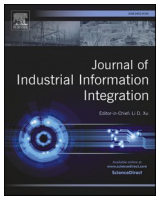
Journal of Industrial Information Integration 46 (2025) 100854

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.sciencedirect.com/journal/journal-of-industrial-information-integration



Full Length Article

# Collaborative and trustworthy fault diagnosis for mechanical systems based on probabilistic neural network with decision-level information fusion

Zifei Xu [a], Kaicheng Zhao [b], Wanfu Zhang [b], Weipao Miao [b], Kang Sun [b], Jin Wang [c], Musa Bashir [a,*]

[a] Department of Civil and Environmental Engineering, University of Liverpool, Liverpool, UK
[b] School of Energy and Power Engineering, University of Shanghai for Science and Technology, Shanghai, China
[c] Liverpool Logistics Offshore and Marine Research Institute, Liverpool John Moores University, Liverpool, UK

ARTICLE INFO

*Keywords:*
Trustworthy diagnostic
Uncertainty quantification
Information fusion
Reliability
Probabilistic neural network

ABSTRACT

Fault diagnosis is a critical component of prognostics and health management, enhancing machinery reliability and ensuring operational efficiency by enabling proactive maintenance strategies. However, achieving this requires high data fidelity to accurately predict the full spectrum of faults and structural degradation for reliable assessments. AI-driven fault diagnostics based on machine learning often face challenges in reliability due to uncertainties arising from variations in data distribution, caused by changing operating conditions and noise interference. These factors undermine the trustworthiness of such methods. To address these challenges in accuracy and reliability for mechanical systems, such as gearboxes, this study proposes a Trustworthy Intelligent Diagnostic (TID) model. The TID model incorporates a multi-scale probabilistic neural network, and a decision fusion module based on uncertainty quantification (UQ). Specifically, three UQ-based decision fusion strategies are introduced to enhance diagnostic reliability by effectively managing uncertainty in fault diagnosis. Building upon the TID model, a cooperative fault diagnosis framework is further proposed to facilitate fault knowledge sharing and alleviate the limitations posed by data scarcity. The proposed approach is validated using both experimental data and real-world wind turbine gearbox failure datasets, demonstrating significant improvements in diagnostic accuracy and a notable reduction in false alarm rates. These results highlight the effectiveness, reliability, and superiority of the proposed method.

## 1. Introduction

### 1.1. Background

Wind power, as one of the most promising clean energy sources, has the potential to accelerate decarbonization and achieve net-zero emissions. Wind turbines, primarily composed of rotating machinery that converts wind energy into electrical energy, serve as the main driving force behind wind power generation technology. However, several significant obstacles hinder the continued adoption and development of wind power technology, including manufacturing, installation, and the high costs associated with operation and maintenance (O&M). Reducing O&M costs can significantly enhance the sustainability of wind energy. Conventional maintenance approaches can extend the service life of wind turbines, but they do so at significantly higher costs compared to predictive maintenance methods. Predictive maintenance, as one of the advanced approaches in Prognostics and Health Management (PHM) technology, offers the most effective way to reduce O&M costs [1,2]. Fig. 1 presents the flowchart of a general PHM process, highlighting that fault diagnosis plays a critical role in ensuring the reliability and accuracy of diagnostic outcomes. These outcomes directly influence the effectiveness of the entire PHM process, including maintenance decision-making strategies.

Motivated by widespread use of Artificial Intelligence (AI) across various fields, such as aviation, healthcare, navigation, etc., [3–5], new insight have emerged on how to address the challenges in diagnostic technologies. Compared to traditional diagnostics, AI-driven fault diagnostics integrate the advantages of knowledge-based, model-based, and purely data-driven approaches. However, these AI-based methods still suffer from low trustworthiness, and existing solutions have yet to fully overcome these limitations. As a result, AI-driven diagnostics continue to face significant risks, along with low accuracy and reliability

* Corresponding author.
*E-mail addresses:* zifei.xu@liverpool.ac.uk (Z. Xu), m.b.bashir@liverpool.ac.uk (M. Bashir).

https://doi.org/10.1016/j.jii.2025.100854

Available online 7 May 2025

in real-world applications. Therefore, the primary objective of this research is to develop a trustworthy AI-driven diagnostic system to enhance the reliability and safety of AI application in PHM process [6].

### 1.2. Related works

Modern predictive tools are increasingly developed using Machine Learning (ML), an important subset of AI, as part of PHM to enable intelligent fault diagnosis. Several diagnostic tools developed based on ML have been successfully applied in medical diagnostics [7], and a number of studies have extended this concept to the fault diagnosis of mechanical systems, such as wind turbines [8]. In contrast to medical applications, where ML models often rely on pixel segmentation, fault diagnosis in wind turbines typically involves analyzing vibration signals [9].

However, since vibration signals are easily obscured by noise, changes in rotational speed can lead to shifts in the data distribution of these signals. This introduces a new challenge in applying ML to wind turbine fault diagnosis: how to establish a meaningful and robust relationship between features representing healthy operating conditions and those indicative of failure modes [10–15]. To address the above issues, Xu *et al.* combined the strengths of Variational Mode Decomposition (VMD) and Convolutional Neural Networks (CNN) to develop a bearing fault diagnosis model. Their research demonstrated the reliability of integrating multi-physical signals with neural networks, providing a foundation for automated fault diagnosis. However, despite the potential, combining such algorithms to develop diagnostic models has been shown to negatively impact maintenance strategies in terms of both accuracy and efficiency [16]. Thus, Huang et al. proposed the use of a parallel convolutional kernels with different sizes to directly obtain the multi-scale information from the vibration signal of bearings. The results indicate that the fusion of multi-scale features can improve the diagnostic accuracy of the model [17]. However, a similar study conducted by Jiang *et al.* considered the impact that the depth of feature fusion has on the diagnostic performance of the NN-based model. The study used the multi-scale advanced features as the input for the fully connected layer to calculate the diagnostic probabilities [18]. As more in-depth studies, Xu *et al.* and Bashir *et al.* used the different contributions from multi-scale features from the predicted probabilities to calculate more accurate diagnostic probabilities. The researchers subsequently added attention mechanism to the weights and fused it with the multi-scale features [19,20]. The diagnostic probabilities of each branch of filters in the hybrid multi-scale model are fused using a developed ensemble network. The results show that the methods can reduce the rate of false positives [21]. Besides, as more credible studies in the diagnosis area continue to emerge, domain shift for condition adaptation [22] in diagnosis has attracted attention of researchers in this field of study. Zhang *et al.* developed a fault diagnosis model based on transfer learning that can more accurately detect faults by restraining the transfer of ineffective information [23]. Wang et al. also established an intelligent diagnostic model based on transfer learning. The results indicated that the biases between the marginal and conditional distribution of objectives target, and original source were reduced [24]. Cao et al. developed the Y-net model based on transfer learning network that can overcome the limitations in robustness caused by the domain shift. The result showed that the reduction of the discrepancy between marginal and conditional distribution of the learned features can improve the diagnostic performance [25]. Song *et al.* developed a data-driven model which can maintain the maximum training accuracy to minimize the difference between training and test data to address domain adaption limitations [26]. Meng et al. developed a data-driven diagnosis model based on empirical mode reconstruction to enhance the data in use of training the model for diagnostic reliability [27]. Furthermore, Ragab et al. developed a fault diagnosis framework based on multiple machine learning modules and to make a reliable decision for fault classification for improvement of diagnostic reliability [28]. However, the diagnostic uncertainties caused from the limited data that cannot provide prefect description of the distribution of the target source leading to an out-of-distribution (ODD) problems in the diagnostic problem [29].

AI-driven methods have shown great potential across various real-world applications, particularly where predictive decision-making and system reliability are crucial [30]. For example, in the field of smart healthcare, artificial neural network-based federated learning has been successfully applied to heart stroke prediction [31]. In industrial asset management, AI techniques are widely used in PHM to detect

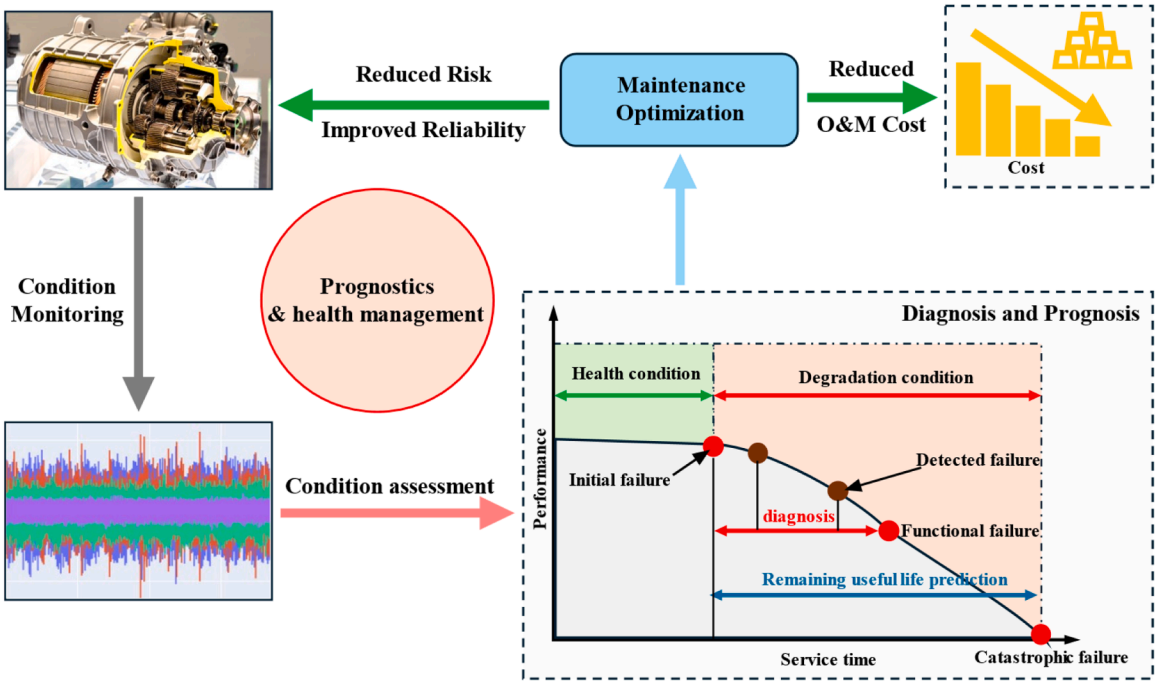



**Fig. 1.** Relationship and importance of diagnosis in prognostics and health management.

early-stage faults and optimize maintenance strategies [32,33]. Furthermore, AI also plays an increasingly important role in energy systems, where it enables real-time anomaly detection, intelligent control, and digital twin development for complex infrastructures such as wind turbines [34].

Uncertainty quantification (UQ) is increasingly becoming an essential component of intelligent neural network-based diagnostic models [35]. has long been a primary source of unreliable decisions in often leading to inaccurate fault or damage estimation by intelligent models, which in turn can pose safety risks to the overall integrity of engineering system [36,37]. Recently, researchers have recently used Bayesian Deep Neural Network (BDNN) to estimate both aleatoric and epistemic uncertainties in fault diagnosis models, demonstrating their promising capabilities [38,39]. However, these studies have primarily focused on proving the feasibility of quantifying uncertainty within BDNN-based fault diagnosis models [40–42].

Zhu et al. developed intelligent diagnosis based on Bayesian transfer learning under limited data. The study demonstrates that using Bayesian neural network can avoid overfitting during statistical ML modelling [43]. Li et al. used Bayesian learning to develop an approach to analyze random vibration of bridge. The examination result indicated that Bayesian approach could stabilize the prediction of vehicle dynamic interaction [44]. Chen et al. used Bayesian optimization to obtain the optimal parameters of Bayesian neural network to address fault diagnosis model for bearing diagnosis [45]. The results showed that the method can achieve high diagnosis accuracy. Xu et al. developed a damage prediction model based on Bayesian neural network to estimate the degradation capability of rolling bearing [46].

Soualhi et al. investigated the uncertainty in the Remaining Useful Lifetime (RUL) prediction to develop a data-driven model to combine the direct and recursive RUL predictions for reliability of the predicted RUL [47]. Lin et al. developed an intelligent diagnosis model based on variational mode decomposition and probabilistic neural network to improve the diagnostic results in noisy environments [48]. Yao and Wang established the fault diagnosis model based on fractal theory and probabilistic neural network to improve the diagnostic capability for the complex system. Their model considers the nonlinear dynamic characteristics of the complex (mechanical) system [49]. Fang et al. designed a fault diagnosis model based on Bayesian CNN that can quantify the uncertainty, especially when the data contains noise [50]. A similar study by Peng et al. constructed a 2-D Bayesian CNN to estimate bearing deflections and the uncertainty in the prediction [51]. However, this study used images as input of the model. Similarly, Fang et al. used the spectrum transferred by short time Fourier transformation (STFT) as the input of the model [48]. The study led to some loss in diagnostic capabilities through the data transformation when the uncertainty was enlarged. Thus, it should be noted that using raw data as the input of NNs for diagnostic modelling has some inherent limitations. Furthermore, more advanced studies have been conducted by Zhou et al. using raw vibration signals to develop the BDNN diagnostic model with UQ capability to analysis to improve the reliability of the diagnosis model [52]. Han et al. established a fault diagnosis approach based on BDNN with ensemble layer to integrate the probability outputs from multiple deep learners to improve uncertainty estimation. It has been determined that the BDNN-based fault diagnosis model can accurately classify the working conditions and predict the uncertainty for decision making [53]. Consequently, the above studies only re-emphasized the importance of UQ in fault diagnosis. More importantly, intelligent models should be able to provide high confidence level when diagnosing known or unknown faults. The results of the UQ should have high accuracy for them to be timely used as feedback in the decision-making system as an alarm. Accordingly, this paper presents the first research to propose the use of uncertainty to develop and implement trustworthiness in fault diagnosis decision-making using a multi-scale Bayesian neural network. The study shows how to quantify, control, and calibrate the uncertainty for diagnostic reliability in order to improve the reliability of the PHM system of wind turbine.

### 1.3. Summary, motivation and contribution

Recent advances in AI-driven fault diagnosis have achieved significant success in industrial applications. Through techniques such as transfer learning and generative modeling, intelligent diagnostic systems have demonstrated impressive performance in small-sample scenarios, effectively addressing many challenges in industrial environments. However, as AI-driven diagnostic decisions become more integrated into real-world operations, ensuring the reliability and trustworthiness of these conclusions is critical, especially in high-stakes systems. Motivated by the need to enhance diagnostic reliability under uncertainty and improve the credibility of AI-based decisions in out-of-distribution (OOD) environments, this study aims to improve the trustworthiness of intelligent fault diagnosis conclusions to ensure the reliability of PHM. This, in turn, supports the structural health monitoring and assessment of wind turbine rotating machinery.

