) \end{aligned} \tag{2}$$

where Embedding(•) is used to embed the input sequence of length l, Projection(•) is employed to map the input to the target prediction, L denotes the number of layers in the iTransformer, LN(•) refers to layer normalization, and FFN(•) represents the feed-forward network.

2.2 LSTM model

As the local modeling branch in the proposed dual-stream architecture, the Long Short-Term Memory (LSTM) network is responsible for capturing fine-grained temporal degradation patterns from sequential data. LSTM effectively manages long sequential data by modeling long-term dependencies through its gating mechanism. Specifically, LSTM receives the hidden state h^t and cell state C^t from the previous time step as inputs and processes data sequentially through the gating mechanism to determine when to store, discard, or retrieve information. This architecture allows LSTM to preserve localized temporal context, making it well-suited for modeling gradual degradation and dynamic behaviors in time-series data. The gating mechanism calculation formula is presented as follows:

$$\begin{aligned} &i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \\ &f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \\ &\tilde{C} = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \\ &o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\ &C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C} \\ &h_t = o_t \odot \tanh(c_t) \end{aligned} \tag{3}$$

where i_t denotes the output of the input gate, f_t represents the output of the forget gate, $\widetilde{C}_t$ expresses the candidate cell state, o_t indicates the output of the output gate, C_t defines the cell state at the current time step, and h_t specifies the current hidden state.

2.3 KAN model

KAN (Kolmogorov-Arnold Network) is a neural architecture inspired by the Kolmogorov-Arnold representation theorem [15], fundamentally differing from traditional MLPs in its parameterization. While MLPs apply fixed activation functions at nodes, KANs implement learnable activation functions on edges (weights), enabling dynamic feature adaptation. This design is particularly advantageous for modeling the complex nonlinear interactions between fused global–local features, as the theorem guarantees that any continuous multivariate function can be decomposed into univariate function compositions:

$$f(x) = f(x_1, x_2, \cdots, x_n) = \sum_{q=1}^{2n+1} \phi_q \left(\sum_{p=1}^{n} \phi_{q,p}\left(x_p\right) \right) \tag{4}$$

where $\varnothing_{q,p}(\bullet)$ denotes the univariate function that transforms each input variable, defined as $\varnothing_{q,p}[0,1] \rightarrow R$. By leveraging this theoretical foundation, KAN enhances both prediction accuracy and generalization in high-dimensional spaces. The detailed derivation and full formulation of the KAN layer are provided in Online Resource 1.

3 Overall model architecture

This study proposes a novel temporal modeling framework, termed iTransformer–LSTM. Figure 4 illustrates the architecture of the proposed iTransformer–LSTM model. This framework adopts a dual-stream design that simultaneously processes multivariate time-series data through an iTransformer encoder and an LSTM network. The model is developed to effectively capture both global degradation trends and localized temporal dynamics in bearing signals, thereby improving the accuracy and robustness of RUL prediction.

In the iTransformer branch, each variable sequence is treated as a token, and the self-attention mechanism is applied to model inter-variable dependencies and long-range temporal relationships. This structure allows the model to decouple the time dimension from variable interactions, improving its ability to learn global degradation patterns. The encoder comprises stacked transformer layers with multi-head attention and a feedforward network, enabling deep global feature extraction. In parallel, the LSTM branch takes the same input and focuses on modeling sequential patterns. The LSTM consists of multiple layers and hidden units, where gating mechanisms regulate the flow of temporal information. This design enables the extraction of localized degradation features and preserves the temporal continuity inherent in mechanical degradation signals.

The outputs from the iTransformer and LSTM branches are fused through a cross-attention mechanism that dynamically treats the iTransformer's global embeddings as queries and LSTM's local features as keys/values. By computing scaled attention weights with softmax normalization and dropout regularization, this mechanism selectively enhances

