This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,
IEEE INTERNET OF THINGS JOURNAL

Machinery Multimodal Uncertainty-Aware RUL Prediction: A Stochastic Modeling Framework for Uncertainty Quantification and Informed Fusion

Yuan Wang, Yaguo Lei, *Senior Member, IEEE*, Naipeng Li. *Senior Member*, *IEEE,*
Ke Feng, *Member, IEEE*, Zidong Wang, *Fellow, IEEE*, Yang Tan, Huitong Li

***Abstract*— Accurate prediction of machinery's remaining useful life (RUL) is essential for preventing catastrophic breakdowns and supporting predictive maintenance. Although RUL prediction has been extensively studied, most literature develops on unimodal data, which provides a limited and often biased perspective. Multimodal monitoring, which collects multiple sensor data, enables a more comprehensive understanding of degradation processes. While promising, significant challenges are encountered in existing methods: 1) point yet deterministic predictions are predominantly produced which, while potentially erroneous, tend to exhibit overconfidence, thereby lacking the dynamic uncertainty informing; 2) the processing of heterogeneous data and the achievement of physically interpretable fusion remain challenging; and 3) anomalies in the operation process are not appropriately identified. To address these issues, a new multimodal uncertainty-aware RUL prediction framework is proposed, grounded in stochastic modeling. Fractional stochastic differential equation-controlled subnets process each modality independently, wherein layer-wise transformations are modeled as state evolution in stochastic dynamical systems, allowing modality-specific uncertainty to be quantified without requiring parameter priors. A Lagrange multiplier-based fusion module is subsequently employed to perform explicit uncertainty-based fusion, enabling an interpretable and synergistic integration. Validation on harmonic drive reducers for robots demonstrates the superiority of the proposed framework, achieving an average improvement of 26.6% in RMSE and a 16.6% reduction in MAPE compared to state-of-the-art benchmarks. Furthermore, the method significantly reduces prediction uncertainty variance by 21.3%, offering more reliable insights into system degradation.**

***Index Terms*—Multimodal learning, Remaining useful life prediction, Uncertainty quantification, Stochastic modeling, Harmonic drive reducers**

I. Introduction

ADVANCES in sensor technology, the Internet of Things (IoT), and cyber-physical systems have ushered in the era of the Internet of Machinery [1]. In remaining useful life (RUL) prediction, major challenges are encountered by traditional methods in extracting valuable fault-related features and identifying degradation patterns from vast volumes of data. As a result, deep learning techniques have emerged as a promising solution [2, 3]. Extensive research has been conducted to explore their transformative potential in this domain. For instance, Deutsch et al. [4] presented a deep belief network (DBN)-feedforward neural network that enables automatic feature extraction and RUL prediction from vibration data of gears and bearings. Liu et al. [5] introduced a feature-attention mechanism that assigns higher weights to critical inputs. Subsequently, a bidirectional gated recurrent unit (GRU) was employed for relevant feature extraction, followed by fully connected layers (FCLs) for RUL prediction. Li et al. [6] focused on addressing the degradation alignment issue by proposing a cycle-consistent learning scheme to align sensor data at similar degradation levels. RUL predictions were then obtained using a first predicting time determination approach. Cao et.al [7] proposed a temporal cascade broad learning system for continuous learning with incoming data. A ridge regression method was used to update the network weights and support subsequent RUL predictions.

While most existing methods have demonstrated a certain level of effectiveness, a significant limitation is that they are developed based on unimodal data. Unimodal data provide only a limited and potentially biased perspective [8], making it highly likely that critical aspects of a machine's health state are overlooked and that the machinery degradation process is inadequately represented. Furthermore, unimodal data are susceptible to external interference, and their effectiveness is highly dependent on the sensitivity and characteristics of the specific sensor employed [9]. To overcome these limitations, multimodal monitoring has received significant attention and application. This approach incorporates diverse sensor types and extends monitoring coverage to collect more comprehensive and abundant information about machine health states. Fig. 1 presents an example of sensor arrays used for

This work was supported in part by the National Science Fund for Distinguished Young Scholars of China under Grant 52025056, in part by the National Natural Science Foundation of China under Grant 52435003, 52375121 and in part by the Fundamental Research Funds for the Central Universities of China. *(Corresponding author: Naipeng Li).*

Y. Wang, Y. Lei, N. Li, K. Feng, Y. Tan and H. Li are with the Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi'an Jiaotong University, Xi'an, 710049, China. (e-mail: oliveryuan@stu.xjtu.edu.cn; yaguolei@mail.xjtu.edu.cn; naipengli@mail.xjtu.edu.cn; kefeng@xjtu.edu.cn; tanyang@stu.xjtu.edu.cn; huitongli@stu.xjtu.edu.cn).

Z. Wang is with the Department of Computer Science, Brunel University, London, Uxbridge UB8 3PH, UK (e-mail: Zidong.Wang@brunel.ac.uk).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,

monitoring in an industrial robot, which is a representative mechanical system. The diverse and enriched data acquired through multimodal monitoring provide deeper insights into machine degradation patterns, thereby improving RUL prediction accuracy. However, these multimodal data present both opportunities and challenges. On the one hand, they contain abundant information about machine health states, enabling models to access a broader perspective. On the other hand, differences in the statistical properties of each modality and the complex nonlinear relationships among their low-level representations pose challenges for effective data utilization and fusion [10]. Some researchers have attempted to explore the potential of multimodal monitoring data for RUL prediction. A method for outlier removal, based on the μ-3σ principle was developed by Guo et al. [11], followed by the implementation of a multi-scale convolutional attention network to fuse the cleaned data from milling cutters for RUL prediction. Yang et al. [12] proposed a multi-branch network designed to process diverse data modalities, including sensor measurements, images, and inspection records. Features extracted from each branch were subsequently combined and fed into a regression layer to predict the RUL of steam generators. Li et al. [13] developed a multi-scale feature extraction module that dynamically weights multi-sensor measurements based on their similarity across different time scales. These weighted features were then projected onto turbofan engine RULs using a hybrid approach combining local and global attention mechanisms. Wang et al. [14] introduced a framework in which data from multiple modalities (e.g., current and sound) were processed in parallel using subnets equipped with distillation blocks. These subnets extracted features, which were then fused and input into fully connected layers for the final RUL prediction.

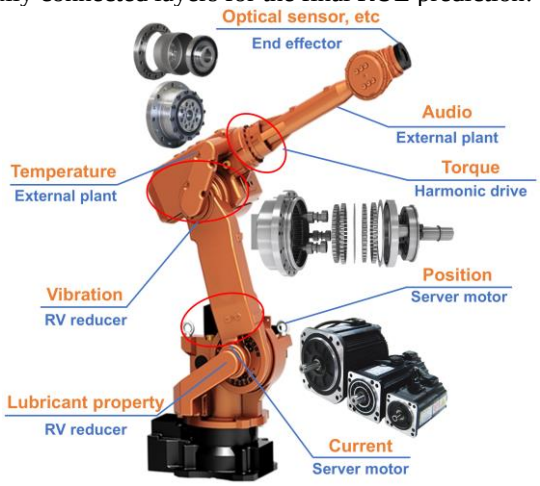


Fig. 1 A typical example of sensor array deployment on an industrial robot.

So far, research efforts have been dedicated to leveraging multimodal data and integrating them for RUL prediction. Unfortunately, the available results are limited to providing point predictions without corresponding uncertainty levels. Consequently, erroneous yet overly confident predictions may be produced, even for out-of-distribution (OOD) data. This limitation has significant implications across various industries, particularly those that require high standards of safety and control. For instance, stringent safety standards are mandated in the aerospace industry, efficient workflows and product quality are prioritized in smart manufacturing, and driver and passenger safety is emphasized in the autonomous driving sector. Existing methods for uncertainty quantification in neural networks (NNs) primarily fall into two categories [15]. The first category comprises Bayesian neural networks (BNNs), which impose probability distributions over model parameters, with related works including [16, 17]. Although BNNs offer a principled approach, the exact probabilistic inference of parameter posteriors is typically infeasible. Furthermore, the specification of parameter priors for deep neural networks (DNNs) is intractable due to the large and complex parameter space. The second category consists of non-Bayesian methods, among which model ensembling is the most widely used. In this approach, multiple DNNs are trained with different initializations to estimate uncertainty [18, 19]. However, training multiple DNNs for ensembling can be prohibitively expensive and inefficient in practical applications.