To achieve this goal, we propose a novel Trustworthy Intelligent Diagnostic (TID) model, which consists of a multi-scale probabilistic neural network and a decision fusion module based on UQ. This model demonstrates strong extrapolation capability and robustness when performing fault diagnosis in complex environments. Furthermore, by leveraging the TID model and its UQ-based decision fusion module's ability to distinguish uncertainty between OOD and In-Distribution (ID) data, we introduce a collaborative diagnostic framework. This framework enhances fault diagnosis in OOD scenarios by utilizing the collaborative effects of multiple models, significantly reducing the number of samples that require additional manual inspection. The key contributions of this study are as follows:

(1) A newly developed multi-scale probabilistic neural network model has been created by integrating Bayesian neural networks with a multi-scale fault diagnosis framework. This model efficiently extracts valuable information from raw vibration signals, resulting in enhanced overall performance.
(2) Based on UQ of predictions, three UQ-based decision fusion modules are designed, including UQ-based hard voting, UQ-based soft voting, and UQ-based Bayesian causal inference. These decision fusion modules provide more confident inferences for ID diagnoses, resulting in lower epistemic uncertainty. Conversely, UQ-based hard voting and UQ-based soft voting amplify epistemic uncertainty for OOD diagnoses, thereby enabling a clear distinction between ID and OOD cases.
(3) A UQ-based collaborative diagnostic framework is proposed, which leverages accurate OOD identification to trigger an additional Trustworthy Intelligent Classifier (TIC) for further recognition of remaining OOD cases. This approach significantly improves the accurate identification of unknown faults, ultimately reducing labor costs.

The remainder of the paper is structured as follows. Section 2 presents the preliminary methods used in the proposed approach. Section 3 introduces the proposed methodologies, including the trustworthy intelligent diagnostic model, the decision-level fusion modules based on uncertainty quantification, and the collaborative diagnostic framework. Section 4 presents the experimental results and discussion. Section 5 concludes the study, while Section 6 discusses future directions for reliable intelligent diagnosis.

## 2. Preliminaries

### 2.1. Bayesian neural network

The fault diagnosis method based on traditional deep NNs falls under the category of point estimation problems, which suffer from inherent

overconfidence in predictions during diagnosis [54]. The classical fault diagnosis approach can be defined as the model parameters $w$, which is optimized to obtain the maximum likelihood estimation (MLE):

$$\mathrm{w}^{*} = \underset{\mathrm{w}}{\operatorname{argmax}}\,\mathrm{P}(\mathrm{D}|\mathrm{w}) = \underset{w}{\operatorname{argmax}} \sum\nolimits_{i} \log p(\mathrm{y_i}|\mathrm{x_i}, w) \tag{1}$$

In the BDNN model, the distribution parameters are defined as $P(w|D)$ to enable a neural network model with interval estimation. The model's capability to quantify uncertainty is derived from the random parameters in the BDNN [55]. Fig. 2 shows the topological structure of a traditional BDNN used in this study for benchmarking purposes.

The Bayes rule is used to find the posterior distribution over the weights $p(w|D) = \frac{p(D|w)p(w)}{p(D)=\frac{p(D|w)p(w)}{\int_w p(D|w)p(w)dw}}$ where $p(w|D)$ is the likelihood estimated probability based on the dataset, $p(w)$ is the prior distribution, and $p(D)$ is the marginal likelihood. The Gaussian distribution is generally set as the prior in the study. For a set of $w$, the $p(D|w)$ and $p(w)$ are trackable. On the contrary, the marginal likelihood $p(D)$ is intractable because $\int_w p(D|w)p(w)dw$ is very difficult to calculate. Thus, in this study, variational inference is used to address this problem and to approximate the $p(D)$.

### 2.2. Variational inference

Variational inference defines a distribution $q(w|D)$ to assimilate with the unknown distribution $p(w|D)$. Generally, the $q(w)$ is much easier to be collected than $p(w|D)$. To make the $q(w|D)$ much closer to the distribution $p(w|D)$, the Kullback-Leibler (KL) divergence is used to estimate the similarity between the two distributions [56], $p(x)$ and $q(x)$. **KL** is defined as $\mathrm{KL}[q(x) \parallel p(x)] = E_{q(x)}\log\left(\frac{q(x)}{p(x))} = \int q(x)\log(q(x)/(p(x))dx$. The true posterior distribution $p(w|D)$ is calculated by minimizing the **KL** divergence and the variational distribution $q(w)$. The approximation problem is transformed into an optimization problem as $\theta_{opt} = \underset{\theta}{\operatorname{argmin}}\,\mathrm{KL}[q(w|D) \parallel P(w|D)]$. However, it is difficult to directly solve the **KL** except through maximizing the evidence lower bound (ELBO) to correspond with an equivalently minimized **KL**.

$$\theta_{opt} = \underset{\theta}{\operatorname{argmin}} E_{q_\theta(w|D)}\left[\frac{q_\theta(w|D)}{p_\theta(w|D)}\right]$$

$$= \underset{\theta}{\operatorname{argmin}} E_{q_\theta(w|D)}[\log q_\theta(w|D)] - E_{q_\theta(w|D)}[\log p_\theta(w|D)]$$

$$= \underset{\theta}{\operatorname{argmin}} E_{q_\theta(w|D)}[\log q_\theta(w|D)] - E_{q_\theta(w|D)}\left[\log\frac{p(w,D)}{p(D)}\right]$$

$$= \underset{\theta}{\operatorname{argmin}} \; -E_{q_\theta(w|D)}[\log p(w,D)] - E_{q_\theta(w|D)}[\log q_\theta(w|D)] + \log p(D) \tag{2}$$

In order to minimize the computational complexity, $\theta_{opt}$ is addressed through sampling from the variational posterior distribution $q_\theta(w)$. It is assumed that K samples are extracted from the $q_\theta(w)$ following Eq. (3). The loss function for backpropagation to optimize the parameters of the BDNN consists of the **KL** and the cross entropy [57].

$$\begin{aligned}\theta_{opt} = &\underset{\theta}{argmin} \; -E_{q_\theta(w|D)}[\log p(w,D)] - E_{q_\theta(w|D)}[\log q_\theta(w|D)] + \log p(D) \\ \approx &\underset{\theta}{argmin} \sum_{k=1}^{K} \log p(w,D) - \log q_\theta(w|D) + \log p(D)\end{aligned} \tag{3}$$

### 2.3. Uncertainty quantification

MC dropout is used to conduct uncertainty predictions and quantify the uncertainty of this dynamic prediction results. Assuming the test data is represented by $\mathbb{x}$, to sample the model parameters $K$ times, for the $k^{\mathrm{th}}$ sampling, the prediction of the intelligent diagnostic model is $\hat{\mathbb{y}}^k = \mathrm{model}(\mathbb{x}|\omega^k)$, where $\hat{\mathbb{y}}^k$ represents the probability distribution of the estimated status. In the case of $\hat{\mathbb{y}}^k$ equals to $k^{\mathrm{th}}$ output of the AI-driven model, the average prediction after sampling K times is $\overline{\mathbb{y}}$. Based on the entropy, the model is able to quantify the prediction uncertainty [38]. The total uncertainty PU for each $\mathbb{x}$ can be quantified by Eq. (4).

$$\mathrm{PU} = \mathbb{H}(\overline{\mathbb{y}}|\mathbb{x}) = -\sum_{c=1}^{C} \mathbb{p}(\overline{\mathbb{y}} = c|\mathbb{x}) \;\times\; \log[\mathbb{p}(\overline{\mathbb{y}} = c|\mathbb{x})] \tag{4}$$

By considering total uncertainties, PU can be decomposed into aleatoric (AU) and epistemic uncertainties (EU) that respectively refer to the uncertainty inherent to the input samples and model parameters. The AU can be approximated by Eq. (5).

$$\mathrm{AU} \approx -\frac{1}{\mathrm{K}}\sum_{k=1}^{K}\sum_{c=1}^{C} \mathbb{p}\left(\overline{\mathbb{y}}^k = c|\mathbb{x}\right) \;\times\; \log\left[\mathbb{p}\left(\overline{\mathbb{y}}^k = c|\mathbb{x}\right)\right] \tag{5}$$

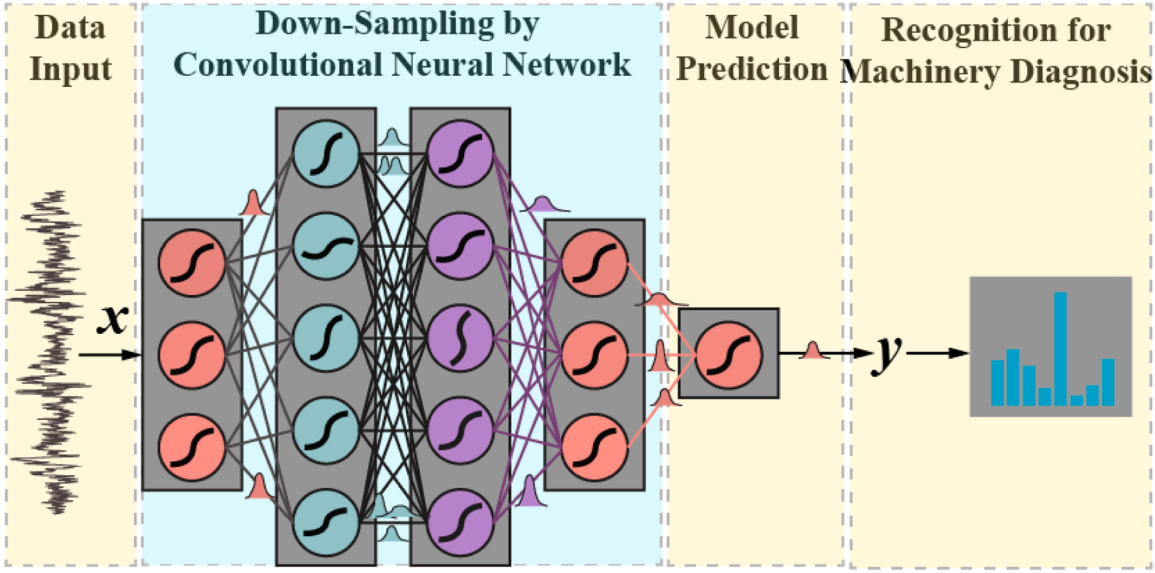



**Fig. 2.** Illustration of DBNN topological architecture.

According to the definition, epistemic uncertainty is the difference between total uncertainty and aleatoric uncertainty, which can be approximated as:

$$\mathrm{EU} \approx -\sum_{c=1}^{C} \mathbb{p}(\overline{\mathbb{y}}=c|\mathbb{x}) \times \log[\mathbb{p}(\overline{\mathbb{y}}=c|\mathbb{x})] + \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} \mathbb{p}(\overline{\mathbb{y}}^k=c|\mathbb{x}) \times \log\left[\mathbb{p}(\overline{\mathbb{y}}^k=c|\mathbb{x})\right] \tag{6}$$

## 3. Proposed methodologies

### 3.1. Trustworthy intelligent diagnostic model

In this study, a TID model was developed based on a Bayesian multi-scale convolutional neural network and the proposed UQ-based decision-level fusion module. The framework for trustworthy intelligent diagnostics was designed based on the works of Xu et al. [58]. who demonstrated that the diagnostic accuracy of AI-driven models can be enhanced by incorporating a multi-scale feature extraction layer into the framework and employing an appropriate information fusion strategy.

Fig. 3 illustrates the concept of the proposed TID framework. It comprises three key components: a multi-scale feature extractor, multiple parallel Bayesian neural networks (BNNs) for feature extraction, and a decision-fusion module. The multi-scale extractor enhances the model's capability by providing richer diagnostic information, improving its ability to handle complex patterns. Bayesian neural networks facilitate uncertainty-aware diagnosis by quantifying diagnostic uncertainty, enabling reliability assessment of the inference process to determine its trustworthiness. The decision-fusion module integrates information at the decision level, leveraging quantified uncertainty to assess the reliability of each sub-network operating at different scales. This approach ensures a more robust and reliable diagnostic outcome. To further enhance diagnostic trustworthiness, three UQ-based decision-fusion strategies are proposed in Section 3.2, specifically designed to manage uncertainty and improve the reliability of diagnostic decisions.

The TID model directly acts on the raw data collected by the sensor from the rotating machinery, which contains a *multi-scale extractor* to capture multi-scale features from raw vibration. The multi-scale extractor is addressed by average pool, for each data point, the multi-scale feature can be calculated by Eq. (7)

$$\mathrm{output}[l'] = \frac{1}{\tau}\sum_{l=l'}^{l'+\tau-1} \mathrm{input}[l] \tag{7}$$

The input is a raw vibration $x(t)$, where $l$ represents the position over the time series $x(t)$, and $\tau$ is the kernel size of the average pool, representing the multi-scale factor. Padding procedure is used to keep the length of the output sequence same as the input sequence. In this study, the maximum scale factor sets three which is the same as the multiscale diagnostic research [59,60]. In that case, there should be a total of three BNNs models applied on each scale vibration features$x_\tau(t)$. In the down-sampling module, consisting of BNNs, each independent network $f_\tau$ developed based on stack of layers including Bayesian convolutional layer BayesianConv1d$(\cdot)$, a dropout layer Dropout$(\cdot)$, an activation layer ReLU$(\cdot)$, and Bayesian batch normalization layer BayesianBatchnorm1d$(\cdot)$ works on each single $x_\tau(t)$.

$$\mathrm{BayesianConv1d}(x) = \mathbb{W} * x + \mathbb{b},\ \mathbb{W} \sim \mathcal{N}(\mu_{\mathbb{W}}, \sigma_{\mathbb{W}}^2),\ \mathbb{b} \sim \mathcal{N}(\mu_{\mathbb{b}}, \sigma_{\mathbb{b}}^2) \tag{8}$$

$$\mathrm{Dropout(x)} = \begin{cases} 0 \\ \dfrac{\mathrm{BayesianConv1d}(x)}{1-\mathbb{p}} \end{cases},\ \mathbb{p} \sim \mathrm{Bernoulli}(\theta) \tag{9}$$

$$\mathrm{BayesianBatchnorm1d(x)} = \frac{\mathrm{BayesianConv1d}(x) - E(\mathrm{BayesianConv1d}(x))}{\sqrt{\mathrm{Var}(\mathrm{BayesianConv1d}(x)) + \varepsilon}} \tag{10}$$

$$\mathrm{ReLU(x)} = \max(0,\ x) \tag{11}$$

The hidden features will be decoded by a fully connected layer with a softmax function to produce the diagnosis result $\widehat{y}$. Algorithm 1 presents