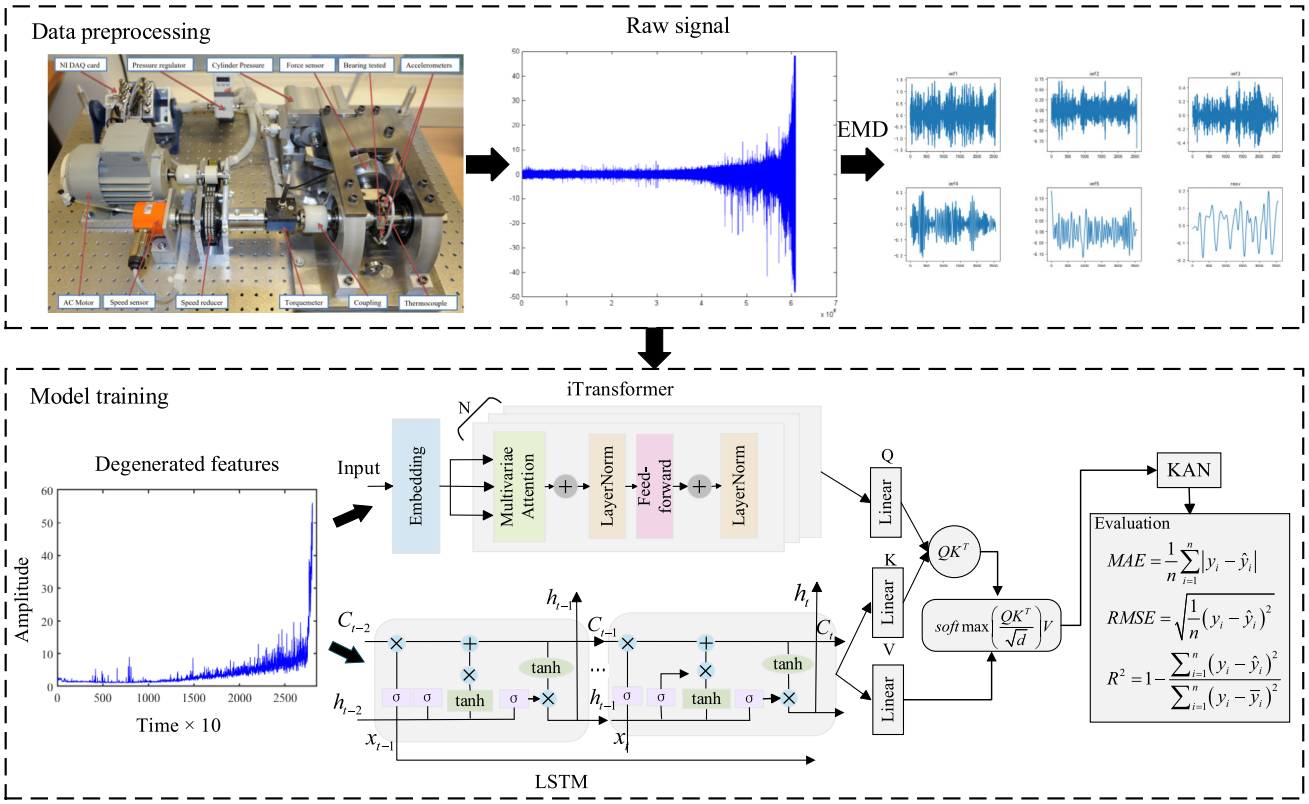

Fig. 4 Overall model architecture

complementary temporal features while suppressing redundancy—evidenced by the improved inter-stage separation in Fig. 9 where fused features show minimal cluster overlap compared to single-stream outputs. The refined features are then processed by a Kolmogorov-Arnold Network (KAN) whose learnable edge activation functions provide adaptive nonlinear mappings, ultimately generating the RUL prediction through the first time step of KAN's output. This unified architecture effectively combines global attention modeling, local temporal dynamics capture, redundancy-aware feature fusion, and flexible nonlinear representation to comprehensively characterize rolling bearing degradation processes.

4 Experimental section

4.1 Experimental platform description

In this experiment, the PHM2012 Challenge dataset provided by the FEMTO-ST Institute was utilized. The data were collected using the PRONOSTIA accelerated bearing degradation platform, which simulates bearing wear under variable operating conditions while continuously recording health-related signals such as rotational speed, load, temperature, and vibration. A photo of the experimental platform is provided in Online Resource 2. The dataset includes three different operating conditions with a sampling frequency of 25.6 kHz, as shown in Table 1. Data are recorded every 10 s, with each sample containing 2,560 sampling points in both the horizontal and vertical directions. Studies have shown that, compared to vertical signals, horizontal signals are more effective in detecting bearing conditions [16]. Therefore, in this study, horizontal bearing data under an operating condition of 1800 r/min and a load of 4000 N were used for experimental analysis. Based on the PHM2012 rolling bearing dataset obtained from the challenge [17], the vibration signals generated by Bearing1-1 and Bearing1-2 under the same operating conditions were found to be the most representative. Therefore, these data were used as the training set,

Table 1 PHM2012 bearing dataset

Test Condition	Working Condition 1	Working Condition 2	Working Condition 3
Load/N	4000	4200	5000
Rotational Speed/(r·min^{-1})	1800	1650	1500
Test Bearings	1–1–1–7	2–1–2–7	3–1–3–3

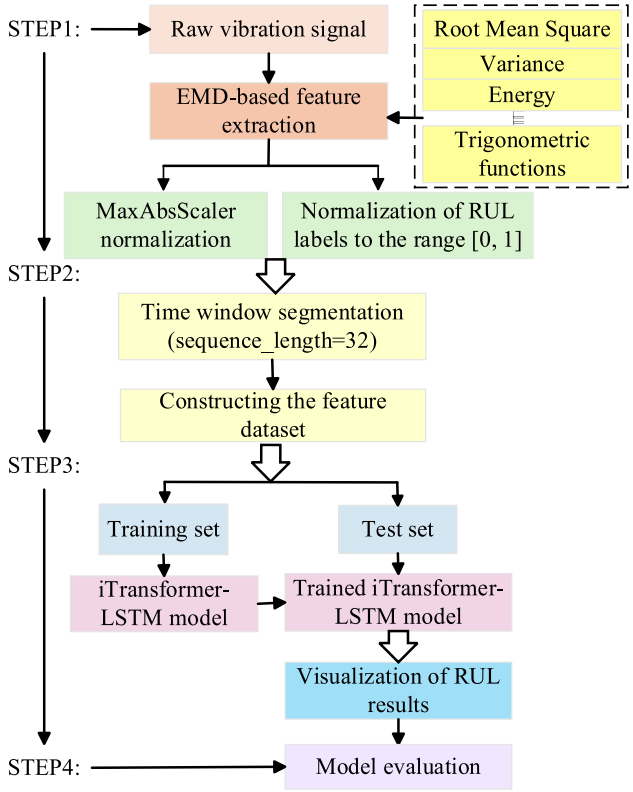

Fig. 5 Experimental procedure

while Bearing1-3 and Bearing1-4 were designated as the test set.