To achieve more effective uncertainty quantification in RUL prediction, a new multimodal uncertainty-aware RUL prediction framework based on stochastic model updating is proposed. Within this framework, the uncertainty of each modality is modeled using a customized fractional stochastic differential equation (FSDE) subnet. Modality-specific features are then dynamically fused according to their respective uncertainties. The final predictions, along with their associated uncertainties, are obtained through regression reasoning. Experimental results on harmonic drive reducers life-testing data show that, compared with four advanced uncertainty quantification methods, the proposed method achieves up to average 26.62% improvement in RMSE, and 20.94% reduction in uncertainty variance under different working conditions. These quantitative gains enable more reliable decision-making in predictive maintenance scenarios.

The main contributions of this paper can be summarized as follows.

1) The inherent uncertainty in RUL predictions for each modality is quantified through stochastic modeling. DNN forward transformations are modeled as the state evolution of a nonlinear stochastic dynamical system over time. Fractional Brownian motion (FBM) is specifically incorporated to capture correlated noise effects. These ensure that the framework is straightforward to implement, computationally efficient, and does not require prior specification of model parameter distributions.
2) A modality-specific FSDE subnet parallel processing strategy is developed. Each modality's data is processed separately to recognize specific degradation patterns while explicitly quantifying uncertainty. Two types of embedded uncertainty sources are identified to facilitate on-site detection and localization.
3) A Lagrange multiplier-based multimodal fusion module is introduced, providing two primary benefits. First, a more comprehensive understanding of machine health states is enabled. Second, explicit multimodal fusion is achieved, guided by modality-specific uncertainty and a well-defined optimization objective, thereby endowing the fusion process with a degree of physical interpretability.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,

II. Prediction Uncertainty Modeling

A. Sources of uncertainty

The sources of uncertainty in RUL predictions are first examined. For a commissioned system, its multimodal monitoring data can be formulated as,

$$\boldsymbol{\mathcal{D}} = \{\boldsymbol{X}_t^m \in \mathbb{R}^{c_m \times d_m}\}_{t=1}^{T} \in \mathbb{R}^{T \times M \times C \times D} \quad m = 1, \cdots, M \quad (1)$$

where T is the number of data-label pairs, M is the number of modalities, $C = \{c_m\}_{m=1}^{M}$ and $D = \{d_m\}_{m=1}^{M}$ indicate the number of channels and sampling datapoints per timestep for each modality, respectively.

To enable RUL prediction at each timestep, a model $\mathcal{M}$ is trained to establish a relationship between the monitoring data $\boldsymbol{\mathcal{D}}_t$ and the predicted RUL outcome RUL_t. This process is then scrutinized to identify the sources of uncertainty. First, uncertainty arises from the inherent variability or randomness that is characteristic of all systems. This type of uncertainty is referred to as aleatory uncertainty (AU). The second type, epistemic uncertainty (EU), results from the model's limited knowledge of the system or constraints in valuable data availability. By identifying and distinguishing these two sources of uncertainty, maintenance technicians can promptly locate the root causes of operational issues. If AU exhibits an increasing trend, it suggests that the data acquisition process should be inspected for potential issues such as sensor malfunctions or data transmission errors. Conversely, if an increase in EU is observed in prediction results, it may indicate a shift in online data distribution compared to the model's training data, suggesting the necessity of acquiring additional online data to enhance the model's knowledge.

B. Uncertainty qualification for each modality

After identifying the source of uncertainty, an attempt is made to model them. According to the investigation in [20], the forward pass in NNs can be viewed as the state evolution of a dynamical system. Therefore, this evolution can be modeled using a parameterized ordinary differential equation (ODE). Specifically, for the hidden state $\boldsymbol{h}$ at layer l, the state of the next layer can be expressed as:

$$\boldsymbol{h}_{l+1} = \boldsymbol{h}_l + b(\boldsymbol{h}_l, \boldsymbol{\theta}_b) \quad (2)$$

where $b(\cdot)$ is the mapping function of the processing branch parameterized by $\boldsymbol{\theta}_b$.

Eq. (2) illustrates a typical processing paradigm in NNs. By replacing 1 with Δl, it can be rewritten as:

$$\frac{\boldsymbol{h}_{l+1} - \boldsymbol{h}_l}{1} = b(\boldsymbol{h}_l, \boldsymbol{\theta}_b) \Longrightarrow \frac{\boldsymbol{h}_{l+\Delta l} - \boldsymbol{h}_l}{\Delta l} = b(\boldsymbol{h}_l, \boldsymbol{\theta}_b) \quad (3)$$

Here, Δl represents the step size of the state between two adjacent layers. In NNs, forward inference is performed by sequentially stacking multiple layers to process the data flow. In the limit, Δl can be regarded as an infinitesimal quantity. Consequently, the relation in Eq. (3) can be reformulated as:

$$\lim_{\Delta l \to 0} \frac{\boldsymbol{h}_{l+\Delta l} - \boldsymbol{h}_l}{\Delta l} = b(\boldsymbol{h}_l, \boldsymbol{\theta}_b) \Longrightarrow \frac{d\boldsymbol{h}_l}{dl} = b(\boldsymbol{h}_l, \boldsymbol{\theta}_b) \Leftrightarrow d\boldsymbol{h}_l = b(\boldsymbol{h}_l, \boldsymbol{\theta}_b)dl \quad (4)$$

Based on the above transformations, the computation flow in a NN can reasonably be regarded as a dynamical system controlled by an ODE. This ODE-controlled model can capture system dynamics without pre-supposing prior distributions, promoting more efficient training. However, the ODE-controlled model remains deterministic, providing only point predictions and failing to quantify EU. To enable the model to produce predictions with corresponding uncertainty, a stochastic differential equation (SDE) can be introduced to replace the deterministic formulation. Thus, Eq. (4) can be reformulated as,

$$d\boldsymbol{h}_l = \mu(\boldsymbol{h}_l, l)dl + \sigma(\boldsymbol{h}_l, l)dB_l \quad (5)$$

where $\mu(\cdot)$ represents the drift term which denotes the main trend of predictions, $\sigma(\cdot)$ is the diffusion term characterizing model uncertainty in a stochastic environment, dB_l is the differential increment of a stochastic process.

The SDE serves as a mathematical formulation that describes continuous random fluctuations in a prediction process over time which indicates that, over a small interval Δl, the change $d\boldsymbol{h}_l$ in state $\boldsymbol{h}_l$ is influenced by both drift and diffusion factors. However, solving Eq. (5) is challenging due to the absence of a closed-form solution for the random variable $\boldsymbol{h}_l$, leading to estimation issues when high-order numerical methods are employed [21]. A promising approach to addressing this issue is to reformulate the problem from an NN perspective. Specifically, a drift net, denoted as $f(\cdot)$, is introduced to parameterize the drift term $\mu(\cdot)$, while a diffusion net $g(\cdot)$, is designed to model the diffusion term $\sigma(\cdot)$.

In Eq. (5), the term B_l models stochastic fluctuations in observations. Standard Brownian motion is a common choice, especially when observations are assumed to follow independent and identically distributed (i.i.d.) patterns. However, this assumption rarely holds in real-world industrial systems, where degradation processes exhibit temporal correlations, drift coupling, and non-Markovian behavior. Consider a typical industrial scenario: as a machine operates, increasing failure severity degrades the performance of various components (e.g., reduced rotation, stiffness loss). These changes introduce escalating disturbances and inter-component coupling, resulting in growing correlations among the increments of the degradation process over time. Moreover, even when observations are assumed to follow i.i.d. distributions, inter-layer correlations can still emerge during model training. From this perspective, FBM provides a more realistic framework for modeling noise in such systems.

FBM incorporates the Hurst index, enabling a well-established stochastic calculus framework for time series with memory effects. This characteristic makes FBM especially suitable for RUL prediction tasks. As shown in Fig. 2, FBM exhibits increased regularity at higher Hurst indices.

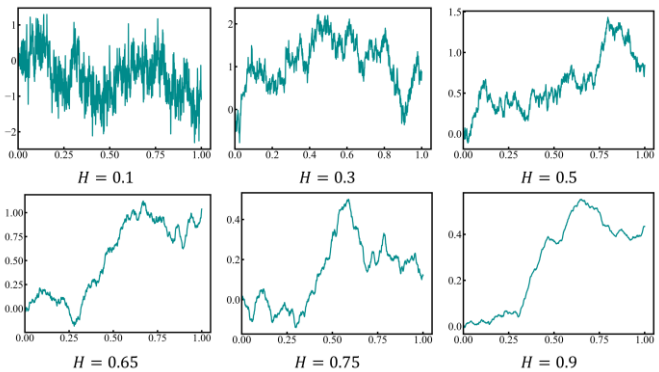


Fig. 2 Realization of FBM with different Hurst indices.