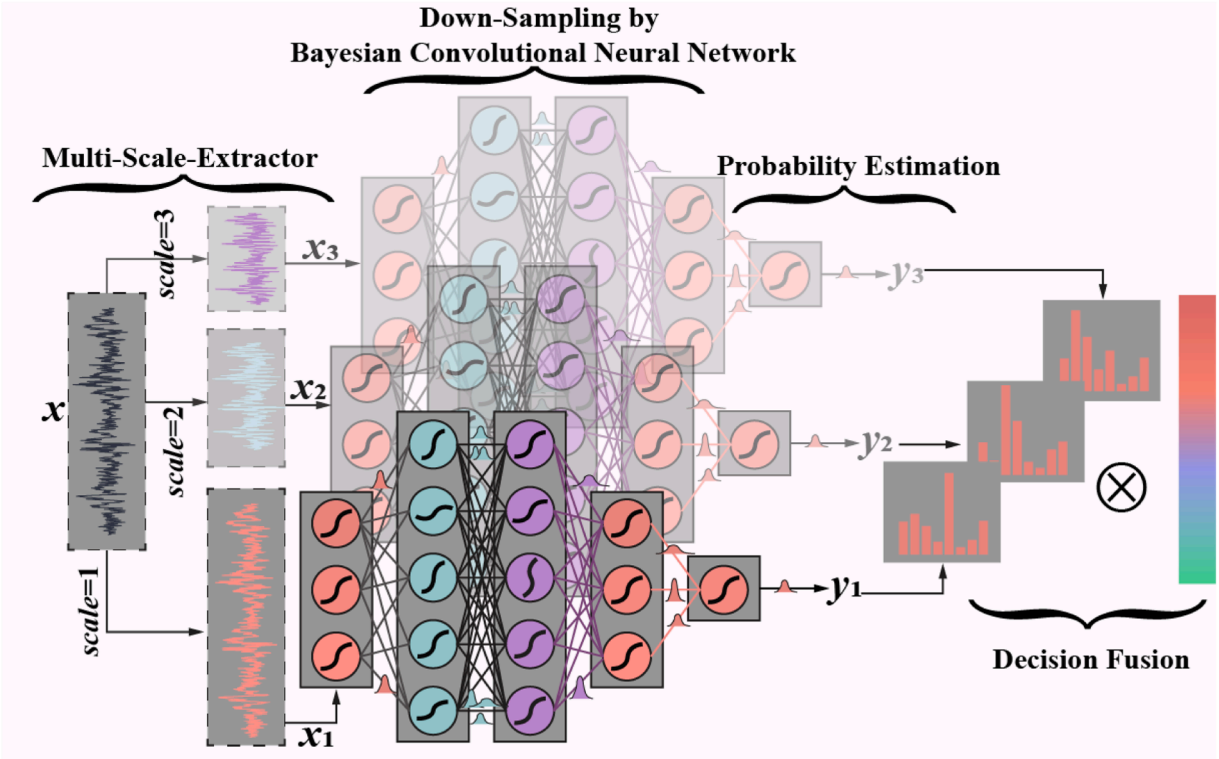



**Fig. 3.** Illustration of the trustworthy intelligent diagnostic (TID).

**Algorithm 1**
TID model optimization.

Model Optimization

**Input**: Training Data $\mathscr{D}_{train} = \{\mathscr{X},\ \mathscr{Y}\}$ Models $f_j(\cdot)$ with initialized parameters $\omega_j$, learning rate $\eta$.
**Output**: $f_j(\cdot|\widehat{\omega}_j)$, $j$ is number of single sub-network in TID model, corresponding to multi-scale factor $\tau$
**Training TID branch network**:
for *epoch* in maximum epochs:
Samples $[\boldsymbol{x}, y)]$ in $(\mathscr{X}, \mathscr{Y})$
**Parallel Down sampling by Encoders**: $f_j(\cdot|\omega_j)$
**Fault prediction by each parallel network**: $\hat{y}_j$
**loss calculation**: by $CrossEntropy\left(\hat{y}_j, y\right) + \mathrm{KL}\ (\omega_j,\ \mathscr{D})$ *by Equation* (3)
(By performing $k$ iterations of Monte Carlo sampling, the loss and performance are averaged over the $k$ iterations)
Parameters $\omega_j$ are optimized by optimizer with learning rate $\eta$
Saving the best performance models (By performing $k$ iterations of Monte Carlo sampling, the loss and performance are averaged over the $k$ iterations)
**end**
**Output**: $f_j(\cdot|\widehat{\omega}_j)$
**End**

the training process of the TID model.

### 3.2. Decision-level fusion module based on uncertainty quantification

The method of information fusion significantly impacts the decision-making process of intelligent models. For multi-scale AI models, while it has been verified that integrating multi-scale information can improve fault diagnosis performance, existing fusion approaches do not account for the reliability of decision information and its effect on the final diagnosis outcome. Therefore, this section introduces a decision-level information fusion module based on UQ. Specifically, three decision fusion strategies are proposed: UQ-based hard voting, UQ-based soft voting, and a decision fusion module based on Bayesian causal inference. In training process of BCI, First, we use the training dataset $\mathscr{D}_{train} = \{\mathscr{X},\ \mathscr{Y}\}$ and the TID model $f_j(\cdot|\widehat{\omega}_j)$ to generate an estimated $\widehat{\mathbb{y}}_i$ and the corresponding epistemic uncertainty $\mathrm{EU}_i$ for each sample. Theye are the observed data to fit the Bayesian network's structure and parameters, which are subsequently utilized during the inference phase. Algorithm 2 illustrates how the Trustworthy Intelligent Diagnostic (TID) model performs decision fusion during fault diagnosis.

**Algorithm 2**
Decision-level fusion module based on UQ.

**Uncertainty Quantification and UQ-based decision fusion module:**
**Input:** Testing Data $\mathscr{D}_{test} = \mathscr{X}_{test}$, TID model: $\mathrm{TID}(\cdot|\widehat{\omega}_i)$, sampling time K, **Output:** Diagnostic result $\widehat{\mathscr{Y}}$, Bayesian network
**for** $\mathbb{x}$ in $\mathscr{X}_{test}$
**Sampling** K times from the branch network of TID model: $TID(\cdot|\widehat{\omega}_i)$
where $k^{\text{th}}$ prediction is $\widehat{\mathbb{y}}_i^k = \mathrm{TID}(\mathbb{x}|\widehat{\omega}_i^k)$
**Record** $\widehat{\mathbb{y}}_i = \left\{\widehat{\mathbb{y}}_i^k | k \in N,\ 1 \le k \le \mathrm{K}\right\}$
**Obtain** epistemic uncertainty $\mathbb{EU} = \{\mathrm{EU}_i | i \in N,\ 1 \le k \le \mathrm{n}\}$ by Equation (4) to Equation (6)
**Trustworthy decision-level fusion:**
**if** ***'hard-vote'***:
**Trustworthy index:** $idx = \mathrm{argmin}(\mathbb{EU})$
**Trustworthy diagnosis** for $\mathbb{x}$**:** $\widehat{\mathbb{y}} = \left\{\widehat{\mathbb{y}}_i | i \in N,\ 1 \le k \le \mathrm{n}\right\}[idx]$
**elseif** ***'soft-vote'***:
**Trustworthy weights:** $\mathbb{w} = softmax\left(\mathbb{w}^{-1}/sum\left(\mathbb{w}^{-1}\right)\right)$
**Trustworthy diagnosis** $\mathbb{x}$ : $\widehat{\mathbb{y}} = \frac{1}{N}\sum \mathbb{w}_i \cdot \widehat{\mathbb{y}}_i$
**else: (Bayesian causal inference):**
**Trustworthy diagnosis** for $\mathbb{x}$**:** $\widehat{\mathbb{y}} = \mathrm{P}\left(\mathbb{y}\middle|\widehat{\mathbb{y}}_1\widehat{\mathbb{y}}_2,..\widehat{\mathbb{y}}_n, \mathrm{EU}_1, \mathrm{EU}_2, ...\mathrm{EU}_n\right)$
**end**
**end**
**Diagnosis results** for $\mathscr{X}_{test}$: $\widehat{\mathscr{Y}} = \{\widehat{\mathbb{y}}\}$
**End**
**Output: Fault diagnostic:** $\widehat{\mathscr{Y}}$
**End**

### 3.3. Collaborative diagnostic framework based on uncertainty quantification

A collaborative fault diagnosis framework based on TID models is proposed in this study, leveraging UQ. In industrial applications, there is often a large amount of out-of-distribution (OOD) data that must be accurately detected, yet each AI-driven diagnostic model has limited coverage of potential faults. To address this, our framework calls upon a second model whenever the first model identifies an input as OOD, enabling more OOD data to be recognized. This complementary cooperative diagnosis not only substantially reduces the misidentification of OOD samples but also eliminates the need for retraining both models, thereby greatly saving computational resources. The process diagram of the collaborative fault diagnostic framework is presented as follows.

Fig. 4 illustrates a hierarchical fault diagnosis framework that leverages trustworthy Intelligent Diagnostic Models with diverse knowledge sources. This framework integrates uncertainty quantification and decomposition to identify OOD data, progressively enhancing diagnostic coverage through automated and intelligent diagnosis. The process follows these steps: 1) Sensor data collection for the Task-Informed Diagnosis (TID) model. 2) Initial diagnosis is performed by TID1, analysing fault patterns in the input data.3) Uncertainty quantification and decomposition are applied to classify uncertainty into epistemic uncertainty and aleatoric uncertainty. 4) OOD identification is conducted by comparing the epistemic uncertainty against a predefined threshold of 1.5 times the interquartile range (IQR), derived from historical fault records. 5) If the epistemic uncertainty exceeds the threshold, the data is labelled as OOD; otherwise, it is considered valid diagnostic data. 6) The OOD data is transferred to the next TID model (TID2) for further analysis, enabling multiple models to collaborate and improve diagnostic accuracy. 7) The TID2 model analyses the received OOD data and attempts to classify its fault pattern. 8) *Re*-evaluation of uncertainty: The TID2 diagnosis result undergoes another 1.5 × IQR OOD evaluation. 9) Non-OOD data is directly output as a diagnosis result. Remaining OOD data is passed to the next available TID model for further processing until no extra TID model can use. For complex fault diagnosis tasks, this iterative approach eliminates the need to train a large monolithic model, while reducing manual effort in verifying OOD samples. Through collaborative multi-model inference, the system enhances diagnostic completeness and efficiency.

## 4. Experiment and discussion

### 4.1. Descriptions of datasets and experimental setup

Gearbox fault dataset is used to examine and analyze the capability of the proposed model for uncertainty quantification in this study. The gearbox examination focused on investigating two different working

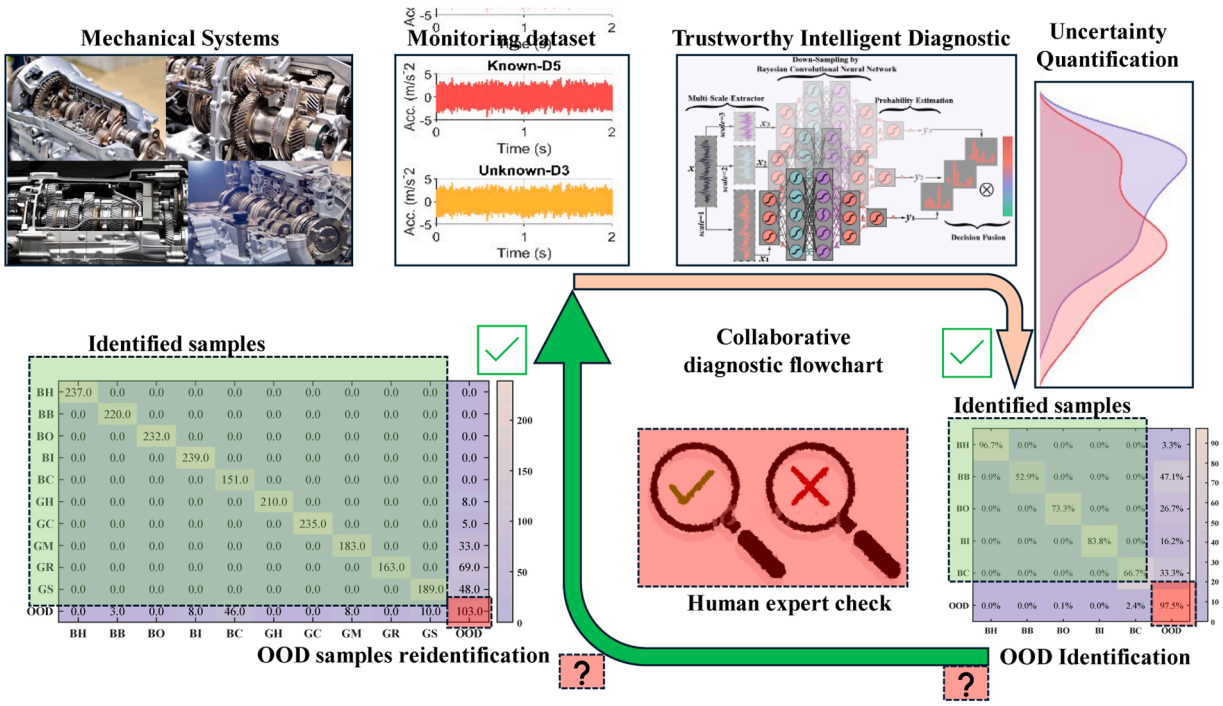



**Fig. 4.** Collaborative diagnostic framework based on TID and UQ.

conditions in which the loads on the system are set at 20 Hz-0 V or 30 Hz-2 V. This method allows for an accurate representation of the gearbox's health status. Details of bearing and gear types and their respective diagnosed faults are presented in Table 1 [61].

Based on the above categories of the working conditions, several experiments are constructed to examine the diagnosis and uncertainty quantification. Details of these experiments are presented in Table 2.

Wind turbine condition monitoring benchmarking dataset, provided by National Renewable Energy Laboratory (NREL), is used to examine the proposed TID model. The test turbine drive train configuration and the vibration sensor locations are shown in Fig. 5 [62].

Table 3 presents a list of the actual damage on the gearbox that have been detected through vibration analysis. The desired sensors for the intermediate speed (IS) and high speed (HS) are designated as AN 5 to AN 9. The relationships between these sensors and the components are determined based on the location and proximity to the rotating components, as listed in Table 3 [63].

Using this dataset, a test environment based on a real-life wind turbine gearbox damage cases was constructed to examine the reliability and validity of the proposed method. Details of the examination are presented in Table 4.

In the table, H indicates that the data is collected from a healthy condition while D represents data from a damaged condition. Numbers 1 to 5 indicate the five categories. In each experimental scenario, the training data is collected on day 1 and day 2. The testing data is collected from days 3 to 10. Each sample has 2048 data points, without any overlap between the samples.