4.2 Experimental procedure

The overall experimental procedure involves data preprocessing, dual-stream feature extraction, prediction via KAN, and final error evaluation, as outlined in Fig. 5. Detailed textual descriptions of each step are provided in Online Resource 3.

4.3 Features and performance evaluation metrics

Time-domain features capture statistical properties and instantaneous signal states, while frequency-domain features reflect periodicity and hidden fault characteristics. The combination provides a comprehensive view of bearing behavior during its entire lifecycle [19]. Under normal conditions, vibration signals show low energy, near-zero mean, and low RMS, indicating stable operation. As degradation progresses, impact-induced components increase, raising energy, RMS, and mean values. These trends are illustrated in Fig. 6 and Fig. 7 for Bearing1_1.

As shown in the figures, towards the end of the bearing's lifespan, key indicators such as characteristic values and RMS exhibit a distinct degradation trend. This result

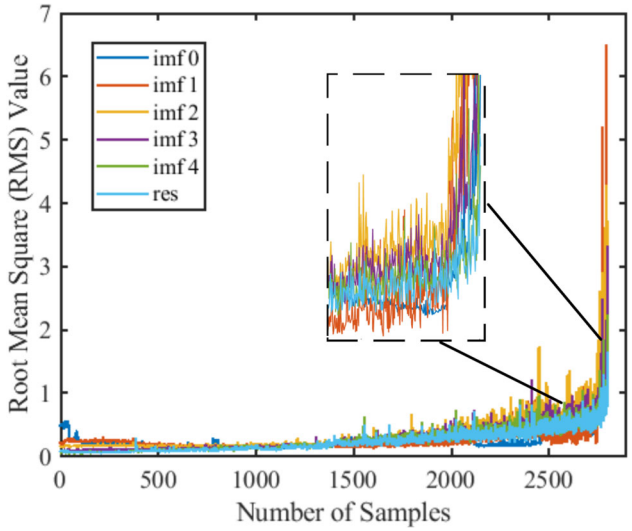

Fig. 6 Comparison of RMS trends

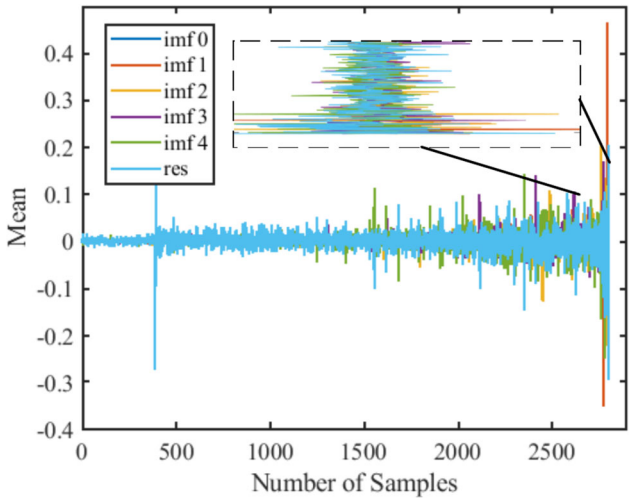

Fig. 7 Comparison of Mean trends

demonstrates the feasibility of predicting the remaining useful life (RUL) of bearings based on time-domain and frequency-domain features, effectively capturing the degradation process.

To provide a more intuitive assessment of prediction performance, root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) are selected as evaluation metrics [20]. The detailed formulations of these metrics are provided in Online Resource 4.

4.4 Parameter configuration

The model training was conducted according to predefined procedures, with hyperparameter selection and optimization strategies detailed in Online Resource 5 (see Tables 1 and 2 therein). Hyperparameter optimization was performed

Table 2 Model comparison results

		LSTM	iTransformer	Informer	SCINet	CNN-LSTM	Transformer-LSTM	iTransformer-TCN	The proposed method
Bearing1-3	MAE	0.092	0.086	0.081	0.079	0.074	0.068	0.099	0.049
	RMSE	0.111	0.102	0.095	0.093	0.106	0.124	0.104	0.073
	R^2	0.611	0.682	0.720	0.734	0.751	0.530	0.601	0.876
Bearing1-4	MAE	0.147	0.140	0.138	0.135	0.197	0.250	0.157	0.130
	RMSE	0.182	0.174	0.169	0.164	0.198	0.300	0.174	0.152
	R^2	0.369	0.448	0.462	0.480	0.420	0.445	0.359	0.627

using the Optuna framework [21], a Python-based automated Hyperparameter Optimization (HPO) tool, with the explicit objective of minimizing validation MAE. Based on the parameter search space defined in Online Resource 5 (Table 2), the optimization process employed the Tree-structured Parzen Estimator algorithm across 100 parallel GPU-accelerated trials. Optimization terminated when 20 consecutive trials exhibited less than 0.5% MAE improvement, and the resulting optimal configuration is also provided in Online Resource 5.