The definition of FBM is introduced by first defining the Gaussian process (GP). For a stochastic process $\{X_t : t \geq 0\}$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,
IEEE INTERNET OF THINGS JOURNAL

which consists of a collection of real-valued random variables X_t, if every its finite subset $\boldsymbol{X}_{t_1,\ldots,t_k} = (X_{t_1}, \ldots, X_{t_k})$ follows a multivariate Gaussian distribution, i.e., any linear combination of $(X_{t_1}, \ldots, X_{t_k})$ has a univariate Gaussian distribution, then the process is classified as a GP. The definition of FBM is then given as follows. A real-value Gaussian process $\mathcal{B}^H = \{\mathcal{B}_t^H : t \geq 0\}$ is defined as FBM if it satisfies the following conditions:

$$\mathcal{B}_0^H = \mathrm{E}[\mathcal{B}_t^H] = 0 \tag{6}$$

$$\mathrm{E}[\mathcal{B}_t^H \mathcal{B}_s^H] = \frac{1}{2}\{|t|^{2H} + |s|^{2H} - |t-s|^{2H}\}, s, t \geq 0 \tag{7}$$

where $H \in (0,1)$ is the Hurst index, which describes the dependence structure of FBM increments. FBM serves as a generalization of standard Brownian motion; when $H = 1/2$, it reduces to standard Brownian motion.

The fundamental properties of FBM include stationary increments, the self-affinity property, and the self-correlation property, which are defined as follows:

1) Stationary increments: $\mathcal{B}_t^H - \mathcal{B}_s^H \sim \mathcal{N}(0, |t-s|^{2H})$.

2) Self-affinity property: $\{\mathcal{B}_{t+\tau}^H - \mathcal{B}_t^H\} \triangleq \{k^{-H}[\mathcal{B}_{t+k\tau}^H - \mathcal{B}_t^H]\}$

3) Self-correlation property: For the increment of FBM: $\Delta\mathcal{B}_{t,s}^H = \mathcal{B}_t^H - \mathcal{B}_s^H$, for every $h > 0$, when $H > 1/2$ it holds that: $\sum_{n=1}^{\infty} \left|\mathrm{Cov}\left(\Delta\mathcal{B}_{0,h}^H, \Delta\mathcal{B}_{(n-1)h,nh}^H\right)\right| = \infty$.

After introducing FBM, Eq. (5) can be transformed into the FSDE form as:

$$d\boldsymbol{h}_l = f_{\boldsymbol{\theta}_1}(\boldsymbol{h}_l, l)dl + g_{\boldsymbol{\theta}_2}(\boldsymbol{h}_l, l)d\mathcal{B}_l^H \tag{8}$$

where $d\mathcal{B}_l^H$ is the differential increment of FBM, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the parameters in the drift net and diffusion net, respectively.

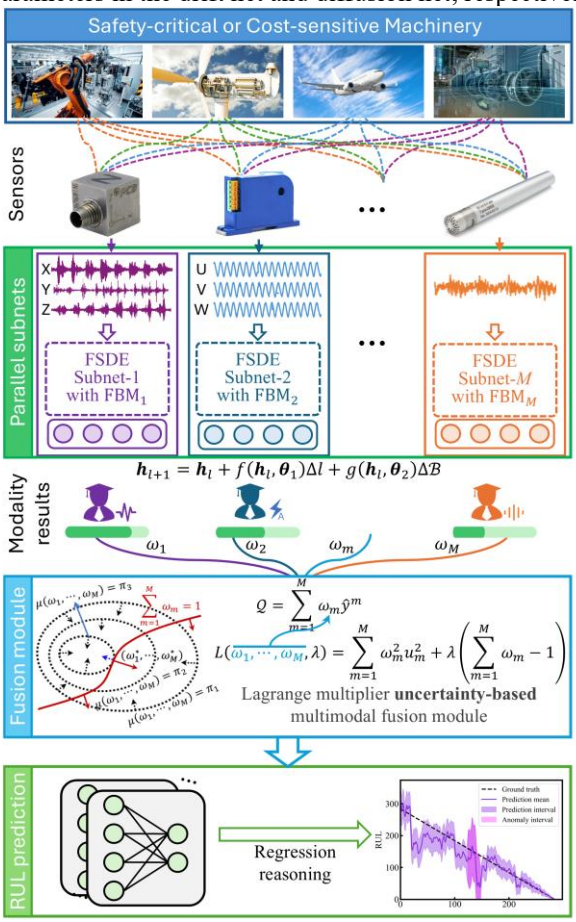


Fig. 3 Whole computation flow of the proposed framework.

The FSDE integrates a drift net $f_{\boldsymbol{\theta}_1}(\boldsymbol{h}_l, l)$ for accurate RUL prediction and a diffusion net $g_{\boldsymbol{\theta}_2}(\boldsymbol{h}_l, l)$ to induce high diffusion for data beyond the existing distribution. Specifically, for in-distribution (ID) data, where the model has sufficient observations, the variance of the stochastic process remains small, and the drift term predominantly governs the system's behavior. For OOD data, where the system possesses limited knowledge, the variance becomes significantly higher, and the diffusion term assumes a crucial role.

In summary, the derivation follows the progression: Feedforward → ODE → SDE → FSDE (with FBM), reflecting a shift in assumptions from deterministic to stochastic and from memoryless to memory-aware formulations.

III. Computation Flow of the Proposed Framework

Fig. 3 illustrates the computational workflow of the proposed framework for machinery RUL prediction and uncertainty quantification under multimodal monitoring data. Each modality is processed in parallel through a customized subnet, where each subnet is governed by its corresponding FSDE. Both ID and OOD paths are incorporated, enabling predictions with associated AU and EU. Subsequently, these modality-specific results are integrated using a weighted fusion scheme designed to minimize the overall uncertainty in the final predictions. The detailed structure of the framework is described as follows.

A. Network construction for each modality

Multimodal data are subject to different sampling principles and transmission paths during signal acquisition. Therefore, a separate processing subnet is constructed for each modality to facilitate specific feature extraction. Fig. 4 illustrates the architecture of an individual subnet. The monitoring data flow is initially processed using a sliding sampling algorithm to establish the sequence length dimension. Subsequently, the data follow distinct paths in training and prediction modes. In training mode, the data are divided into two paths: ID samples and OOD samples. ID samples are primarily utilized to update the weights of the drift net, along with a portion of the diffusion net. Meanwhile, OOD samples contribute to updating the diffusion net weights, guiding the model to produce high uncertainty. The main structures of the drift net and diffusion net are detailed below.

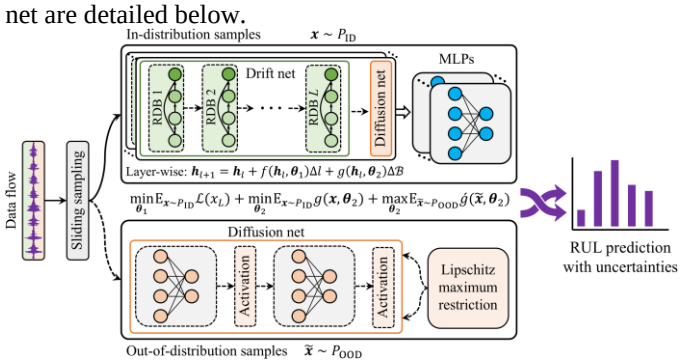


Fig. 4 The processing flow and structure of each modality-specific FSDE subnet.

I. Structure of the drift net

The drift net, serving as the main pipeline in the subnet, is designed to extract the health states of the monitored object. Its core unit, the representation distilling block (RDB), is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,
IEEE INTERNET OF THINGS JOURNAL

specifically structured to enhance feature extraction. As depicted in Fig. 5, the RDB consists of four layers: data structuring (DS), short-term localized feature refinement, long-term temporal trend exploitation, and downsampling. A residual connection is incorporated to capture cross-layer information and facilitate gradient flow. Specifically, a data sequence $\mathbf{x}$ is first structured by the DS layer to standardize to a comparable scale. Mathematically,

$$\mathrm{DS}(\mathbf{x}) = \gamma_{\mathrm{scale}} \odot \frac{\mathbf{x} - \widehat{\boldsymbol{\mu}}_{\delta}}{\widehat{\boldsymbol{\sigma}}_{\delta}} + \gamma_{\mathrm{shift}} \tag{9}$$

where $\odot$ denotes the Hadamard product operator, $\widehat{\boldsymbol{\mu}}_{\delta}$ and $\widehat{\boldsymbol{\sigma}}_{\delta}$ are the mean and standard deviation of mini-batch δ samples, respectively, $\gamma_{\mathrm{scale}}, \gamma_{\mathrm{shift}}$ are the learnable scale parameter and shift parameter, respectively.