Fig. 6 presents kernel density estimation plots of sensor data distributions collected across different days from five sensors (AN5 to AN9). Each curve represents data collected on a specific day (D1 to D10), where H indicates data from a healthy condition and D represents data from a damaged condition. Notably, for sensors AN5, AN6, AN7, and AN8, the distributions exhibit variations over time, reflecting the inherent instability in the equipment's operating conditions. In this study, the training data is derived from the first two days (D1 and D2), while the test data spans days 3 to 10. This experimental setup ensures a rigorous evaluation of the model's performance in both ID and OOD scenarios. The ID testing accounts for variations in normal operating conditions, whereas OOD testing assesses the model's ability to generalize when encountering data from different time periods, capturing the evolving nature of the system. By incorporating data from different days and considering real-world operational fluctuations, this approach effectively evaluates the robustness and generalization capability of the diagnostic model. It ensures that the model is not merely overfitting to a specific dataset but can adapt to varying conditions, making it more reliable for real-world applications.

To further verify the reliability and generalizability of the proposed method for fault diagnosis applications, it is also applied to an aero-engine-based test platform for intelligent diagnosis validation. The structure of the motor system is illustrated in Fig. 7.

Fig. 7 shows the experimental test rig designed based on a real dual-rotor aero-engine. The system includes a motor drive system, a lubricant

**Table 1**
Gearbox fault types description.

| Type | Description | Type | Description |
|---|---|---|---|
| GH | Gear healthy | BH | Bearing Healthy |
| GC | Gear feet crack | GM | Missing gear feet |
| GR | Gear root feet crack | GS | Gear surface wear |
| BB | Ball crack | BO | Outer race crack |
| BI | Inner race crack | BC | Inner race and outer race crack |

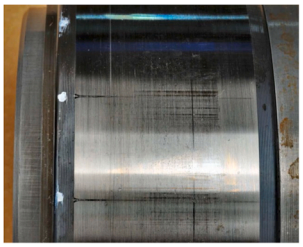
**Table 2**
Description of experiments.

| Experiment | Training | Testing (In Domain) | Testing (Out of Domain) |
|---|---|---|---|
| No.1 | 20 Hz-0V:GH, GC, GR, GM, GS | 20 Hz-0V:GH, GC, GR, GM, GS (−10 dB ~ 10 dB) | - |
| No.2 | 20 Hz-0V:BH, BB, BI, BO, BC | 20 Hz-0V:BH, BB, BI, BO, BC (−10 dB ~ 10 dB) | - |
| No.3 | 20 Hz-0V:GH, GC, GR, GM, GS | - | 20 Hz-0V:BH, BB, BI, BO, BC |
| No.4 | 20 Hz-0V: BH, BB, BI, BO, BC | - | 20 Hz-0V: GH, GC, GR, GM, GS |

(a) Wind turbine drive train configuration

(c) Sun spline fretting corrosion

(b) HS-ST gear scuffing



(d) IMS-SH upwind bearing Assembly damage

**Fig. 5.** The NREL test wind turbine and real damage pictures.

**Table 3**
Actual Gearbox Damage Deemed Detectable through Vibration Analysis.

| Sensor | Name of Sensor | Bearing | Damage |
|---|---|---|---|
| AN5 | IS gear | LSS upwind and downwind bearing | LSS bearings defect and ISS gear defect |
| AN6 | IS pinion and HS gear | ISS | ISS bearings defect, IS pinion defect, HS gear defect |
| AN7 | HS pinion | HSS | HSS bearings defect, HS pinion defect |
| AN8 | HS pinion | HSS upwind bearing | HS pinion defect, HSS upwind bearing defect |
| AN9 | HS pinion | HSS downwind bearing | HS pinion defect, HSS downwind bearing defect |

**Table 4**
Description of NREL experiments for fault diagnosis.

| Experiment | Training (Day 1 and 2) | Testing (Day 3–10) |
|---|---|---|
| No.1 | H1, H2, H3, H4, H5, D1, D2, D3, D4, D5 | H1, H2, H3, H4, H5, D1, D2, D3, D4, D5 |
| No.2 | H1, H2, H3, H4, H5 | D1, D2, D3, D4, D5 (as OOD) |

system, and a monitoring system. The modified aero-engine retains critical components such as the low-pressure and high-pressure rotors and the inter-shaft bearing, while components like rotor blades and combustion chambers have been removed. Artificial faults, including outer and inner ring damage, were introduced into the inter-shaft bearing using precision wire-cutting methods. During testing, six sensors were installed, including four accelerometers to capture casing acceleration responses. A total of 28 sets of operating conditions, with varying LP/HP speeds, were applied.

The vibration signals collected from this aero-engine-based test rig do not exhibit clear fault characteristic frequencies in either the spectrum or envelope spectrum, unlike signals in commonly used experimental datasets. This suggests that the dataset more accurately reflects complex, real-world industrial scenarios, where fault features are often weak, nonlinear, or obscured by noise. Therefore, using this dataset to validate the proposed TID method is highly meaningful and persuasive, as it demonstrates the method's effectiveness in more challenging and realistic conditions.

To comprehensively evaluate the robustness and generalization ability of the proposed method under varying operational conditions, five experimental scenarios are designed based on different combinations of training and testing rotational speeds (rpm). Table 5 summarizes the speed configurations for each scenario. In Scenario 1, the method is evaluated across the full-speed range with added noise during testing, simulating real-world disturbances. Scenarios 2 to 5 are designed to test cross-speed generalization, where the testing speeds are partially or entirely unseen during training. These scenarios assess the diagnostic performance and reliability of the TID method under domain shift conditions.

The damage diagnosis cases in this study are conducted using a 64-bit Windows server with 64GB RAM, 12th Gen Intel CPU (i9–12,900 K) and NVIDIA RTX A5500 GPU, and by using a Pytorch library [64–66]. The baseline models developed based on Bayesian NN, are of the same architecture as those used for ConvNet [52], ResNet[67], MSDNN[19] and MSCNN[18] for comparison with the proposed Bayesian MS-ACNN to prove its reliability and effectiveness. The parameters of each unit baseline model are kept the same as those in Ref [52]. These parameters are of kernel sizes $\{32 \times 1, 16 \times 1, 8 \times 1, 3 \times 1, 2 \times 1\}$, respectively with corresponding kernel numbers {16, 16, 32, 32, 64}. A max-pooling procedure would be applied after batch normalization with the pooling size of $2 \times 1$. After the hyper parameter examination, an optimizer with a Root Mean Squared Propagation (RMSProp) and a learning rate of 0.001, decay weights of $5 \times 10^{-5}$ and the mini batch is 128 is used. The prior distribution of the parameters in each kernel is yielded by $N \sim (0, 0.05^2)$.

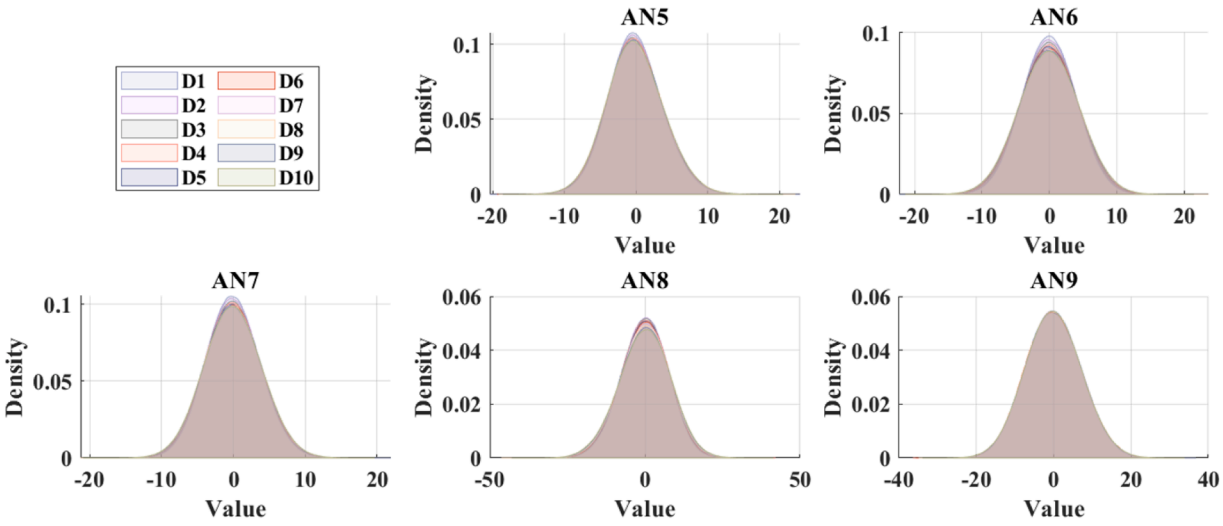
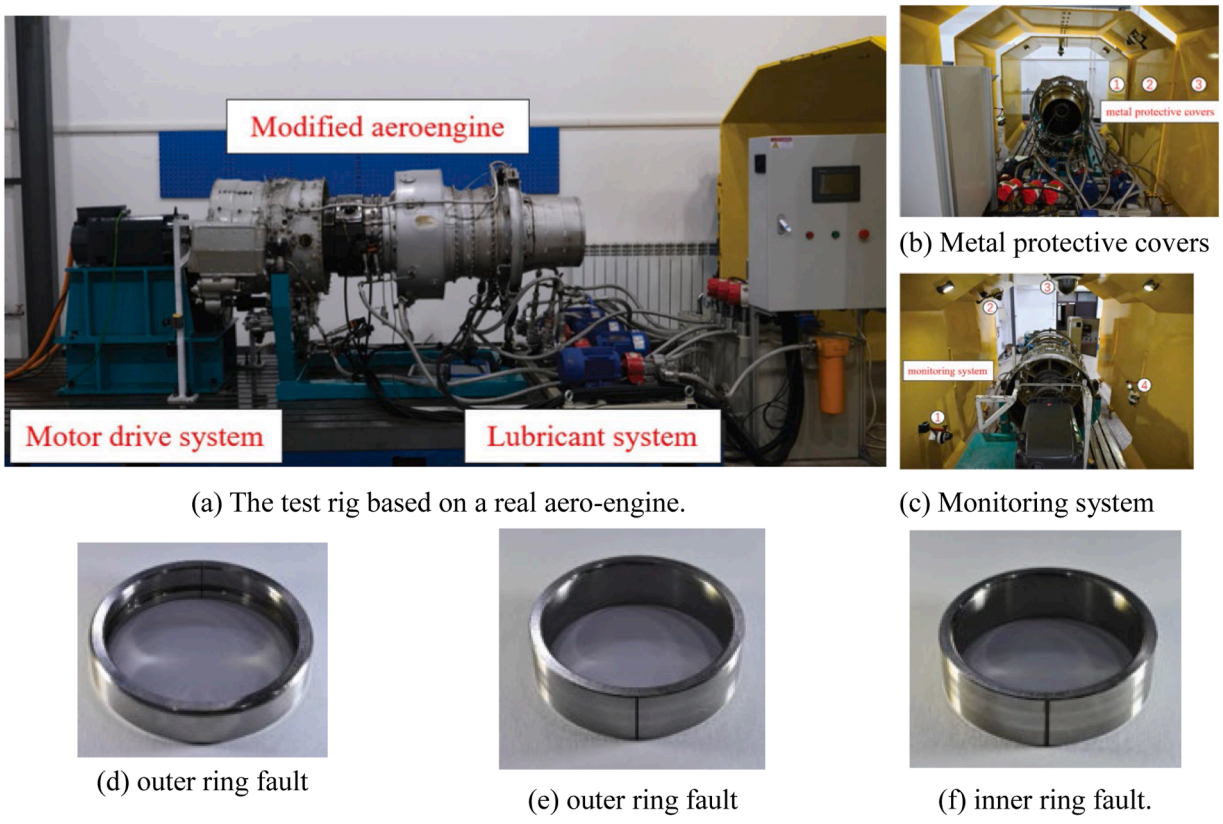



**Fig. 6.** Data distribution changes vary time.




(a) The test rig based on a real aero-engine.

(b) Metal protective covers

(c) Monitoring system

(d) outer ring fault

(e) outer ring fault

(f) inner ring fault.

**Fig. 7.** Permanent magnet synchronous motors test bench.

**Table 5**
Description of aero-engine experiments for fault diagnosis.

| Scenario | Training | Testing |
|---|---|---|
| No.1 | 1000–5000 rpm | 1000–5000 rpm + noise (0 dB) |
| No.2 | 3000,3200,3400,3600,3800,4000 rpm | 3100, 3300, 3500, 3700 rpm |
| No.3 | 2000, 3000, 4000, 5000 rpm | 2500, 3500, 4500 rpm |
| No.4 | 3000,3100,3200,3300,3400 rpm | 3500,3600,3700 rpm |
| No.5 | 3700,3600,3500,3400,3300 rpm | 3200,3100,3000 rpm |

### *4.2. Results analysis*

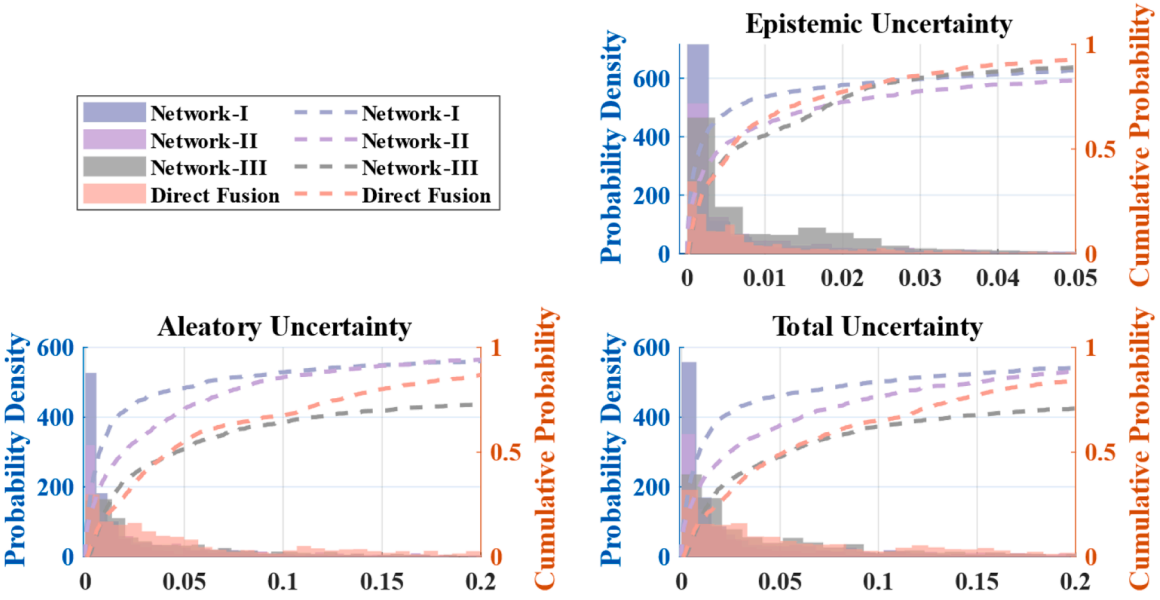
The diagnostic performance of the proposed method through a series of experiments we are analyzed and discussed in this section. Specifically, three representative case studies are conducted, including (i) gearbox failure experiments, (ii) real-world fault diagnosis of wind turbine drivetrains, and (iii) fault detection in aero-engine systems. The analysis focuses on uncertainty quantification (UQ), comparative diagnostic performance, reliability, and collaborative diagnosis enabled by UQ. The detailed results and insights are organized into the following subsections.