4.5 Analysis of results from different RUL prediction methods

The RUL prediction results for bearings Bearing1-1 to Bearing1-4 are illustrated in Fig. 8. The results suggest that the proposed model is capable of effectively capturing degradation characteristics of bearings and generating relatively accurate RUL predictions. To comprehensively evaluate the performance of the proposed method, several representative models from the literature—Informer [22], SCINet [23], CNN-LSTM [24], Transformer-LSTM [25], and iTransformer-TCN [26]—were selected for comparison. Informer adopts a ProbSparse self-attention mechanism to efficiently model long-sequence dependencies, making it well-suited for time-series forecasting with high temporal complexity. SCINet employs a recursive multi-scale decomposition architecture to capture hierarchical temporal features across different scales. The CNN-LSTM model combines CNNs for local feature extraction with LSTM for temporal sequence modeling. The Transformer-LSTM model utilizes self-attention to capture long-range dependencies, which are then modeled by LSTM to enhance temporal representation. The iTransformer-TCN model first extracts temporal features via the iTransformer and then refines long-term dependencies through the temporal convolutional network (TCN).

All comparative models were trained under consistent input dimensions and time steps to ensure fairness in experimental comparison. To reduce the influence of randomness, average results from ten independent runs were reported as the final evaluation. Performance metrics of each model are summarized in Table 2.

Experimental results suggest that the proposed iTransformer–LSTM hybrid architecture achieves superior performance in predicting the remaining useful life (RUL) of bearings. As shown in Table 2, the model achieves the lowest MAE (0.049) and RMSE (0.073), along with the highest R^2 (0.876), outperforming all baseline models on the Bearing1-3 dataset. Specifically, compared to Informer, the proposed method reduces MAE and RMSE by approximately 39.5 and 23.2%, respectively, while improving R^2 by 21.7%. Relative to SCINet, the reductions in MAE and RMSE are 38.0 and