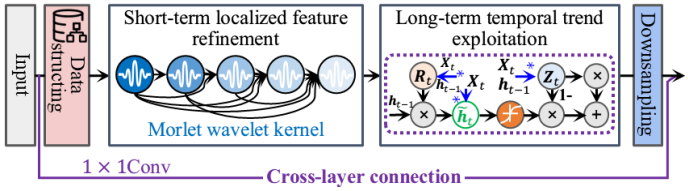


Fig. 5 The architecture of RDB in the drift net.

Subsequently, the short-term localized feature refinement layer enables the model to extract fine-grained, discriminative features. Its structure is introduced mathematically as follows. Recall that ResNet decomposes a function as:

$$f(\mathbf{x}) = \mathbf{x} + \kappa(\mathbf{x}) \tag{10}$$

This decomposition results in an additive combination of a linear term and a nonlinear term, leading to coarse feature extraction. To enhance information capture, additional terms beyond simple addition can be incorporated. For the inference function $f(x)$ at $x = x_0$, applying Taylor expansion yields:

$$f(x) = f(x_0) + f'(x_0)x + \frac{f''(x_0)}{2!}x^2 + \cdots \tag{11}$$

Here, $f(x)$ is decomposed into higher-order terms. Thus, $\mathbf{x}$ can be mapped to its transformed values through a sequence of increasingly complex functions, expressed as:

$$\mathbf{x} \to [\mathbf{x}; f_1(\mathbf{x}); f_2([\mathbf{x}, f_1(\mathbf{x})]); \dots] \tag{12}$$

These functions are realized using wavelet convolution. Mathematically, the calculation for the j-th channel of the output c_{out} for a mini-batch input δ is formulated as:

$$s\left(c_{\mathrm{out}_j}\right) = (\mathbf{x} *_d \boldsymbol{W})\left(c_{\mathrm{out}_j}\right) = \sum_{k=1}^{c_{\mathrm{in}}} \mathbf{x}(k) *_d \boldsymbol{W}\left(c_{\mathrm{out}_j}, k\right)$$
$$\text{s.t.} \quad \boldsymbol{W} = \psi_{b,a}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{13}$$

where $\boldsymbol{W}$ represents the wavelet kernel weights, $*_d$ denotes dilation with coefficient d, $\psi(\cdot)$ is the wavelet basis function, and c_{in} and c_{out} are the input and output channels, respectively.

Wavelet kernel convolution is a promising approach to addressing the non-stationarity inherent in monitoring data and enabling multi-domain analysis, potentially enhancing model performance. Additionally, it facilitates effective multi-frequency response and scales efficiently with the receptive field size without the risk of over-parameterization. In this model, the Morlet wavelet is employed due to its efficiency and simplicity.

The long-term temporal trend exploitation layer is designed to track the evolution of monitoring data over time, particularly changes in degradation patterns. Tracking this evolution is essential for identifying trends and patterns may be obscured in static snapshots. Specifically, the computation flow is controlled in a gated manner. For the input $\mathbf{x}_t$ at timestep t, the output is given by:

$$\boldsymbol{y}_t = \boldsymbol{y}_{t-1} \odot \boldsymbol{z}_t + \widetilde{\boldsymbol{\phi}}_t \odot (1 - \boldsymbol{z}_t) \tag{14}$$

where $\boldsymbol{y}_{t-1}$ represents the output from the previous timestep, $\boldsymbol{z}_t$ is described later, and $\widetilde{\boldsymbol{\phi}}_t$ is the stored state that integrates multi-resolution and attention mechanisms to effectively capture long-range contextual cues. This is formulated as:

$$\widetilde{\boldsymbol{\phi}}_t = \mathrm{HAP}(\boldsymbol{\varphi}) + \boldsymbol{\varphi}$$
$$\text{s.t.} \quad \boldsymbol{\varphi} = \mathrm{Conv}_{1\times1}([\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}; \boldsymbol{x}_t] + pe) \tag{15}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ denotes the 1×1 convolution, pe is the positional encoding term, and $\mathrm{HAP}(\cdot)$ denotes the hierarchical attention pooling (HAP) operation. This operation expands the receptive field, allowing the model to integrate information from a broader contextual window. First, the input feature maps are divided into b branches across hierarchical scales. Each branch is processed using self-attention and subsequently aggregated via a linear transformation. Mathematically,

$$\mathrm{HAP}(\boldsymbol{\varphi}) = [\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_b]\boldsymbol{W}_{\boldsymbol{\omega}} \tag{16}$$

where $\boldsymbol{W}_{\boldsymbol{\omega}} \in \mathbb{R}^{d_v \times d_v}$ are the learnable parameters in the linear transformation, and $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_b$ are individually calculated as $\boldsymbol{\omega} = \mathrm{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})\boldsymbol{V}$, where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ are the queries, keys, and values, respectively.

$\boldsymbol{z}_t$ and $\boldsymbol{r}_t$ function as gates restricted within (0, 1) to perform convex combinations, which are computed as:

$$\boldsymbol{r}_t = \varsigma(\mathrm{Conv}_{3\times3}^{s=2}[\boldsymbol{h}_{t-1}; \mathbf{x}_t], \boldsymbol{W}_r) \tag{17}$$
$$\boldsymbol{z}_t = \varsigma(\mathrm{Conv}_{3\times3}^{s=2}[\boldsymbol{h}_{t-1}; \mathbf{x}_t], \boldsymbol{W}_z) \tag{18}$$

where $\varsigma(\cdot)$ is the sigmoid activation function, $\mathrm{Conv}_{3\times3}^{s=2}(\cdot)$ denotes a convolution operation with stride $s = 2$ and kernel size 3×3, and $\boldsymbol{W}$ corresponds to the respective weight matrix.

The RDB concludes with a downsampling layer that aggregates high-level representations and reduces computational overhead without compromising signal fidelity. Downsampling is applied over divided regions, R, as:

$$\boldsymbol{o}_{i,j} = \max_{(i,j)\in R} \boldsymbol{y}_{i,j} \tag{19}$$

where $R \subset \{1,2,\dots,\Upsilon_{\mathrm{in}}\}^2$ for each $(i,j) \in \{1,\dots,\Upsilon_{\mathrm{out}}\}^2$, and the operation is performed using a specialized stochastic step that depends on the target output size.

Multiple RDBs constitute the core module of the drift net, progressively refining the health state information embedded within the monitoring data. Following this, an MLP-based module is employed to smooth the extracted representation and project it onto a lower-dimensional space.

II. Structure of the diffusion net

Another essential component in constructing the modality-specific subnet is the diffusion net. Built upon an MLP backbone, it is constrained by the Lipschitz maximum restriction [22]. Specifically, the diffusion net adheres to Lipschitz continuity, which necessitates the use of Lipschitz-compliant nonlinear activation functions, for which ReLU is selected. Additionally, to prevent excessively large values for OOD samples that could introduce instability during optimization, the maximum output of the diffusion net is regulated using a sigmoid function and the hyperparameter $\varepsilon_{\max}$. Consequently, the diffusion net is represented as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,

$\acute{g}(\mathbf{x}_B, \boldsymbol{\theta}_2)$, where $g\acute{(}\cdot) = \varepsilon_{\max} \times \varsigma(g(\cdot))$. $\varepsilon_{\max}$ is initially set to 0.1 and gradually increased to 0.5 at epoch 30.

B. Training paradigm for each subnet

After detailing the structures of the drift net and the diffusion net, the layer-wise principle for constructing subnets for ID samples is derived. By applying Euler discretization to Eq. (8), we can obtain:

$$\boldsymbol{h}_{l+1} = \boldsymbol{h}_l + f(\boldsymbol{h}_l, \boldsymbol{\theta}_1)\Delta l + g(\boldsymbol{h}_l, \boldsymbol{\theta}_2)\Delta\mathcal{B} \tag{20}$$

where $\Delta l = L/N$ represents the step size, N is the number of stochastic forward passes, and $\Delta\mathcal{B} = \mathcal{B}_{l+1}^H - \mathcal{B}_l^H$ is the increment of each step, sampled from the FBM.