*4.2.1. Uncertainty quantification on diagnosis*

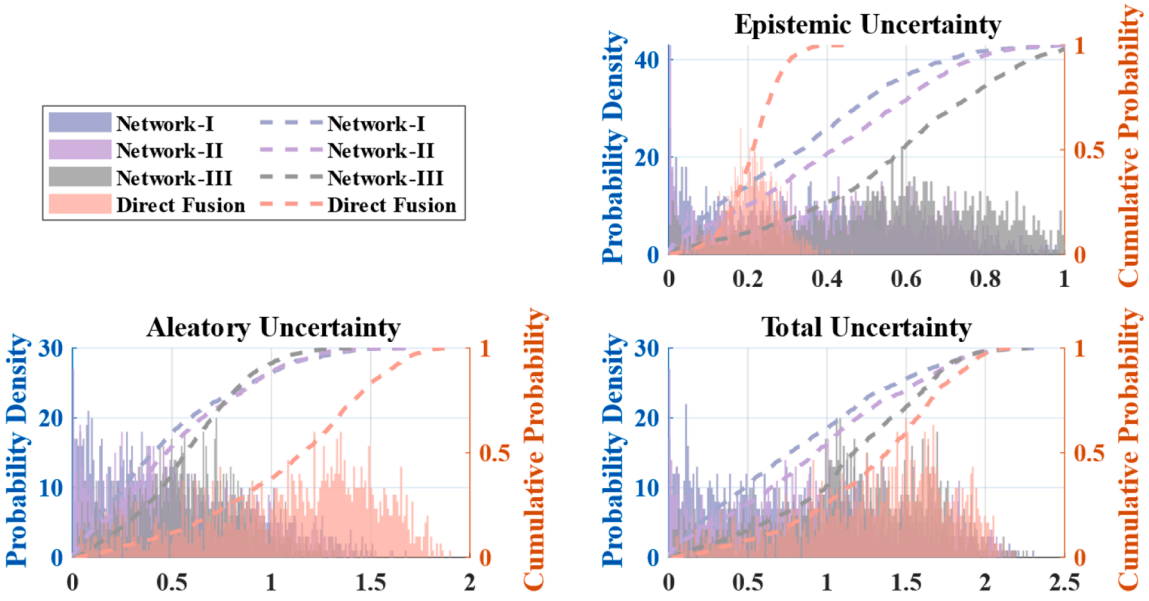
In real-world industrial applications, AI-driven fault diagnosis methods must account for uncertainty when identifying both known and unknown data. Since rotating machinery often operates under variable conditions, unknown faults and significant operational differences can undermine the reliability of AI-driven predictions. These reliability limitations should be reflected in the model's uncertainty estimation. Therefore, Fig. 8 first evaluates the proposed method's ability to quantify prediction uncertainty for both known and unknown datasets.

In Fig. 8(a), "Direct Fusion" serves as the baseline model in this study, indicating that the diagnostic results of each network are fused directly. "Network-I," "Network-II," and "Network-III" each represent a single network at different scales. In terms of epistemic uncertainty, each single network displays its own unique PDF characteristics; however, they share the same overall trend of assigning low uncertainty to most known cases.

By contrast, the PDF peak of the Direct Fusion method shifts slightly to the right and is somewhat lower, yet its tail is shorter. Meanwhile, in the domain of relatively higher uncertainty, the CPD of the Direct Fusion method rises more rapidly than that of the single networks, although Network-I exhibits the fastest rise in the lower-uncertainty region. This suggests that the Direct Fusion method is more effective in dealing with




(a) The uncertainty model for known conditions

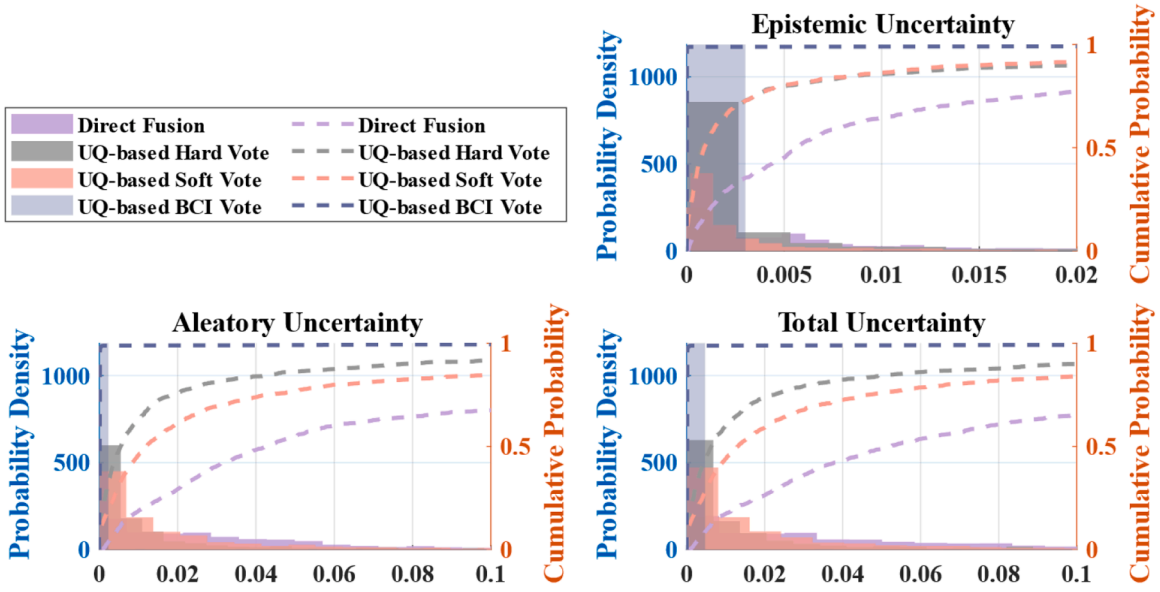



(b) The uncertainty model for unknown conditions

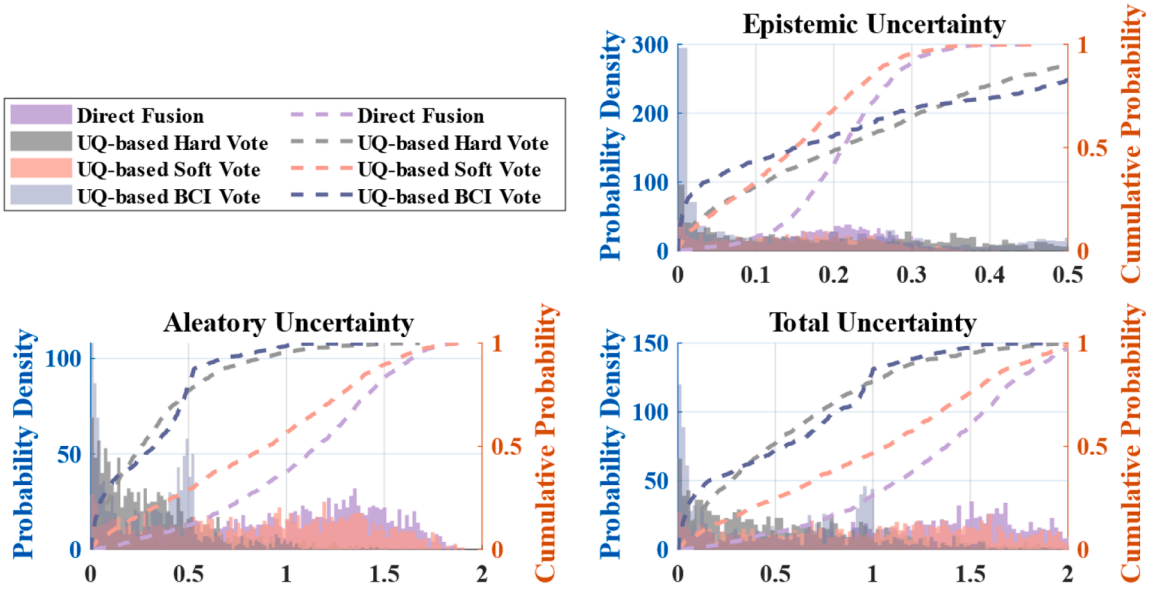
**Fig. 8.** Uncertainty quantification of different AI-driven diagnostics.

hard-to-diagnose cases compared to using a single network.

The reason lies in the fact that Direct Fusion combines information from different scales, providing complementary knowledge for diagnosis and resulting in greater reliability when handling challenging samples. In terms of aleatory uncertainty, all models exhibit a peak in the very low aleatory uncertainty region, indicating that most in-domain samples are considered to have minimal noise effects. However, differences emerge in the tail region (representing medium to high uncertainty): some single networks show a longer tail, suggesting that they struggle to handle samples with significant noise or random factors consistently. In contrast, the Direct Fusion method, by "seeing more" through multiple branches, may incorporate the differences or noise from other branches, resulting in an increased uncertainty for certain samples. Observing from CPD, if the Direct Fusion method also exhibits a relatively wide tail, it indicates that its assessment of aleatory uncertainty is higher for some samples, as it captures more "noise signals" from the different branches. Conversely, if one of the single networks shows an especially high tail, it implies that the model is completely overwhelmed by certain noisy samples, causing its aleatory uncertainty to spike. These are because of in fault diagnosis, sensor noise, operating condition fluctuations, and other factors introduce randomness into the signals. When more features are fused, the model may be better equipped to detect these differences or anomalies, which can lead to an increase in the aleatory uncertainty. On contract, A single network might overlook some noise features or




(a) The fusion methods comparation under known conditions




(b) The fusion methods comparation under unknown conditions

**Fig. 9.** The Uncertainty quantification of different decision fusion methods.

may not capture noise sufficiently across certain frequency bands or scales. As a result, it could either be overly confident or, when faced with pronounced noise, its uncertainty might spike dramatically. The total uncertainty indicates that total uncertainty is relatively reduced when multiple sub-models agree on the same sample. However, the fusion model will also be alerted and increase uncertainty when multiple sub-models have large differences on some difficult samples. For in-domain failures, since the sample pattern is relatively clear, Direct Fusion has multiple perspectives that can mostly complement each other and reduce extreme uncertainty. This is why its tail (high uncertainty part) is relatively smaller and CPD rises faster in the middle and late stages.

In Fig. 8(b), Single networks tend to exhibit a dichotomy when facing unseen distributions. In contrast, by integrating multi-scale information, Direct Fusion adopts a more cautious stance toward OOD samples, often displaying a broader or more concentrated distribution in the medium-to-high uncertainty region. This can more effectively reveal the model's doubts about unfamiliar patterns. When dealing with OOD data, an appropriate increase in uncertainty helps identify potential anomalies or unknown fault types, allowing for subsequent manual checks or online updates. Direct Fusion generally holds an advantage in OOD scenarios: they can avoid excessive confidence in single networks while also reducing extreme uncertainty, thereby demonstrating more robust diagnostic performance. Overall, in OOD situations, both individual networks and Direct Fusion show higher uncertainty levels. While fusion approaches may be more sensitive to noise, they are also more capable of recognizing unfamiliar samples and preventing overconfidence, which is beneficial for the safety and robustness of fault diagnosis.

Direct Fusion is used as the baseline for comparison with the proposed UQ-based decision-fusion methods, including both hard-vote and soft-vote approaches. The uncertainty quantification and decomposition results are shown in Fig. 8.

As shown in Fig. 9(a), for three kinds of uncertainty, the probability density function (PDF) of the UQ-based decision-level fusion methods exhibits a pronounced peak near zero uncertainty, indicating that these approaches demonstrate high confidence when processing many known samples. In contrast, the PDF for the direct fusion method shows relatively elevated uncertainty, as reflected by a more gradual increase in its cumulative distribution function (CDF), suggesting a higher prevalence of samples with substantial uncertainty. This discrepancy can be attributed to the fact that the UQ-based decision-level fusion methods maintain low uncertainty when most sub-networks are in agreement, and they only increase uncertainty when there is significant divergence among the sub-networks. Conversely, the direct fusion method does not explicitly quantify the level of disagreement among sub-networks, which can result in unresolved conflicts and consequently lead to a higher proportion of probability mass in the medium to high uncertainty range.

Analysis between hard-vote and soft-vote approaches, with regards to epistemic uncertainty, the PDFs of both soft-vote and hard-vote methods appear similar, although the CDF of the soft-vote method suggests a slightly higher level of confidence. In contrast, for aleatory uncertainty, the hard-vote method outperforms soft-vote. This is because, when the sub-networks largely agree on their inferences, the hard-vote method maintains an extremely low uncertainty level. On the other hand, the soft-vote method, by smoothly integrating the confidence scores from each sub-model, produces a more nuanced distribution across various sample types. Therefore, for ID scenarios, the hard-vote method is more reliable than the soft-vote approach.

As shown in Fig. 9(b), with regards to epistemic uncertainty direct fusion has a noticeable concentration in a lower to mid-range uncertainty region but also shows a moderate tail extending toward higher uncertainty values. On the contrary, UQ-based decision-level fusion approaches appear more spread out in the medium and high uncertainty zones, reflecting a heightened sensitivity to OOD inputs. Their peaks may be slightly lower but broader, indicating they detect more samples as unknown or less certain. So that epistemic uncertainty stems from the model's knowledge gaps. UQ-based decision-level fusion approaches explicitly account for sub-network disagreement, often pushing OOD samples toward higher uncertainty, while direct fusion merges feature without highlighting which samples sub-networks fundamentally disagree on. Besides, with regards to aleatory uncertainty direct fusion shows a distribution centered in a moderate uncertainty zone, with a tail into the higher range. On the contrary, UQ-based decision-level fusion approaches are more responsive to random noise or inherent variability in the OOD data. When sub-networks diverge significantly on noisy inputs, the UQ-based decision-level fusion approaches will elevate aleatory uncertainty more sharply. In summary, UQ-based decision-level approaches tend to push uncertainty higher for OOD problems, as they detect and amplify sub-network disagreements, which are more conservative, labeling more OOD samples as uncertain, which can be safer for fault diagnosis.