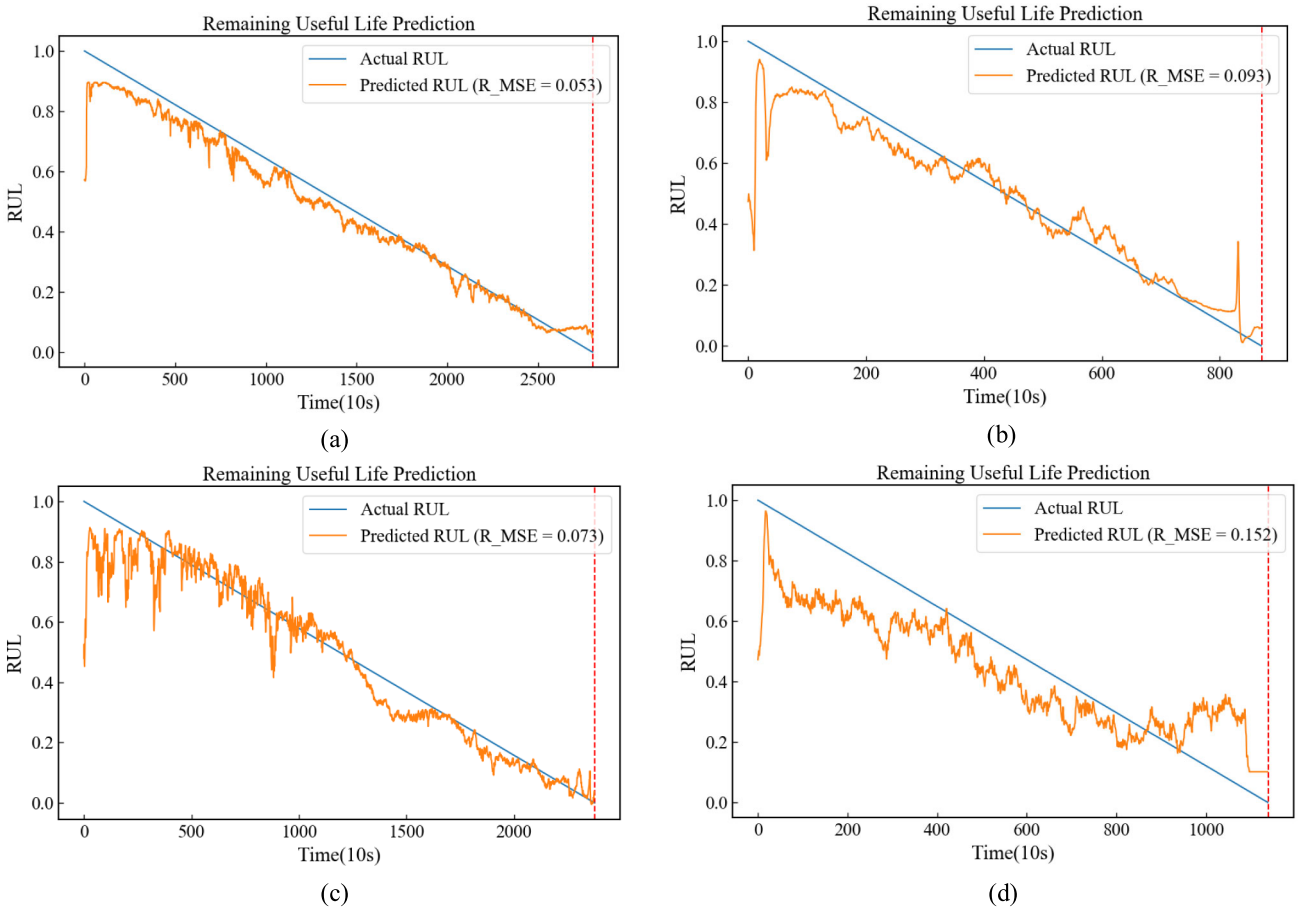

Fig. 8 Prediction results: **a** Bearing1-1; **b** Bearing1-2; **c** Bearing1-3; **d** Bearing1-4

21.5%, and R^2 increases by 19.3%. Compared with CNN-LSTM, MAE and RMSE are reduced by 33.8 and 31.1%, and R^2 improves by 16.6%. When compared to Transformer-LSTM, the proposed model achieves a 27.9% reduction in MAE, a 41.1% reduction in RMSE, and a 65.3% increase in R^2. Similarly, it outperforms iTransformer-TCN, reducing MAE and RMSE by 50.5% and 29.8%, respectively, and increasing R^2 by 45.8%.

On the more challenging Bearing1-4 dataset, the proposed model achieves an MAE of 0.130, RMSE of 0.152, and R^2 of 0.627. These results correspond to an average relative reduction in MAE and RMSE of approximately 23.7% and 19.6%, respectively, and an average R^2 improvement of 38.2% over the baseline models. These findings demonstrate that the dual-stream design—combining global attention modeling from the iTransformer with local temporal feature extraction from the LSTM—effectively improves prediction accuracy, robustness, and the ability to model complex degradation processes.

4.6 Ablation study on dual-stream architecture

To evaluate the effectiveness of the proposed dual-stream architecture, an ablation study was conducted by constructing two single-stream variants: an LSTM-only model for capturing local temporal dependencies and an iTransformer-only model for modeling global inter-variable attention. These variants share the same input and output structure but omit the cross-attention and KAN modules. As shown in Table 2, the dual-stream model consistently outperforms both baselines across all metrics. On Bearing1-3, the MAE of the dual-stream model is 0.049, which is notably lower than 0.092 and 0.068 obtained by LSTM and iTransformer, respectively. Similar improvements are observed in RMSE and R^2 metrics, demonstrating that fusing local and global degradation representations leads to more accurate RUL prediction.

The computational cost was further assessed in terms of trainable parameters and average training time per epoch. As reported in Table 3, the dual-stream model introduces higher complexity (34.83×10^5 parameters, 6.48 s/epoch) than iTransformer (23.55×10^5, 4.25 s) and LSTM ($10.04 \times$