To streamline the training of each FSDE subnet, the objective function is formulated with three components, each serving a distinct purpose. The first term penalizes RUL prediction error using samples within the known distribution (ID). It guides the drift net to regress toward accurate predictions. The second term regularizes the diffusion net to minimize uncertainty estimates for well-known samples (i.e., suppress overestimation on ID data). The last term encourages high uncertainty for OOD inputs, enabling the model to recognize unfamiliar data and mitigate overconfident predictions. The objective function is formulated as follows:

$$\begin{aligned}\mathcal{O}_m = &\min_{\boldsymbol{\theta}_1} \mathrm{E}_{\boldsymbol{x}\sim P_{\mathrm{ID}}} \mathcal{L}(x_L) + \min_{\boldsymbol{\theta}_2} \mathrm{E}_{\boldsymbol{x}\sim P_{\mathrm{ID}}} g(\boldsymbol{x}, \boldsymbol{\theta}_2) + \\ &\max_{\boldsymbol{\theta}_2} \mathrm{E}_{\tilde{\boldsymbol{x}}\sim P_{\mathrm{OOD}}} \acute{g}(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}_2)\end{aligned} \tag{21}$$

where L denotes the total layers in the subnet, P_{ID} and P_{OOD} represent the distributions of ID and OOD samples, respectively, and $\mathcal{L}(\cdot)$ refers to the RUL loss function [14].

Together, these terms enforce an adaptive uncertainty profile, improving prediction confidence when appropriate while maintaining caution under data mismatch.

C. Uncertainty quantification in each subnet

In the constructed framework, monitoring data from each modality are processed using a customized subnet. Each subnet is responsible for performing RUL prediction and quantitative uncertainty modeling. Specifically, for the m-th trained subnet, its prediction $\hat{\rho}^m$ is obtained as follows:

$$\hat{\rho}^m = \frac{1}{S}\sum_{s=1}^{S} \hat{\rho_s}^m \tag{22}$$

where S represents the number of stochastic forward passes during testing, and $\hat{\rho_s}^m$ denotes each prediction result.

For uncertainty quantification, as previously discussed, two types of uncertainty are considered: AU and EU, where AU is computed as:

$$u_{au}^m = \frac{1}{S}\sum_{s=1}^{S} (\hat{\rho_s}^m - \hat{\rho}^m)^2 \tag{23}$$

The second type, EU, is determined as the variance of the final solution and is given by:

$$u_{eu}^m = \mathrm{Var}(\hat{\rho_s}^m) \tag{24}$$

D. Multimodal fusion based on respective uncertainty

Multimodal fusion, a core component of multimodal learning research, integrates information from multiple modalities into a stable, unified representation. In practice, factors such as data quality, transmission paths, and sensor capacity cause variations in the amount of degradation-related information present in monitoring data across different modalities. This degradation information typically corresponds to the uncertainty associated with each modality. Therefore, a dynamic uncertainty-based weighted fusion approach is more suitable for module design. In this approach, the fusion process dynamically allocates weights based on the uncertainty of each modality, assigning lower weights (i.e., lower importance) to modality-specific features with higher uncertainty. To achieve this, a Lagrange multiplier-based fusion method is developed. This fusion approach is explicitly derived through numerical optimization, providing interpretability to the fusion process.

Notably, within each subnet, an MLP-based regression module is employed to reason predictions and their corresponding uncertainty. However, during multimodal fusion, the original modality-specific features are directly fused. Specifically, for m-th modality, the corresponding subnet provides a prediction $\hat{\rho}^m$ along with an associated uncertainty value u^m, which is computed as:

$$u^m = \alpha \cdot u_{au}^m + \beta \cdot u_{eu}^m \tag{25}$$

where α and β are scaling coefficients that can be adjusted based on the application to reflect relative importance. The unified representation is then obtained as:

$$\mathcal{Q} = \sum_{m=1}^{M} \omega_m \hat{y}^m \tag{26}$$

where ω_m is the weight for m-th modality, satisfying $\sum_{m=1}^{M} \omega_m = 1$.

According to the computational law of uncertainty propagation [15], when the measurements from each sensor are independent, the uncertainty of the final prediction u_Δ is given by:

$$u_\Delta = \sqrt{\omega_1^2 u_1^2 + \omega_2^2 u_2^2 + \cdots + \omega_M^2 u_M^2} \tag{27}$$

where $u_1, u_2, \cdots, u_M$ are the uncertainty for each modality.

To simplify computation, the goal of minimizing u_Δ can be equally substituted for optimizing ${u_\Delta}^2$. Therefore, the following function is formulated:

$$L(\omega_1, \cdots, \omega_M, \lambda) = \sum_{m=1}^{M} \omega_m^2 u_m^2 + \lambda\left(\sum_{m=1}^{M} \omega_m - 1\right) \tag{28}$$

where λ is the Lagrange multiplier.

The optimized weights can now be derived by taking the partial derivatives with respect to $\omega_1, \omega_2, \ldots, \omega_M$ and setting them to zero. Thus, the final weight for each modality is:

$$\omega_m = \frac{1}{u_m^2 \sum_{m=1}^{M} \frac{1}{u_m^2}} \tag{29}$$

IV. Experimental Validation

To evaluate the effectiveness and superiority of the proposed framework for RUL prediction and uncertainty quantification, experimental data from accelerated life-testing of harmonic drive reducers (HDRs) for robots are utilized. This section first provides data descriptions, followed by data preprocessing steps and model configurations. Finally, we present validation results, comparisons, and relevant analyses.

A. Data descriptions

HDRs utilize the controllable deformation of a flexible element to transmit motion and power, offering advantages such as high precision and high load capacity. To investigate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,
IEEE INTERNET OF THINGS JOURNAL

their degradation process, a test bench was designed following GB/T 30819-2014 and GB/T 40729-2021 standards with two Panasonic servo motors. The tested harmonic drive reducers are of type WH-CS-32-80-I. Fig. 6 illustrates the experimental setup and sensor arrangement, while Fig. 7 presents a schematic diagram of the transmission chain. In the experimental setup, a servo motor drives the HDR, which is loaded by another servo motor with torque amplified through an RV reducer. During the experiment, five types of monitoring data were collected. Internal current data were recorded via data tracking using the OMRON NJ301-1100 PLC. External signals included triaxial vibration acceleration data (PCB 356A15), torque data (LONGLV WTQ1050B), three-phase current data (WBI411N95), and two-channel acoustic emission (AE) data (PAC WD). The experimental parameters and descriptions are summarized in TABLE I.

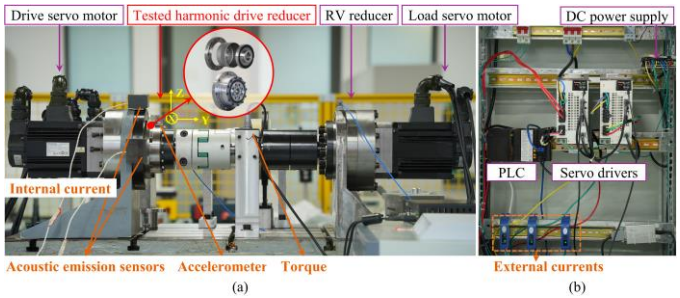


Fig. 6 Experimental test bench and sensor installation layout. (a) Mechanical connection section. (b) Power control section.

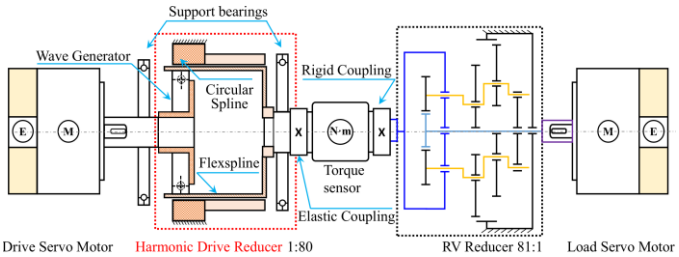


Fig. 7 Schematic diagram of the transmission chain for the test bench.

TABLE I
DESCRIPTIONS OF ACCELERATED LIFE-TESTING OF HARMONIC DRIVE REDUCERS.