#### 4.2.2. Diagnostic results comparisons of methods

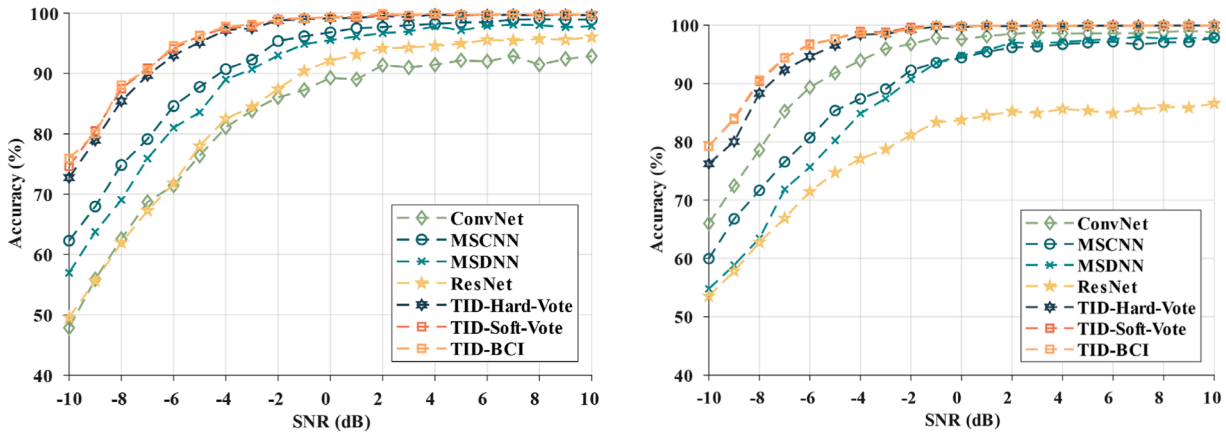
The efficiency and accuracy of an AI-driven model for fault diagnosis are significantly impacted by noise from in data from the system vibration in most practical applications. Therefore, in this section, the diagnostic capability of the proposed AI-driven diagnostic framework is compared with the state-of-the-art methods and baseline model to prove its reliability and efficiency in dealing with noisy environment. Fig. 9 respectively present the diagnostic performance (evaluated by accuracy and F1 score) under noise levels ranging from −10 dB to 10 dB. In this setup, TID-Hard-Vote employs a UQ-based hard-vote method in the decision-level fusion, TID-Soft-Vote utilizes a UQ-based soft-vote approach in the decision-level fusion, and the decision-fusion module of TID-BCI is designed based on Bayesian causal inference and uncertainty quantification.

Fig. 10 illustrates the diagnostic accuracy of different models under varying Signal-to-Noise Ratio (SNR) conditions. The comparison includes four SOTA (State-of-the-Art) methods: ConvNet, MSCNN, MSDNN, ResNet, as well as the proposed UQ-based decision fusion methods (TID-Hard-Vote, TID-Soft-Vote, TID-BCI). Under the low SNR examination, ConvNet, MSCNN, MSDNN and ResNet exhibit relatively low accuracy under low-SNR conditions, with ConvNet achieving only about 45 % accuracy at SNR = −10 dB. But TID-Hard-Vote, TID-Soft-Vote, and TID-BCI maintain relatively high accuracy (>60 %) even at SNR = −10 dB, with Soft-Vote and BCI performing the best. Examination at medium SNR, all methods show a rapid increase in accuracy, but traditional SOTA models like MSCNN and MSDNN still lag behind ResNet and UQ-based fusion methods. TID-Soft-Vote and TID-BCI consistently achieve the highest accuracy across all SNR values, demonstrating superior robustness to noise compared to SOTA methods. In summary, Traditional SOTA methods suffer from significant performance degradation at low SNR, whereas the UQ-based decision fusion methods (TID-Soft-Vote and TID-BCI) exhibit superior noise robustness. TID-Soft-Vote consistently delivers the most stable performance, maintaining higher accuracy than all other methods across the entire SNR range, particularly excelling at SNR < 0 dB. TID-BCI also performs well but is slightly outperformed by Soft-Vote in certain SNR conditions. Therefore, TID-Soft-Vote emerges as the optimal method, demonstrating the strongest noise resilience and robustness, particularly in low-SNR environments.

#### 4.2.3. Reliability analysis of UQ-based diagnostics

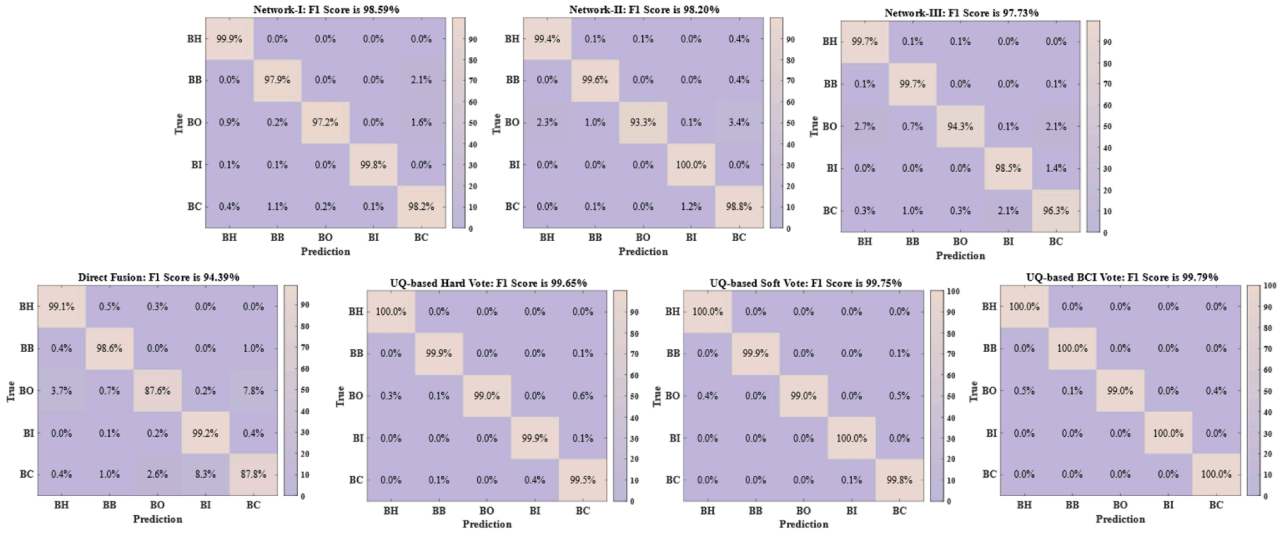
The reliability of the proposed UQ-based decision-level fusion algorithms are proven by using the confusion matrix.

Fig. 11 illustrates the mechanism of the proposed decision-level fusion algorithm, applied to bearing fault diagnosis (top) and gearbox fault diagnosis (bottom). We compare three individual networks (Network-I∼III), Direct Fusion, and three UQ-based decision fusion methods (Hard Vote, Soft Vote, BCI Vote). In case of bearing fault diagnosis, the F1 scores range from 97.79 % to 98.59 %, indicating that the multi-scale network can effectively extract features. However, some

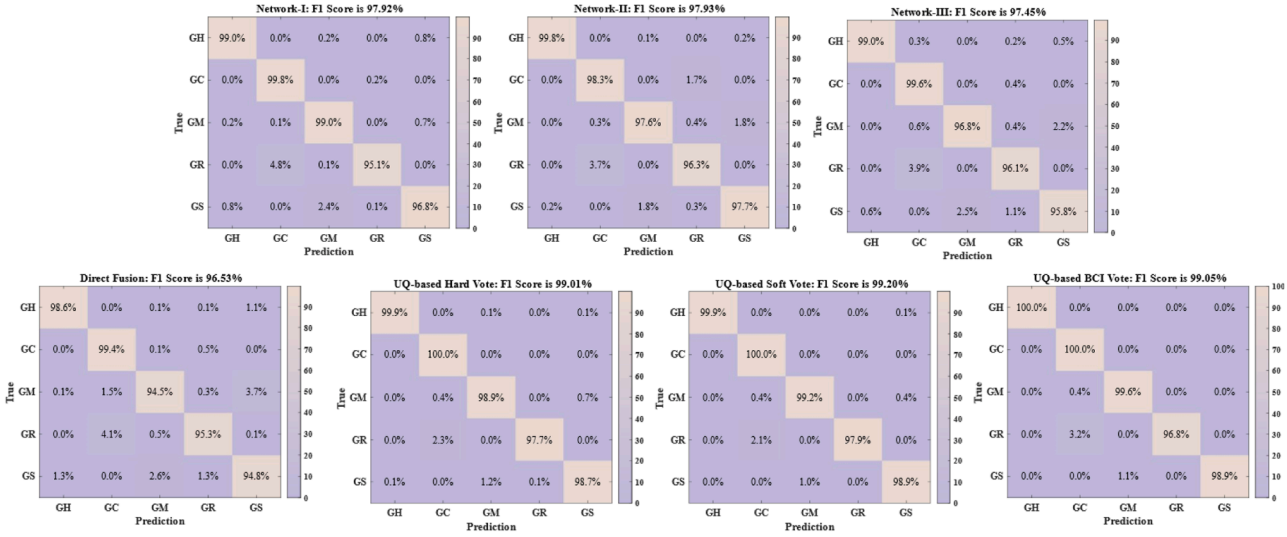



(a) Accuracy score of gear diagnosis (b) Accuracy score of bearing diagnosis

**Fig. 10.** Comparison of the models examined under experiment 1 and 2.




(a) Bearing diagnosis results




(b) Gear diagnosis results

**Fig. 11.** The mechanism of the proposed decision-level fusion algorithm.

misclassifications remain, especially in the BI and BC categories. In terms of gear failures, the F1 scores range from 97.46 % to 97.92 %, showing good performance but with some misclassification in the GR and GM categories. The Direct Fusion method achieves an F1 score lower than all individual networks (94.39 % and 96.53 %), indicating that naive fusion can introduce noise and increase misclassification. For instance, classification accuracy drops in BO and GC. Compared to Direct Fusion, F1 scores improved to 99.66 % and 99.01 %, indicating that hard voting provides more stable predictions. However, certain categories (GR and GM) still exhibit higher uncertainty. Among all methods, Soft Vote achieved the highest F1 scores (99.75 % and 99.20 %) and maintained the lowest misclassification rate across multiple categories. Notably, BI (Inner Race Crack) and BC (Inner and Outer Race Crack) classifications improved significantly, demonstrating the robustness of Soft Vote in handling uncertainty. While using BCI Vote, this method also performed well, with F1 scores close to Soft Vote (99.79 % and 99.06 %), but it still exhibited higher uncertainty in certain categories such as GM and GS. Overall, Direct Fusion performed worse than individual networks, whereas UQ-based decision fusion methods (especially Soft Vote and BCI Vote) significantly improved model stability and accuracy. Soft Vote achieved the lowest misclassification rates across multiple categories while avoiding the high uncertainty issue observed in BCI Vote. Therefore, Soft Vote emerges as the optimal fusion strategy, maintaining high accuracy while reducing misclassification risks in specific fault categories.

In this study, we introduce the Trustworthy Threshold, which represents the uncertainty value at the 99 % CDF of epistemic uncertainty. This metric serves as a key indicator for identifying unreliable predictions. When the inference uncertainty exceeds this threshold, we recommend performing a double-check to enhance diagnostic reliability. As shown in Table 6, the single networks (Network-I∼III) and the Direct Fusion method exhibit significantly higher Trustworthy Threshold values, indicating a higher degree of epistemic uncertainty in certain fault categories. For instance, Network-I shows a threshold as high as 0.409 for GH and 0.415 for GM, suggesting that these methods produce more uncertain predictions in these categories. In contrast, the UQ-based decision fusion methods (TID-Hard-Vote, TID-Soft-Vote, TID-BCI) significantly reduce the Trustworthy Threshold, particularly Soft-Vote and BCI, which show notable reductions in epistemic uncertainty across most fault categories. More specifically, TID-BCI achieves the lowest Trustworthy Threshold in several categories (e.g., GC=0.000, GH=0.029, BH=0.009), indicating highly stable decision-making with minimal inference uncertainty in these cases. However, TID-BCI still exhibits high uncertainty in GM=0.629 and GS=0.421, suggesting that additional verification may be required for these categories. Comparatively, TID-Soft-Vote achieves significantly lower uncertainty in GM=0.102 and GS=0.126 than BCI, demonstrating greater decision stability. Overall, the Direct Fusion method shows relatively high Trustworthy Threshold values, indicating less stable inference, whereas UQ-based methods (TID-Hard-Vote, TID-Soft-Vote, TID-BCI) effectively reduce uncertainty. Among them, TID-Soft-Vote maintains low uncertainty while avoiding the high uncertainty observed in certain categories with BCI, making it the most stable and reliable fusion strategy. Thus, in the following sections on UQ-based collaborative diagnosis analysis, we will compare the performances of Hard-Vote and Soft-Vote exclusively with the baseline Direct Fusion.

#### 4.2.4. UQ-based collaborative diagnosis analysis

The OOD diagnosis can be identified by uncertainty quantification. The inferences of AI-driven model to diagnose that is out of the trustworthy threshold is identified as unknown condition. Fig. 12.

Fig. 12 indicates that the overall F1 score is only 57.12 % from baseline direct fusion method while an intelligent model without UQ-based reidentification. However, when incorporating UQ, the F1 score improves to 84.20 %, suggesting that identifying OOD samples improves overall classification performance. Both UQ-based Hard Vote and Soft Vote significantly outperform Direct Fusion, achieving F1 scores of 93.02 % and 92.78 %, respectively. Hard Vote is more decisive in rejecting OOD samples, as indicated by the clearer separation of OOD from in-distribution classes. This suggests that the Hard Vote is more conservative in uncertain cases, making strong classification decisions. The incorporation of UQ-based decision fusion significantly enhances classification accuracy, particularly for OOD cases. Compared to the baseline Direct Fusion, both methods show remarkable improvements, highlighting the necessity of UQ for reliable diagnostics in complex scenarios.

The diagnosis with high epistemic uncertainty has been identified as OOD results, that needs another diagnostic with extra knowledge to help into re-diagnose them. This, in the next step, the TID that is able to diagnose gear health conditions is called to deal with the OOD samples. Following the collaborative diagnostic framework, the diagnosis result is shown in Fig. 13.