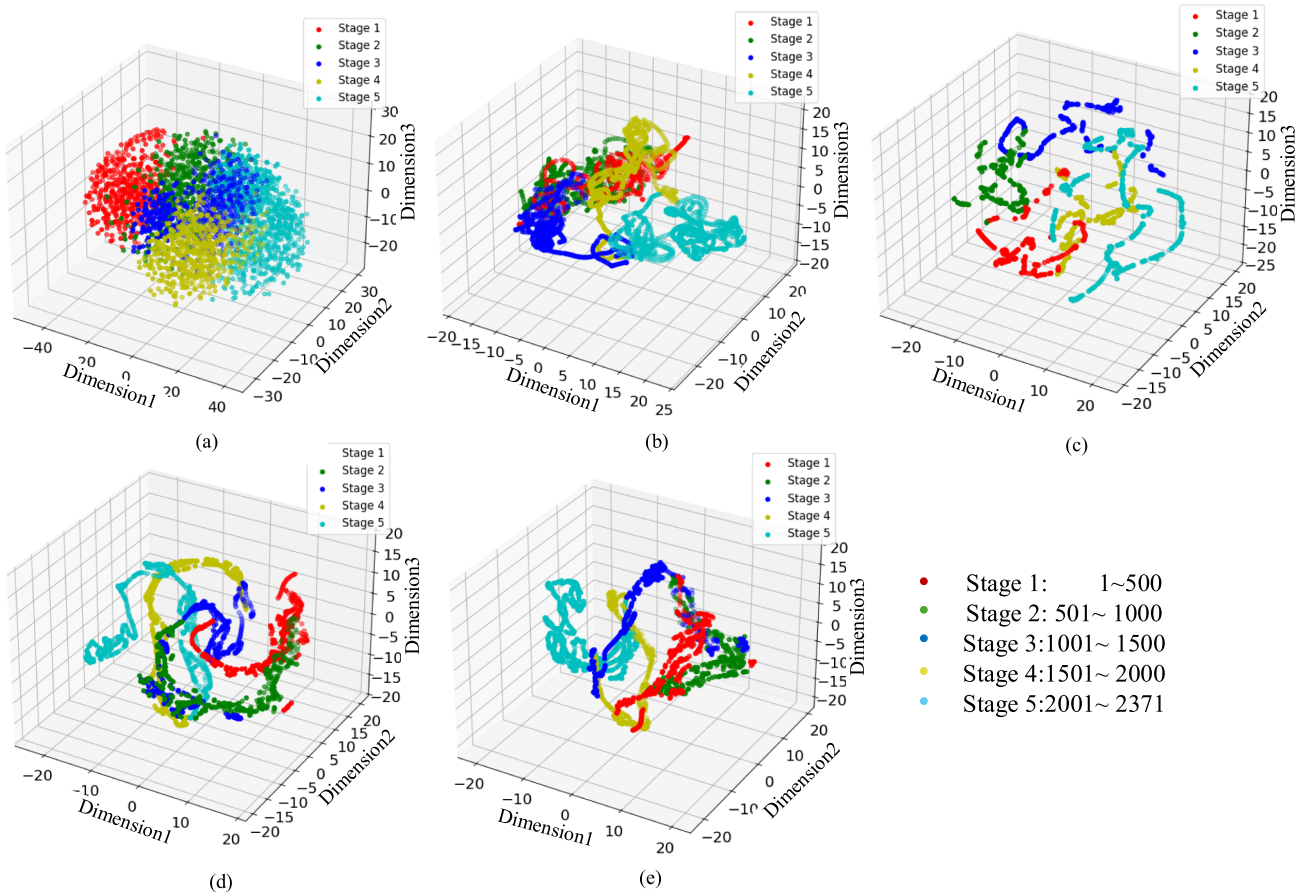

Fig. 9 Feature visualization results: **a** Input features; **b** LSTM layer; **c** iTransformer layer; **d** Cross-attention fusion layer; **e** KAN-enhanced layer

Table 3 Comparison of average training time per epoch and parameter count across different models

Model	Epoch time (s)	Parameters($\times 10^5$)
LSTM	2.26	10.04
iTransformer	4.25	23.55
iTransformer-LSTM	6.48	34.83

10^5, 2.26 s), but the accuracy improvements justify the added overhead.

4.7 Visualization of temporal feature extraction results

To investigate the distribution characteristics of the features extracted by the iTransformer-LSTM model, three-dimensional visualisation of high-dimensional features from different hidden layers was performed using the t-distributed stochastic neighbour embedding (t-SNE) algorithm. Signal samples from Bearing1-3 in the PHM 2012 dataset were selected for analysis. After time-step processing of the raw signal data, a total of 2,371 samples were obtained from Bearing 1_3. These samples were divided into five groups based on the chronological order of failure progression to represent the full lifecycle of the bearing. This segmentation strategy effectively reflects the continuous degradation process and facilitates the visualization of feature evolution across different degradation stages. Figure 9 presents the feature distributions of key layers within the model, including the input layer, LSTM layer, iTransformer layer, feature fusion layer, and the KAN-enhanced representation layer. Different colours and marker shapes correspond to five temporal clusters, highlighting the evolution of feature representations across various degradation stages.

The visualization results in Fig. 9 demonstrate the progressive transformation of feature representations across the model layers and reveal the distinct learning characteristics of the two branches before fusion. The input features are highly overlapped, making it difficult to distinguish degradation stages. After processing by the LSTM branch, a clustering trend begins to emerge. However, the class boundaries remain ambiguous, indicating that LSTM alone struggles to encode the overall degradation progression. In contrast, the

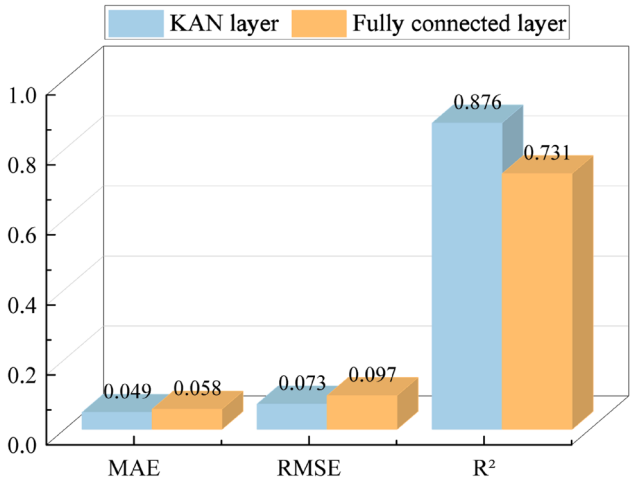

Fig. 10 Comparison of the KAN layer and fully connected layer

iTransformer branch, which models each variable sequence as a token and leverages global self-attention, learns long-range temporal dependencies and inter-variable correlations. This results in a more stretched temporal distribution with increased separation between early and late failure stages. These differences indicate that each stream encodes complementary aspects of the degradation process.