Condition	Speed - Load	RUL interval (recordings)	Train-Test dataset	
Cond-1	2000rpm-193.4Nm	[39.5, 262]	HDR 1-1	HDR 1-2
Cond-2	2000rpm-154.7Nm	[21.22, 74]	HDR 2-1 to HDR 2-3	HDR 2-4
Cond-3	1500rpm-154.7Nm	[21.08, 122]	HDR 3-1, HDR 3-2	HDR 3-3

During the monitoring data collection process, external data were sampled at a frequency of 25.6 kHz for a duration of 2.56 seconds, with an interval of 1 minute between successive recordings. Internal and AE data were continuously sampled: internal data at a frequency of 2 kHz for a duration of 1 second, and AE data at a frequency of 1 MHz for a duration of 0.1 second. The life-cycle plots for each modality in one sample are shown in Fig. 8. As shown in Fig. 8, each modality has a different number of channels, distinct datapoints, varied trend evolution forms, and considerable heterogeneity. Integrating and exploiting degradation-related information from these diverse modalities in a realistic manner presents a significant challenge.

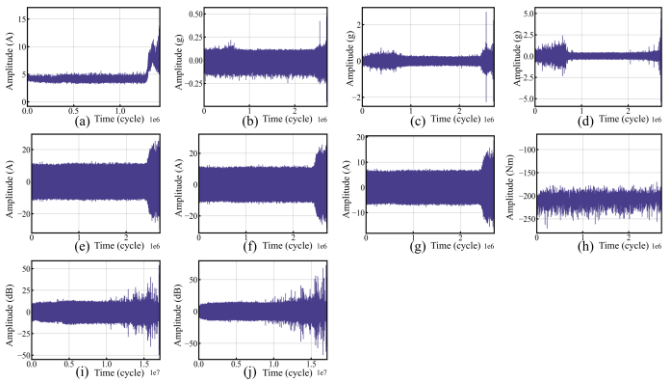


Fig. 8 The whole life-cycle multimodal monitoring data of HDR 2-1. (a) Internal current. (b) Accelerometer in X-axis. (c) Accelerometer in Y-axis. (d) Accelerometer in Z-axis. (e) External current in U-phase. (f) External current in V-phase. (g) External current in W-phase. (h) Torque. (h) AE in X-axis. (i) AE in Z-axis.

B. Data preprocessing and model configurations

In the data processing pipeline of the proposed method, all modalities underwent min-max normalization to scale each channel to the range [0, 1]. Labels were also normalized using the health-degree normalization method to reflect relative degradation. Training samples were generated using a sliding window algorithm with a window size of 5 and a stride of 1. The dataset partitioning is summarized in TABLE I and was kept consistent across all models to ensure fair comparison. During training, popular techniques in the community such as early stopping, weight decay, and gradient clipping are implemented to prevent overfitting and enhance training efficiency. The Adam optimizer is employed with a momentum of 0.9, weight decay of 5×10^{-4}, and mini-batch size of 64. Training is conducted for 120 epoch. The initial learning rate for the drift net is set to $1e^{-4}$ and is reduced after epoch 50, while the learning rate for the diffusion net is set to $1e^{-3}$. The scaling coefficients α and β in Eq. (25) are set to 0.6 and 0.4. In subsequent experiments, common neural network training strategies will be applied to all networks to ensure fair comparisons. Each model is trained 20 times to mitigate the effects of random errors. Additionally, prediction performance is evaluated using two standard metrics, namely, root mean square error (RMSE) and mean absolute percentage error (MAPE), where lower values indicate better performance.

To implement FBM in our framework, we utilize the Davies-Harte method, which is theoretically exact for generating discretely sampled FBM. This method efficiently generates FBM samples by employing fast Fourier transforms to create a circulant embedding of the covariance matrix. During network training and inference, each FSDE subnet uses these pre-generated FBM increments to model noise with long-range dependencies. Specifically, the increment $\Delta\mathcal{B} = \mathcal{B}_{l+1}^{H} - \mathcal{B}_{l}^{H}$ is sampled and scaled by the diffusion net output at each layer transition. This approach maintains computational efficiency while accurately representing the correlated noise structure characteristic of industrial machinery degradation processes.

C. Experimental results of the proposed framework

Following the above settings, the proposed network was applied to the five-modality data. Fig. 9 presents the results for each working condition. In Fig. 9, firstly, the RUL predictions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,
IEEE INTERNET OF THINGS JOURNAL

and corresponding uncertainty for each timestep are provided. The following observations can be made. 1) The proposed network demonstrates satisfactory regression performance, achieving effective RUL prediction with uncertainty quantification across all working conditions. 2) During the prediction of each sample, as the model accumulates more information regarding health states and degradation cues, its prediction performance improves over time. In the critical final stage, the model's estimates closely align with actual RUL curves. 3) When comparing predictions across different working conditions, predictions under Cond-2 exhibit higher accuracy and lower uncertainty than those under Cond-1, indicating better generalization in data-rich conditions. Furthermore, comparing predictions under Cond-2 and Cond-3 reveals that a longer prediction span presents a greater challenge to the model's ability to maintain continuous prediction accuracy, even when the number of available samples is sufficient.

Notably, the detected anomaly intervals are highlighted in Fig. 9, identified by a sudden increase in uncertainty and values exceeding twice those of other intervals. The causes of these anomalies were also traced. Specifically, under Cond-1, the anomaly is attributed to an unusually strong disturbance; under Cond-2, it results from data drift at the anomaly point; and under Cond-3, it is caused by missing data.

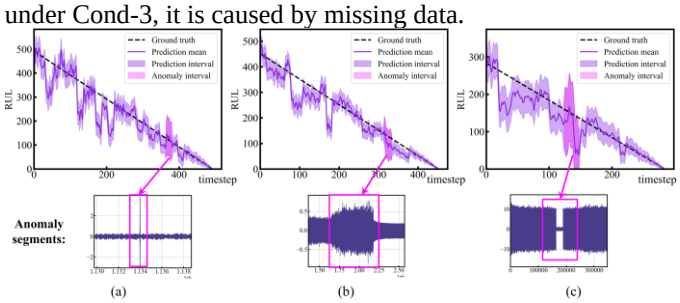
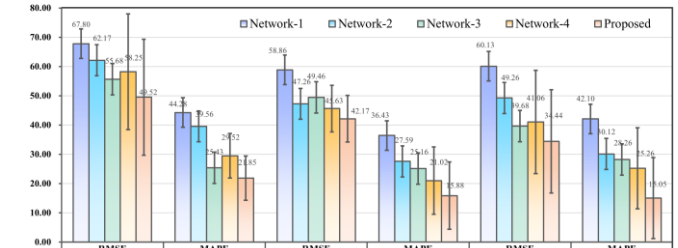


Fig. 9 Prediction results with corresponding prediction intervals, highlighting detected anomaly intervals and their causes. (a) Cond-1. (b) Cond-2. (c) Cond-3.



	Cond-1 RMSE	Cond-1 MAPE	Cond-2 RMSE	Cond-2 MAPE	Cond-3 RMSE	Cond-3 MAPE
Network-1	67.80±29.08	44.28±27.26	58.86±29.86	36.43±26.16	60.13±31.68	42.10±26.43
Network-2	62.17±21.64	39.56±18.85	47.26±15.26	27.59±20.56	49.26±26.14	30.12±21.26
Network-3	55.68±22.16	25.43±12.54	49.46±12.36	25.16±19.26	39.68±24.36	28.26±19.87
Network-4	58.25±21.26	29.52±9.26	45.63±9.56	21.02±15.59	41.06±21.46	25.26±15.68
Proposed	49.52±19.80	21.85±7.57	42.17±8.00	15.88±11.48	34.44±17.66	15.05±13.84

Fig. 10 Comparison of prediction results among the four network variants and the proposed network, each with a different base unit in the drift net.

D. Ablation studies

I. Effectiveness of RDB in modality-specific subnet

To validate the effectiveness of the key component in the drift net, an ablation experiment is conducted. The basic unit RDB in the drift net was replaced while keeping the rest of the architecture unchanged. Four network variants were designed for this purpose: Convolutional bi-LSTM [23] in Network-1, MSAN [11] in Network-2, MSer [24] in Network-3, and MCTAN [25] in Network-4. The prediction comparison results are presented in Fig. 10.

From Fig. 10, it can be observed that the choice of the basic unit in the drift net has a more significant impact on model predictions than on uncertainty quantification. This is primarily because the drift net's main function is to capture the overall trend of model predictions. Specifically, the comparison between Network-2 and Network-1 indicates that incorporating a multi-scale design and an attention mechanism allows the model to extract more localized information about fault signatures, thereby improving the accuracy of Network-2 predictions. The comparison between Network-2 and Network-3 further demonstrates that the inclusion of a multi-head attention mechanism enhances model performance. Additionally, the comparison between Network-2 and Network-4 highlights the importance of temporal information mining for RUL predictions. Temporal information, which tracks data evolution over time, is crucial for improving the model's ability to differentiate similar objects in complex monitoring scenarios. It also helps to reveal trends and patterns that may not be apparent in static snapshots. Based on the overall comparison results, the proposed network achieves the best prediction performance in terms of both accuracy and uncertainty, demonstrating the effectiveness of the constructed RDB unit.