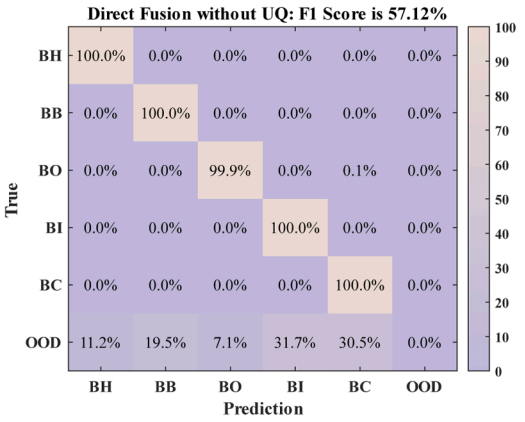
As shown in the above analysis, the OOD identification accuracy exceeds 90 %, meaning that there are only 1000 OOD samples. Manually screening these 1000 OOD samples would still be time-consuming and labor-intensive. Therefore, collaborative fault diagnosis is further performed on the OOD data. Fig. 13 presents the results of a collaborative diagnosis approach, where OOD (out-of-distribution) samples identified in the first stage are re-diagnosed using a second model with additional knowledge reserves. This allows the transition from a bearing-focused fault diagnosis to a combined bearing and gear fault diagnosis system without training extra models. Direct Fusion leaves 308 OOD samples, while UQ-based Hard Vote reduces it to 103, and Soft Vote further lowers it to 114. This suggests that the collaborative method effectively assigns previously unclassified OOD samples to meaningful fault categories, reducing human verification effort. UQ-based Hard and Soft Voting successfully reclassify OOD samples, leading to a more comprehensive diagnosis system, although Direct Fusion achieves a higher F1 score, it fails to address the OOD problem effectively. Overall, Soft Vote provides a balance between making confident classifications and leaving some OOD samples for further inspection.

#### 4.2.5. Fault diagnosis of real wind turbine gearbox

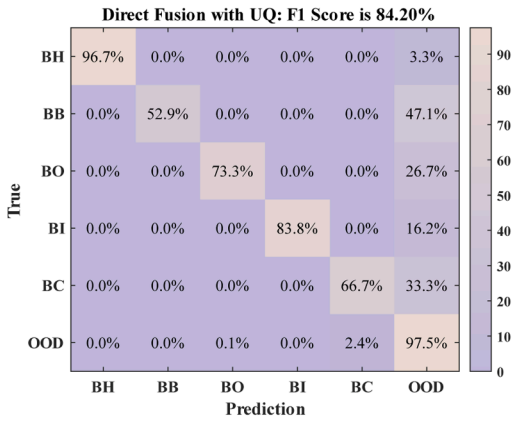
A real-world wind turbine gearbox failure dataset is used to validate the effectiveness and reliability of the proposed collaborative diagnostic framework. Fig. 13 presents the overall performance under both ID and OOD scenarios. In each evaluation, varying conditions, including noise levels and rotational speeds, are considered throughout different times of the day, further demonstrating the robustness of the AI-driven models in real-world applications.

**Table 6**
The average diagnostics performance across from Scenario II-A to I-D.

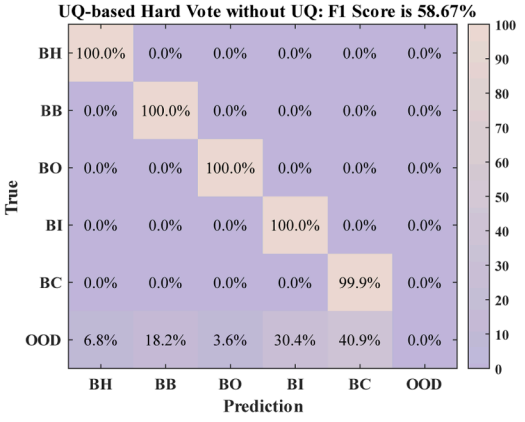
| Model/Trustworthy Threshold | GH | GC | GR | GM | GS | BH | BB | BI | BO | BC |
|---|---|---|---|---|---|---|---|---|---|---|
| Network-I | 0.409 | 0.174 | 0.320 | 0.415 | 0.411 | 0.229 | 0.347 | 0.464 | 0.275 | 0.402 |
| Network-II | 0.227 | 0.358 | 0.428 | 0.469 | 0.405 | 0.198 | 0.322 | 0.542 | 0.118 | 0.404 |
| Network-III | 0.402 | 0.217 | 0.480 | 0.326 | 0.460 | 0.211 | 0.222 | 0.466 | 0.450 | 0.371 |
| Direct Fusion | 0.402 | 0.226 | 0.404 | 0.335 | 0.436 | 0.355 | 0.329 | 0.575 | 0.353 | 0.488 |
| TID-Hard-Vote | 0.188 | 0.081 | 0.316 | 0.250 | 0.238 | 0.061 | 0.098 | 0.279 | 0.040 | 0.266 |
| TID-Soft-Vote | 0.109 | 0.061 | 0.117 | 0.102 | 0.126 | 0.059 | 0.094 | 0.145 | 0.047 | 0.148 |
| TID-BCI | 0.029 | 0.000 | 0.144 | 0.629 | 0.421 | 0.009 | 0.004 | 0.363 | 0.000 | 0.094 |



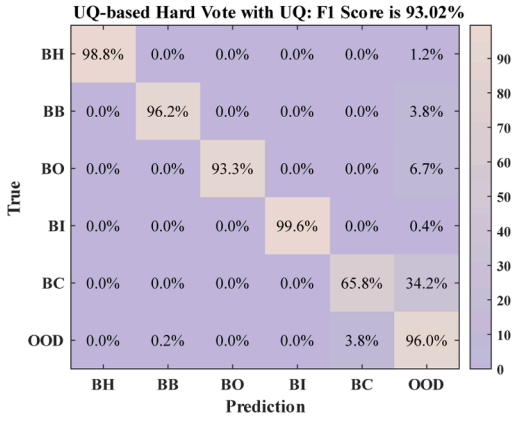

(a) OOD detection without reidentification via trustworthiness examination based on Direct Fusion



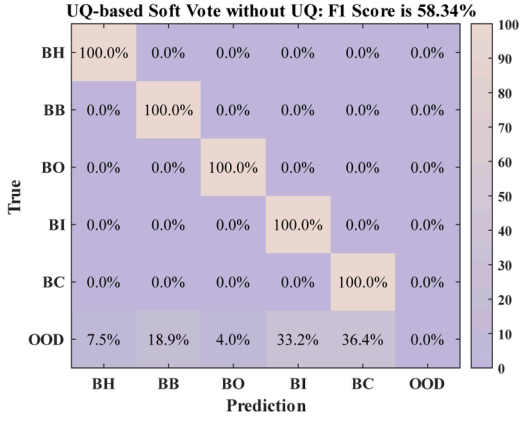

(a) OOD detection with reidentification via trustworthiness examination based on Direct Fusion



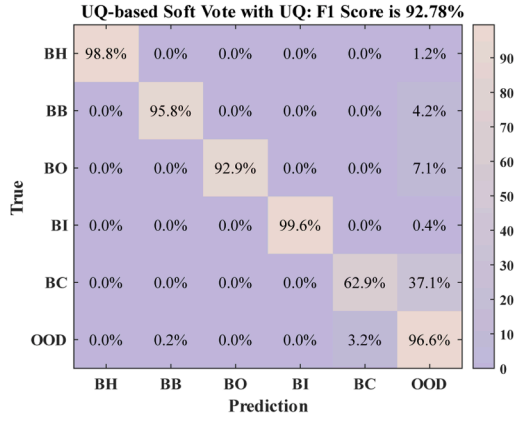

(a) OOD detection without reidentification via trustworthiness examination based on Hard-Vote




(a) OOD detection with reidentification via trustworthiness examination based on Hard-Vote




(a) OOD detection without reidentification via trustworthiness examination based on Soft-Vote




(a) OOD detection with reidentification via trustworthiness examination based on Soft-Vote

**Fig. 12.** Multiple models cooperative diagnosis approach.

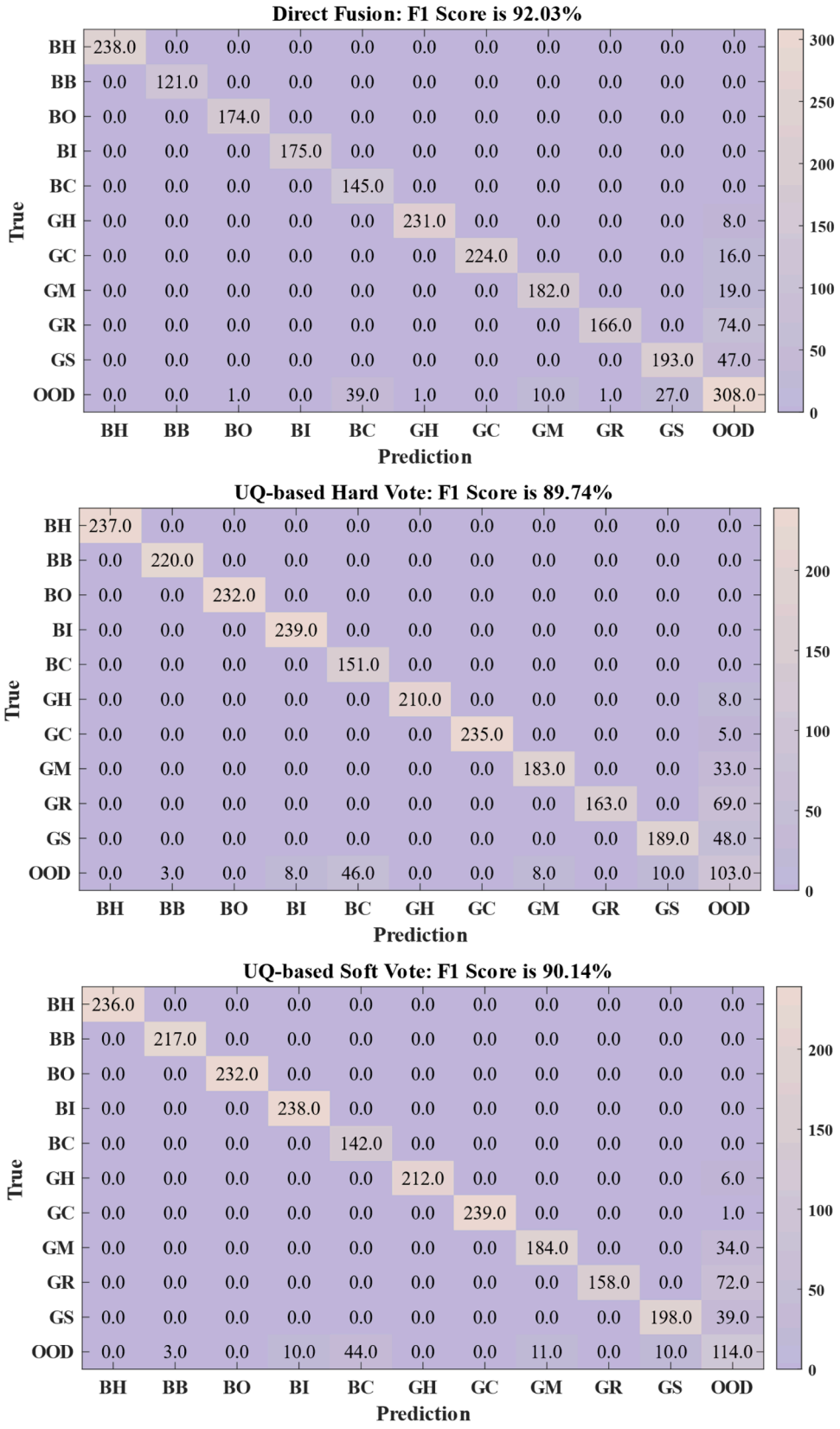
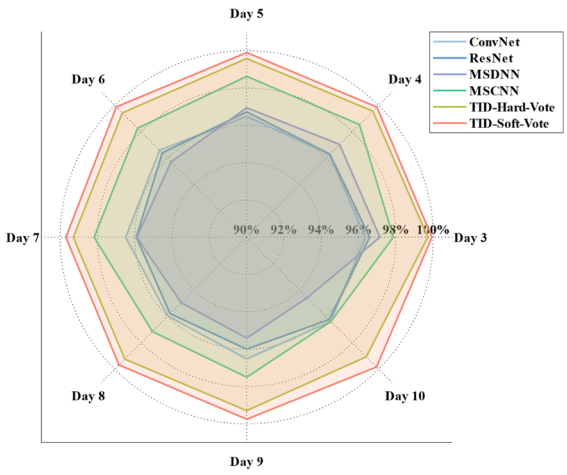
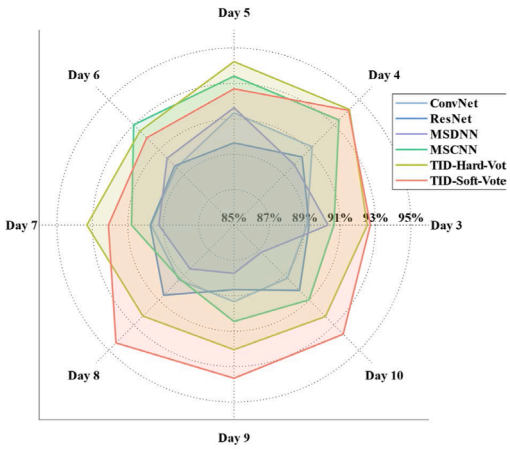



**Fig. 13.** . Collaborative diagnosis results.

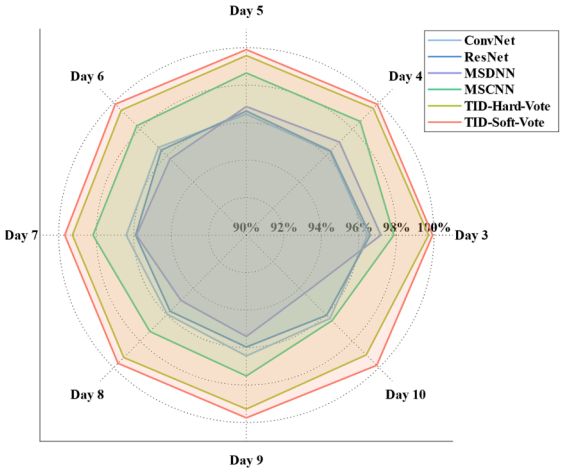
From Fig. 14, it can be observed that the two decision fusion approaches of the TIC method (Soft and Hard) demonstrate more stable performance across all metrics (Accuracy, Recall, Precision) compared to traditional methods such as ConvNet, ResNet, MSDNN, and MSCNN. This stability is particularly prominent in the ID task. In the ID task, as time progresses (from Day 3to Day 10), the performance of all models experiences a slight decline, which aligns with expectations, as the uncertainty in data distribution increases over time. However, the decline