The fused features, obtained via the cross-attention mechanism, exhibit enhanced intra-class compactness and inter-stage separation. This suggests that the fusion module successfully aligns and integrates the LSTM's localized features with the iTransformer's global embeddings, while suppressing redundant patterns. Finally, the KAN layer further transforms the fused representations into a more orthogonal and discriminative space, maximizing inter-cluster separation and improving interpretability. This multi-stage decoupling–fusion–mapping process visually validates the model's multi-scale feature extraction strategy and supports its superior prediction performance.

4.8 Comparative analysis of the output layer

To validate the model's effectiveness, the performance of the output layer using KAN and fully connected layers was compared, with data sourced from bearing1-3. Experimental results indicate that the KAN layer achieves lower MAE (0.049) and RMSE (0.084) compared to the fully connected layer, representing reductions of 15.5% and 24.7%, respectively. The R^2 value of the KAN layer reaches 0.876, compared to 0.731 for the fully connected layer, suggesting enhanced effectiveness in capturing data variability. The results are shown in Fig. 10. This difference arises from structural design: the KAN layer captures complex features through adaptive basis functions and dynamic adjustments, while the fully connected layer relies on static nonlinearity, which is prone to dimensional collapse.

5 Conclusion

In this study, an iTransformer-LSTM dual-stream architecture is developed to estimate the remaining useful life of bearings. Its predictive performance is evaluated using the publicly available PHM2012 bearing degradation dataset. The key findings are summarized as follows:

(1) A model based on the iTransformer-LSTM dual-stream architecture, integrating a multidimensional degradation feature set with time–frequency analysis, was proposed. The degradation characteristics of rolling bearings were extracted, contributing to improved accuracy in the prediction of remaining useful life (RUL).
(2) The iTransformer-LSTM architecture was developed to address challenges such as multivariate information confusion and limited degradation pattern representation, by incorporating an independent variable encoding strategy and the LSTM gating mechanism, which may enhance temporal modeling capability. In addition, the use of a cross-attention mechanism facilitates the integration of complementary feature information, potentially improving the representation of complex temporal patterns. Moreover, the application of the Kolmogorov–Arnold Network (KAN) for high-dimensional feature mapping contributes to the model's nonlinear representation ability and may support improved predictive performance.
(3) Experimental results on the IEEE PHM2012 dataset suggest that the proposed model achieved competitive performance in terms of MAE, RMSE, and other evaluation metrics, indicating its potential effectiveness and generalizability in predicting the remaining useful life of rolling bearings.

Although the proposed iTransformer-LSTM dual-stream architecture has shown promising performance in RUL prediction, several important directions remain for future work. Firstly, the current study validates the model solely on the PHM2012 dataset under a single operating condition. Its generalization capability under varying loads, multiple working conditions, and real-world industrial scenarios—particularly with strong noise and limited fault samples—requires further evaluation. Future work will investigate condition-invariant learning and domain adaptation techniques (e.g., baseline compensation and reference-free methods) [27] to enhance robustness in complex, variable environments.

Secondly, although the dual-stream structure improves prediction accuracy, it introduces additional computational overhead, which may hinder deployment on resource-constrained edge devices. To address this, future work will explore model simplification strategies, such as pruning redundant layers, knowledge distillation, or replacing certain submodules (e.g., Transformer layers) with more efficient alternatives like temporal convolutions. Meanwhile, low-bit quantization and model compression techniques will also be investigated to reduce model size and complexity without significant loss in accuracy, aiming to achieve a better balance between predictive performance and computational efficiency.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11760-025-04617-3.

Acknowledgements The authors would like to express their sincere gratitude to the GitHub website for providing the IEEE PHM 2012 fault bearing experimental datasets, as well as the anonymous reviewers for their constructive suggestions and comments on this paper.

Author contribution Zhigang Chen was responsible for the overall conceptualization and supervision of the study. Mengyao Shi conducted data processing, model construction, and experimental implementation. Yanxue Wang participated in data analysis and result interpretation. Longqiao Chen contributed to the revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding This work was supported by the National Natural Science Foundation of China (Grant No. 52275079).

Data availability The experimental dataset is available in the following repository: https://github.com/wkzs111/phm-ieee-2012-data-challenge-dataset.

Declarations

Conflict of interests The authors declare no competing interests.