II. Advantages of the multimodal fusion module

To evaluate the advantages of the Lagrange multiplier-based uncertainty-driven multimodal fusion module, four variant networks were developed. Network-A utilizes element-wise addition for fusion, while Network-B incorporates the temporal compact bilinear fusion module [26]. Network-C applies multimodal factorized bilinear pooling [27]. The prediction results are presented in Fig. 11.

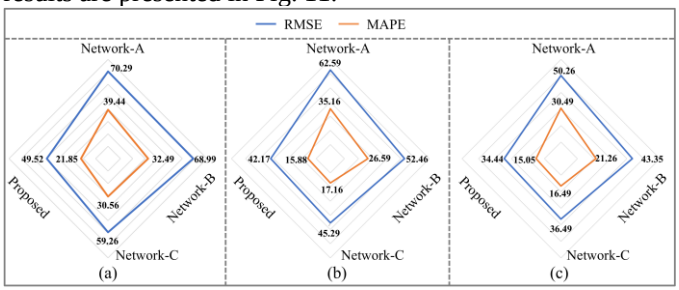


Fig. 11 Results comparison between four networks with different fusion approaches.

As illustrated in Fig. 11, Network-A follows a linear fusion strategy. Due to substantial distributional discrepancies among modalities, linear models struggle to achieve effective fusion, leading to inferior results. In contrast, Networks-B and -C employ bilinear pooling with computational optimizations, surpassing linear methods in performance. However, Network-B exhibits instability, as it relies on high-dimensional feature representations for quadratic expansion to ensure optimal performance. Network-C mitigates this issue by utilizing pair-wise factorized bilinear pooling, transforming bilinear operations into Hadamard products computed from low-rank weight matrices. Nevertheless, these two networks are highly sensitive to data richness. They perform well under data-abundant conditions (Cond-2) but struggle to maintain reliability under data-scarce conditions (Cond-1). The proposed framework, equipped with the uncertainty-based fusion module,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,

outperforms these variant networks. It provides a clear optimization objective guided by modality-specific uncertainty, enabling the network to emphasize informative modalities while suppressing irrelevant ones.

E. Comparison to several state-of-the-art uncertainty quantification prediction methods

To further verify the key characteristic of the proposed network, uncertainty quantification, several state-of-the-art (SOTA) networks commonly used in the community were constructed for comparison. The first network is a deterministic model that utilizes a vision Transformer [28], the second network is a deep ensemble model [29], the third network is a BNN based [30], and the fourth network employs Monte Carlo (MC) dropout [31]. For the deterministic network, predictions and uncertainty estimates are obtained by averaging predictions and computing the standard deviation across multiple random initializations. The prediction results are detailed in TABLE II.

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH FOUR STATE-OF-THE-ART METHODS

Condition	Metric	Deterministic	Deep Ensemble	BNN	MC-dropout	**Proposed**
Cond-1	RMSE	50.06± 36.13	53.26± 31.94	89.64± 10.26	59.64± 23.56	**49.52± 19.80**
	MAPE	25.16± 25.89	24.63± 20.68	41.46± 6.46	26.16± 10.46	**21.85± 7.57**
Cond-2	RMSE	44.16± 28.49	48.26± 26.65	69.26± 8.26	46.59± 9.91	**42.17± 8.00**
	MAPE	17.16± 17.22	18.26± 16.10	39.26± 11.21	20.61± 13.26	**15.88± 11.48**
Cond-3	RMSE	37.02± 27.99	37.58± 26.79	52.16± 18.16	39.95± 20.69	**34.44± 17.66**
	MAPE	20.46± 16.89	21.09± 16.58	29.99± 12.59	19.62± 15.20	**15.05± 13.84**

From TABLE II, the following observations can be made. 1) the deterministic-type network benefits from the powerful capacity of the vision Transformer structure, achieving moderately accurate predictions. However, its predictions exhibit high volatility, making it difficult to ensure on-site accuracy. Without explicit uncertainty quantification, the prediction variance fails to directly reflect model uncertainty. 2) The deep ensemble network improves average uncertainty quantification; however, high uncertainty persists, undermining model reliability. 3) In comparison, the BNN exhibits the lowest uncertainty, indicating high confidence in its predictions. However, its prediction accuracy is not as satisfactory as expected, as the model produces many erroneous but overconfident results. 4) The MC-dropout network delivers reasonably acceptable accuracy and uncertainty. However, the model requires extensive training and struggles to converge in later prediction stages. Overall, the proposed network leverages explicit uncertainty modeling and a well-structured framework to achieve consistently superior performance, reducing RMSE by average 52.22% compared to BNN and 15.69% compared to deterministic network, while simultaneously providing reliable uncertainty bounds that are 16.67% narrower than those from the MC-dropout method.

V. CONCLUSION AND PERSPECTIVE

This paper has presented a new multimodal uncertainty-aware RUL prediction framework, developed through stochastic model updating. Within this framework, data from each modality has been processed in parallel using customized FSDE subnets. Each FSDE subnet has regarded DNN transformations as state evolutions within a stochastic dynamical system, extending model updating from the deterministic to the stochastic domain. Different sources of uncertainty have been identified as AU and EU, with an FBM term introduced to capture EU. Subsequently, modality-specific features have been fused using a Lagrange multiplier-based multimodal fusion module based on their uncertainties. Final predictions, along with corresponding uncertainty estimates, have been generated through regression reasoning. Demonstrations have been conducted using data from accelerated life-testing on HDRs for robots. The results have confirmed the framework's effectiveness and superiority in RUL prediction, demonstrating its capability for uncertainty quantification in both each modality and the final predictions.

While the proposed FSDE-based framework demonstrates strong performance, several limitations exist. First, the use of modality-specific subnets increases model complexity and may challenge real-time deployment in resource-constrained environments. Second, the current fusion strategy requires all modalities to be present during inference, limiting adaptability. Third, FBM sampling introduces additional computational overhead. In future work, we aim to explore lightweight adaptive fusion modules that can dynamically enable/disable modalities, and to investigate online FSDE learning for continuous adaptation in evolving environments.

REFERENCES

[1] Y. Yuan *et al.*, "Data driven discovery of cyber physical systems," *Nature communications*, vol. 10, no. 1, p. 4894, 2019.
[2] G. Zhai, Y. Xu, and B. F. Spencer, "Bidirectional graphics-based digital twin framework for quantifying seismic damage of structures using deep learning networks," *Structural Health Monitoring*, vol. 24, no. 1, pp. 86-110, 2025.
[3] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11-26, 2017.
[4] J. Deutsch and D. He, "Using Deep Learning-Based Approach to Predict Remaining Useful Life of Rotating Components," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 1, pp. 11-20, 2018, doi: 10.1109/tsmc.2017.2697842.
[5] H. Liu, Z. Liu, W. Jia, and X. Lin, "Remaining useful life prediction using a novel feature-attention-based end-to-end approach," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 1197-1207, 2020.
[6] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Degradation alignment in remaining useful life prediction using deep cycle-consistent learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5480-5491, 2021.
[7] Y. Cao, M. Jia, P. Ding, X. Zhao, and Y. Ding, "Incremental learning for remaining useful life prediction via temporal cascade broad learning system with newly acquired data," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 4, pp. 6234-6245, 2022.
[8] Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, and M. Zitnik, "Multimodal learning with graphs," *Nature Machine Intelligence*, vol. 5, no. 4, pp. 340-350, 2023.
[9] Y. Wang, Y. Lei, N. Li, X. Li, and B. Yang, "Multimodal Correlation-Aware Fusion Framework for Enhanced Machinery Health Prognosis With Unlabeled and Low-Quality Data Exploitation," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
[10] G. Zhai *et al.*, "Coupled data/physics-driven framework for accurate and efficient structural response simulation," *Engineering Structures*, vol. 327, p. 119636, 2025.
[11] L. Guo, Y. Yu, H. Gao, T. Feng, and Y. Liu, "Online Remaining Useful Life Prediction of Milling Cutters Based on Multisource Data and Feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,

Learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5199-5208, 2021.
[12] Z. Yang, P. Baraldi, and E. Zio, "A multi-branch deep neural network model for failure prognostics based on multimodal data," *Journal of Manufacturing Systems*, vol. 59, pp. 42-50, 2021.
[13] R. Li *et al.*, "Multiscale Feature Extension Enhanced Deep Global-Local Attention Network for Remaining Useful Life Prediction," *IEEE Sensors Journal*, 2023.
[14] Y. Wang *et al.*, "A Multimodal Dynamic Parameterized Bilinear Factorized Framework for Remaining Useful Life Prediction under Variational Data," *Reliability Engineering & System Safety*, p. 110025, 2024.
[15] S. Bi, M. Beer, S. Cogan, and J. Mottershead, "Stochastic model updating with uncertainty quantification: an overview and tutorial," *Mechanical Systems and Signal Processing*, vol. 204, p. 110784, 2023.
[16] J. Yao, W. Pan, S. Ghosh, and F. Doshi-Velez, "Quality of uncertainty quantification for Bayesian neural network inference," *arXiv preprint arXiv:1906.09686*, 2019.
[17] L. Yang, X. Meng, and G. E. Karniadakis, "B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data," *Journal of Computational Physics*, vol. 425, p. 109913, 2021.
[18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.
[19] Y. Liu, G. Zhao, and X. Peng, "Deep learning prognostics for lithium-ion battery based on ensembled long short-term memory networks," *IEEE Access*, vol. 7, pp. 155130-155142, 2019.
[20] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.
[21] L. Kong, J. Sun, and C. Zhang, "Sde-net: Equipping deep neural networks with uncertainty estimates," *arXiv preprint arXiv:2008.10546*, 2020.
[22] Q. Li, L. Chen, C. Tai, and E. Weinan, "Maximum principle based algorithms for deep learning," *Journal of Machine Learning Research*, vol. 18, no. 165, pp. 1-29, 2018.
[23] J. Liu, L. Xu, and N. Chen, "A spatiotemporal deep learning model ST-LSTM-SA for hourly rainfall forecasting using radar echo images," *Journal of Hydrology*, vol. 609, p. 127748, 2022.
[24] H.-j. Zhu, W. Gu, L.-m. Wang, Z.-c. Xu, and V. S. Sheng, "Android malware detection based on multi-head squeeze-and-excitation residual network," *Expert Systems with Applications*, vol. 212, p. 118705, 2023.
[25] L. Ren, Y. Liu, D. Huang, K. Huang, and C. Yang, "MCTAN: A novel multichannel temporal attention-based network for industrial health indicator prediction," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 9, pp. 6456-6467, 2022.
[26] Y. Wang, Y. Lei, N. Li, T. Yan, and X. Si, "Deep multisource parallel bilinear-fusion network for remaining useful life prediction of machinery," *Reliability Engineering & System Safety*, vol. 231, p. 109006, 2023.
[27] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12, pp. 5947-5959, 2018.
[28] K. Han *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87-110, 2022.
[29] Y. Jiang, T. Xia, D. Wang, X. Fang, and L. Xi, "Spatiotemporal denoising wavelet network for infrared thermography-based machine prognostics integrating ensemble uncertainty," *Mechanical Systems and Signal Processing*, vol. 173, p. 109014, 2022.
[30] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation," *Computational Statistics & Data Analysis*, vol. 142, p. 106816, 2020.
[31] B. Wang, Y. Lei, T. Yan, N. Li, and L. Guo, "Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery," *Neurocomputing*, vol. 379, pp. 117-129, 2020, doi: 10.1016/j.neucom.2019.10.064.

Yuan Wang received the B.S. degree in mechanical design, manufacturing and automation from University of Electronic Science and Technology of China (UESTC), in 2018. He is currently working toward the Ph.D. degree in mechanical engineering at the Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi'an Jiaotong University, P. R. China.

His research interests include multimodal learning, condition monitoring, industrial robot maintenance, and remaining useful life prediction of machinery.

Yaguo Lei (Senior Member, IEEE) received the B.S. and Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, P.R. China, in 2002 and 2007, respectively. He is currently a Full Professor of mechanical engineering at Xi'an Jiaotong University. Prior to joining Xi'an Jiaotong University in 2010, he was a Postdoctoral Research Fellow with the University of Alberta, Canada. He was also an Alexander von Humboldt Fellow with the University of Duisburg-Essen, Germany. His research interests intelligent fault diagnosis and remaining useful life prediction.

Prof. Lei is a Fellow of ASME, IET, and ISEAM. He is currently an Associate Editor or Editorial Board member of more than ten journals, including *IEEE Transactions on Industrial Electronics* and *Mechanical Systems and Signal Processing*.

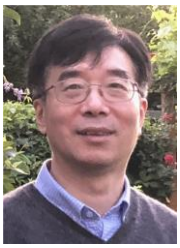
Naipeng Li (Senior Member, IEEE) is currently an Associate Professor of mechanical engineering at Xi'an Jiaotong University. He received the B.S. degree in mechanical engineering from Shandong Agricultural University, P. R. China, in 2012, and the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, P. R. China, in 2019. He was also a visiting scholar of Georgia Institute of Technology, Atlanta, USA.

His research interests include machinery condition monitoring, intelligent fault diagnostics and remaining useful life prediction of machinery.

Ke Feng (Member, IEEE) is a Full Professor at Xi'an Jiaotong University, China. He is a Marie Curie Fellow (Imperial College London & Brunel University London). He received a Ph.D. degree from the University of New South Wales, Australia, in 2021. He worked at the University of British Columbia and the National University of Singapore in 2022 and 2023, respectively.

His main research interests include digital twins, vibration analysis, structural health monitoring, dynamics, tribology, signal processing, and machine learning. He is recognized as the Emerging Leader (2023) by the Measurement Science and Technology journal. He is the Associate Editor and Guest Editor of several journals, including *IEEE Transactions on Industrial Informatics*, *Information Fusion, Mechanical Systems and Signal Processing*, *IEEE Transactions on Industrial Cyber-Physical Systems*, *Journal of Intelligent Manufacturing*, *Structural Health Monitoring*, *Engineering Applications of Artificial Intelligence*, *IEEE Transactions on Instrumentation and Measurement*, *Measurement*, *IEEE Sensors Journal*, *IET Collaborative Intelligent Manufacturing*, *Measurement Science and Technology*, *Advances in Manufacturing*, *Proc Inst Mech Eng B J Eng Manuf, Journal of Central South University*, etc.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3573703,
IEEE INTERNET OF THINGS JOURNAL



Zidong Wang (Fellow, IEEE) received the B.Sc. degree in mathematics from Suzhou University, Suzhou, China, in 1986, and the M.Sc. degree in applied mathematics and the Ph.D. degree in electrical engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1990 and 1994, respectively.

From 1990 to 2002, he held teaching and research appointments in universities in China, Germany and U.K. He is currently a Professor of dynamical systems and computing with the Department of Computer Science, Brunel University London, Uxbridge, U.K. He has authored a number of articles in international journals. He is a holder of the Alexander von Humboldt Research Fellowship of Germany, the JSPS Research Fellowship of Japan, and the William Mong Visiting Research Fellowship of Hong Kong. His research interests include dynamical systems, signal processing, bioinformatics, and control theory and applications.

Prof. Wang is a Member of the Academia Europaea, a Member of the European Academy of Sciences and Arts, an Academician of the International Academy for Systems and Cybernetic Sciences, a Fellow of the IEEE, a Fellow of the Royal Statistical Society and a member of program committee for many international conferences. He serves (or has served) as the Editor-in-Chief for *International Journal of Systems Science*, the Editor-in-Chief for *Neurocomputing*, the Editor-in-Chief for *Systems Science and Control Engineering*, and an Associate Editor for 12 international journals including *IEEE Transactions on Automatic Control*, *IEEE Transactions on Control Systems Technology*, *IEEE Transactions on Neural Networks*, *IEEE Transactions on Signal Processing*, and *IEEE Transactions on Systems, Man, and Cybernetics—Part C*.

Yang Tan received the B.S. degree in mechanical and electronic engineering from Chongqing University, Chongqing, China in 2022. He is currently working toward the M.S. degree in mechanical engineering at the Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi'an Jiaotong University, P. R. China.

His research interests include condition monitoring and remaining useful life prediction of machinery.

Huitong Li received the B.S. degree in mechanical and electronic engineering from Wuhan Institute of Technology, Wuhan, China, in 2022. He is currently working toward the M.S. degree in mechanical engineering at the Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi'an Jiaotong University, P. R. China.

His research interests include condition monitoring and remaining useful life prediction of machinery.