References

1. Zio, E.: 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 (2022)
2. Liu, X., Zhang, Z., Li, Z., Wang, J., Zhu, Y., Ma, H.: Advancements in bearing health monitoring and remaining useful life prediction: techniques, challenges, and future directions. Meas. Sci. Technol. (2025). https://doi.org/10.1088/1361-6501/adafc8
3. Si, X.S., Li, T., Zhang, J., Lei, Y.: Nonlinear degradation modeling and prognostics: a Box-Cox transformation perspective. Reliab. Eng. Syst. Saf. **217**, 108120 (2022)
4. Zhao, X., Zhu, X., Liu, J., Hu, Y., Gao, T., Zhao, L., Yao, J., Liu, Z.: Model-assisted multi-source fusion hypergraph convolutional neural networks for intelligent few-shot fault diagnosis to electro-hydrostatic actuator. Inf. Fusion **104**, 102186 (2024)
5. Ren, L., Sun, Y., Wang, H., Zhang, L.: Prediction of bearing remaining useful life with deep convolution neural network. IEEE Access **6**, 13041–13049 (2018)
6. Wang, C., Jiang, W., Yang, X., Zhang, S.: Rul prediction of rolling bearings based on a DCAE and CNN. Appl. Sci. **11**(23), 11516 (2021)
7. Zhao, R., Wang, D., Yan, R., Mao, K., Shen, F., Wang, J.: Machine health monitoring using local feature-based gated recurrent unit networks. IEEE Trans. Ind. Electron. **65**(2), 1539–1548 (2017)
8. Liu, J., Li, Q., Chen, W., Yan, Y., Qiu, Y., Cao, T.: Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks. Int. J. Hydrogen Energy **44**(11), 5470–5480 (2019)
9. Li, X., Zhang, L., Wang, Z., Dong, P.: Remaining useful life prediction for lithium-ion batteries based on a hybrid model combining the long short-term memory and Elman neural networks. J. Energy Storage **21**, 510–518 (2019)
10. Al-Dahidi, S., Rashed, M., Abu-Shams, M., Mellal, M.A., Alrbai, M., Ramadan, S., Zio, E.: A novel approach for remaining useful life prediction of high-reliability equipment based on long short-term memory and multi-head self-attention mechanism. Qual. Reliab. Eng. Int. **40**(2), 948–969 (2024)
11. Gheini, M., Ren, X., May, J.: Cross-attention is all you need: adapting pretrained transformers for machine translation. arXiv preprint arXiv:2104.08771 (2021)
12. Chen, C., Wang, T., Liu, Y., Cheng, L., Qin, J.: Spatial attention-based convolutional transformer for bearing remaining useful life prediction. Meas. Sci. Technol. **33**(11), 114001 (2022)
13. Mu, H., Zhai, X., Yin, D., & Qiao, F.: A method of remaining useful life prediction of multi-source signals aero-engine based on RF-Transformer-LSTM. In: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2502–2507). IEEE (2022)
14. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: inverted transformers are effective for time series forecasting. arXiv preprint arXiv:2310.06625 (2023)
15. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Tegmark, M.: Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756 (2024)
16. Liu, Z., Zhang, L., Carrasco, J.: Vibration analysis for large-scale wind turbine blade bearing fault detection with an empirical wavelet thresholding method. Renew. Energy **146**, 99–110 (2020)
17. Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., Varnier, C.: PRONOSTIA: an experimental platform for bearings accelerated degradation tests. In: IEEE International Conference on Prognostics and Health Management, PHM'12. (pp. 1–8). IEEE Catalog Number: CPF12PHM-CDR (2012)
18. Yuan, W., Li, X., Gu, H., Zhang, F., Miao, F.: Engine remaining useful life prediction based on PSO optimized multi-layer long short-term memory and multi-source information fusion. Meas. Control. **57**(5), 638–649 (2024)
19. Burda, E.A., Zusman, G.V., Kudryavtseva, I.S., Naumenko, A.P.: An overview of vibration analysis techniques for the fault diagnostics of rolling bearings in machinery. Shock. Vib. **2022**(1), 6136231 (2022)
20. Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., Nandi, A.K.: Applications of machine learning to machine fault diagnosis: a review and roadmap. Mech. Syst. Signal Process. **138**, 106587 (2020)
21. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623–2631) (2019)

22. Zhou, H., Zhang, S., Peng, J., et al.: Informer: beyond efficient transformer for long sequence time-series forecasting. Proc. AAAI Conf. Artif. Intell. **35**(12), 11106–11115 (2021)
23. Wang, Y., Zhao, Y.: Multi-scale remaining useful life prediction using long short-term memory. Sustainability **14**(23), 15667 (2022)
24. Sharma, N.K., Bojjagani, S.: Mechanical element's remaining useful life prediction using a hybrid approach of CNN and LSTM. Multimed. Tools Appl. **83**(31), 75927–75953 (2024)
25. Cai, X., Liu, T.: State of health prediction for Lithium-ion batteries using transformer–LSTM fusion model. Appl. Sci. **15**(7), 3747 (2025)
26. Xie, R., Liang, C., Zheng, X., Zuo, Z., Ouyang, Y., Pan, G.: Short-Term PV cluster power prediction based on fuzzy C-means and iTransformer-TCN. In: 2024 IEEE PES 16th Asia-Pacific Power and Energy Engineering Conference (APPEEC) (pp. 1–5), IEEE (2024)
27. Rezazadeh, N., De Luca, A., Perfetto, D., Salami, M.R., Lamanna, G.: Systematic critical review of structural health monitoring under environmental and operational variability: approaches for baseline compensation, adaptation, and reference-free techniques. Smart Mater. Struct. (2025). https://doi.org/10.1088/1361-665X/ade7db