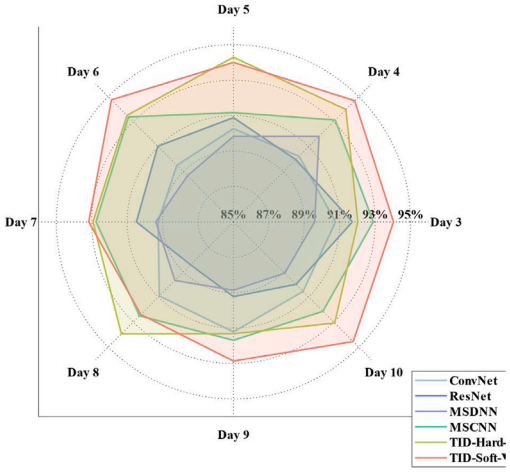

(a) Accuracy on ID task



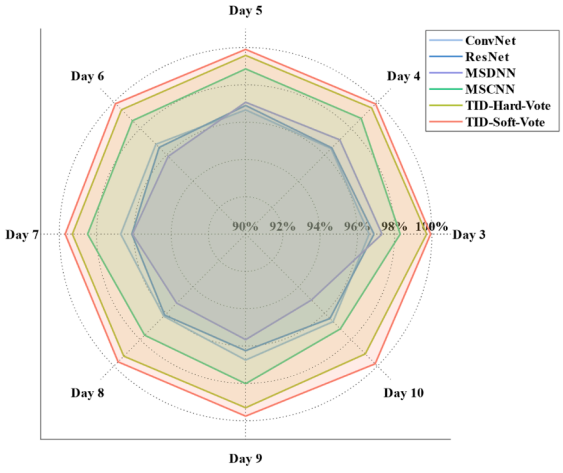

(d) Accuracy on OOD task



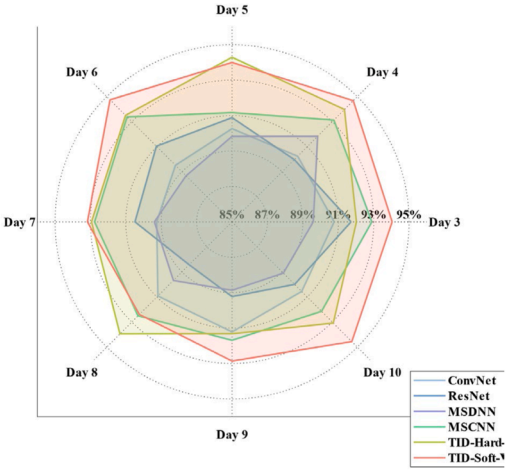

(b) Recall on ID task




(e) Recall on OOD task




(c) Precision on ID task




(f) Precision on OOD task

**Fig. 14.** Diagnostic performance under ID and OOD scenario.

trend of the TIC method is more gradual, indicating its strong generalization ability. In contrast, the performance of the TIC method in the OOD task remains limited, especially due to the chaotic nature of the temporal evolution of data distribution. This suggests that although uncertainty quantification (UQ) techniques help enhance the trustworthiness of the model in unseen distributions, they still cannot fully resolve challenges posed by data shift and distributional uncertainty. The Soft voting mechanism performs relatively better in the OOD task but still falls short of the performance achieved in the ID task. This indicates the need for further optimization of the decision-making mechanism in OOD scenarios. In summary, the TIC method demonstrates high reliability in the ID task, and the soft voting mechanism contributes to improved diagnostic performance. However, in the OOD task, the weak temporal dependency in data distribution leads to greater uncertainty and increased performance fluctuations. While the TIC method offers advantages in the OOD task, it is still insufficient to completely overcome the challenges posed by data distribution shift.

#### 4.2.6. *Fault diagnosis for drivetrains of aeroengine*

To evaluate the robustness and generalization capability of the proposed method, five experimental scenarios (No. 1 to No 5) were constructed using the aero-engine dataset, each involving different configurations of training and testing speeds (as shown in Table 5). The violin plot in Fig. 15 illustrates the distribution of diagnostic accuracy for various models under these scenarios.

In Scenario No 1, where both training and testing data cover the full-speed range (1000–5000 rpm) with added noise (0 dB), the proposed TID-Soft Vote method maintains high accuracy and low variance, demonstrating excellent robustness to noise-induced aleatoric uncertainty. In contrast, baseline models such as ConvNet exhibit high variability. While some outlier cases show high accuracy, the wide spread indicates a lack of consistency and reliability. This highlights that performance peaks alone cannot ensure trustworthiness, especially under uncertainty.

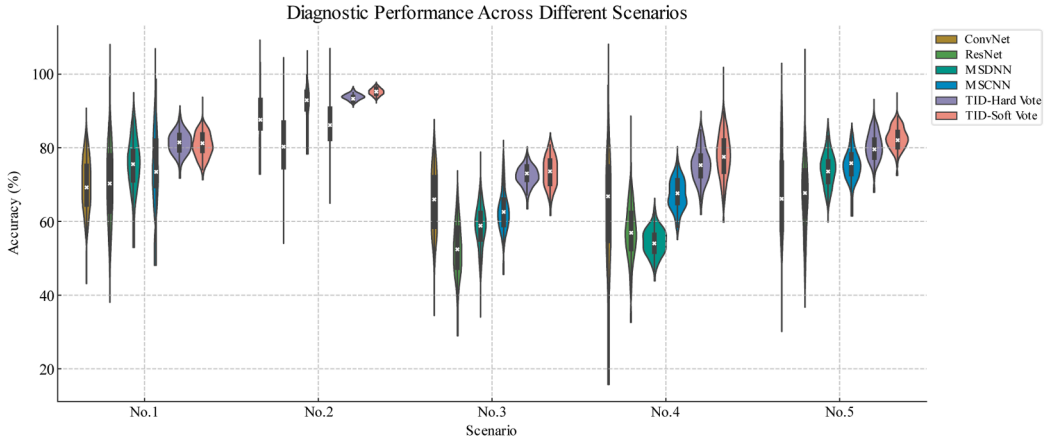
Scenario No 2 introduces a more challenging generalization test by training on discrete speeds (3000–4000 rpm) and testing on unseen intermediate speeds (e.g., 3100, 3300, 3500, 3700 rpm). Most baseline models suffer significant performance degradation due to the domain shift. However, the TID-Soft Vote method consistently delivers higher accuracy with narrower uncertainty bands, thanks to its UQ-based decision fusion module. This reinforces the model's ability to provide confident and reliable predictions even under unseen conditions.

Scenarios No 3 to No 5 are designed to evaluate performance under partial-speed domain shifts. In Scenario No 3, although the training speeds include low, middle, and high rpm (2000, 3000, 4000, 5000), the test speeds (2500, 3500, 4500) are interpolated values not seen during training. Scenarios No 4 and No 5 are even more stringent, with entirely different test speed sets compared to training. The proposed method achieves the highest mean accuracy in all five scenarios while also showing reduced performance variance, confirming its superior reliability and robustness across diverse and uncertain operating conditions.

## 5. Conclusions

This study developed a TID model, which integrates a multi-scale probabilistic neural network with UQ-based decision fusion modules. Three UQ-based decision fusion strategies were designed: hard voting, soft voting, and Bayesian causal inference. The proposed TID model demonstrates higher confidence (lower epistemic uncertainty) in ID fault identification while exhibiting a broader and more dispersed uncertainty distribution in OOD scenarios, signaling the need for additional verification of OOD diagnostic results. Furthermore, leveraging this characteristic, a Collaborative Diagnostic Framework was designed. This framework uses UQ to identify OOD diagnostics and then incorporates additional TID models with supplementary knowledge to re-evaluate these uncertain cases iteratively. The process continues until no further TID models are available. This approach significantly reduces the cost of training large models and minimizes the human effort required to manually inspect a vast number of OOD cases.

The accuracy and reliability of the proposed method were validated using two datasets from prototype and real-world wind turbine models. The gearbox experimental dataset, including both gear and bearing faults, was used to assess the effectiveness of the proposed approach. The superiority of the TID model was demonstrated through comparisons with state-of-the-art multi-scale framework models under diverse operating conditions, particularly in high-noise environments. The results indicate that the proposed TID model can accurately diagnose




| | | | | | |
|---|---|---|---|---|---|
| ConvNet | 69.88±7.32 | 89.52±7.03 | 64.33±8.24 | 64.69±13.58 | 66.64±13.47 |
| ResNet | 70.98±11.23 | 81.03±7.07 | 52.79±6.81 | 57.45±7.81 | 67.57±10.62 |
| MSDNN | 75.17±5.89 | 93.36±4.14 | 57.95±6.24 | 53.93±**3.97** | 73.57±4.28 |
| MSCNN | 73.76±8.57 | 85.59±6.72 | 63.42±5.44 | 67.81±4.45 | 75.72±4.15 |
| TID-Hard Vote | **81.47**±**3.03** | 93.23±**0.69** | 73.25±**3.06** | 73.90±4.18 | 80.23±3.64 |
| TID-Soft Vote | 81.14±3.05 | **95.13**±1.06 | **73.26**±4.41 | **77.03**±5.91 | **81.93**±**3.12** |

**Fig. 15.** illustrates the diagnostic performance of various models under the aero-engine testing scenarios.

faults with lower uncertainty. The aleatoric and epistemic uncertainties of the TID model were analysed to evaluate its performance on known and unknown datasets. The findings highlight that UQ-based fusion methods exhibit broader and higher uncertainty distributions in OOD scenarios, enabling a more cautious approach to data outside the training distribution. Compared to Direct Fusion methods, which may display moderate uncertainty across more samples but lack a mechanism to identify significant sub-network disagreements, the TID model effectively mitigates misestimation in challenging and noisy OOD cases.

## 6. Future works

AI-driven fault diagnosis, RUL prediction, and other PHM tasks hold significant potential to enhance the reliability and safety of industrial equipment. However, it is essential to ensure the reliability of AI-driven methods when performing these tasks. Building upon the foundations established in this study, future research could focus on investigating the propagation of uncertainty within AI models [68], as well as integrating decision-level fusion strategies. These directions aim to further improve the reliability of AI-based diagnostics and to ensure safer applications in industrial environments.

## CRediT authorship contribution statement

**Zifei Xu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Kaicheng Zhao:** Visualization, Software, Data curation. **Wanfu Zhang:** Visualization, Validation, Software, Formal analysis. **Weipao Miao:** Visualization, Software, Resources. **Kang Sun:** Validation, Software, Data curation. **Jin Wang:** Writing – review & editing, Supervision. **Musa Bashir:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

## Declaration of competing interest

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

## Acknowledgments

This research is part of the ULTIMATE project (UKRI/EPSRC: EP/Y014235/2) and has received financial support from UKRI Innovate UK (grant numbers TS/Y006364/1, TS/Y005236/1, and TS/X018407/1). Additional funding was provided by the State Key Laboratory of Mechanical System and Vibration under Grant MSV202411 and by the Shanghai Natural Science Foundation (General Program) under Grant 24ZR1454800. The authors would like to acknowledge the National Renewable Energy Laboratory and the United States Department of Energy for providing the wind turbine gearbox vibration condition monitoring benchmark datasets.

## Data availability

Data will be made available on request.

## References

[1] P. Vrignat, F. Kratz, M. Avila, Sustainable manufacturing, maintenance policies, prognostics and health management: a literature review, Reliab. Eng. Syst. Saf. 218 (2022), https://doi.org/10.1016/j.ress.2021.108140.
[2] F. Folino, G. Folino, M. Guarascio, F.S. Pisani, L. Pontieri, On learning effective ensembles of deep neural networks for intrusion detection, Inform. Fusion 72 (2021) 48–69, https://doi.org/10.1016/j.inffus.2021.02.007.
[3] Wang S., Li Y., Wang D., Zhang W., Chen X., Dong D., et al. Echo state graph neural networks with analogue random resistor arrays 2021. https://doi.org/10.1038/s42256-023-00609-5.
[4] S.H. Singh, F. van Breugel, R.P.N. Rao, B.W. Brunton, Emergent behaviour and neural dynamics in artificial agents tracking odour plumes, Nat. Mach. Intell. 5 (2023) 58–70, https://doi.org/10.1038/s42256-022-00599-w.
[5] E. Zio, Prognostics and health management (PHM): where are we and where do we (need to) go in theory and practice, Reliab. Eng. Syst. Saf. 218 (2022) 108119, https://doi.org/10.1016/j.ress.2021.108119.
[6] Z. Xu, K. Zhao, J. Wang, M. Bashir, Physics-informed probabilistic deep network with interpretable mechanism for trustworthy mechanical fault diagnosis, Adv. Eng. Inform. 62 (2024), https://doi.org/10.1016/j.aei.2024.102806.
[7] Y. Wang, M. Ji, S. Jiang, X. Wang, J. Wu, F. Duan, et al., Augmenting vascular disease diagnosis by vasculature-aware unsupervised learning, Nat. Mach. Intell. 2 (2020) 337–346, https://doi.org/10.1038/s42256-020-0188-z.
[8] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, A.K. Nandi, Applications of machine learning to machine fault diagnosis: a review and roadmap, Mech. Syst. Signal. Process. 138 (2020) 106587, https://doi.org/10.1016/j.ymssp.2019.106587.
[9] A.M. Alaa, D. Gurdasani, A.L. Harris, J. Rashbass, M. van der Schaar, Machine learning to guide the use of adjuvant therapies for breast cancer, Nat. Mach. Intell. 3 (2021) 716–726, https://doi.org/10.1038/s42256-021-00353-8.
[10] W. Mao, W. Feng, Y. Liu, D. Zhang, X. Liang, A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis, Mech. Syst. Signal. Process. 150 (2021) 107233, https://doi.org/10.1016/j.ymssp.2020.107233.
[11] B. Zhao, X. Zhang, H. Li, Z. Yang, Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions, Knowl. Based. Syst. 199 (2020) 105971, https://doi.org/10.1016/j.knosys.2020.105971.
[12] Y. Ding, M. Jia, Q. Miao, Y. Cao, A novel time–frequency transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings, Mech. Syst. Signal. Process. 168 (2022), https://doi.org/10.1016/j.ymssp.2021.108616.
[13] J. Jiao, M. Zhao, J. Lin, K. Liang, A comprehensive review on convolutional neural network in machine fault diagnosis, Neurocomputing. 417 (2020) 36–63, https://doi.org/10.1016/j.neucom.2020.07.088.
[14] H. Shao, J. Lin, L. Zhang, D. Galar, U. Kumar, A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance, Inform. Fusion 74 (2021) 65–76, https://doi.org/10.1016/j.inffus.2021.03.008.
[15] K. Yang, B. Hu, R. Malekian, Z. Li, An improved control-limit-based principal component analysis method for condition monitoring of marine turbine generators, J. Marine Eng. Technol. 19 (2020) 249–256, https://doi.org/10.1080/20464177.2019.1655135.
[16] Z. Xu, C. Li, Y. Yang, Fault diagnosis of rolling bearing of wind turbines based on the variational mode decomposition and deep convolutional Neural networks, Appl. Soft Comput. J. 95 (2020) 106515, https://doi.org/10.1016/j.asoc.2020.106515.
[17] W. Huang, J. Cheng, Y. Yang, G. Guo, An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis, Neurocomputing. 359 (2019) 77–92, https://doi.org/10.1016/j.neucom.2019.05.052.
[18] G. Jiang, H. He, J. Yan, P. Xie, Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox, IEEE Transac. Industr. Electr. 66 (2019) 3196–3207, https://doi.org/10.1109/TIE.2018.2844805.
[19] Z. Xu, C. Li, Y. Yang, Fault diagnosis of rolling bearings using an improved Multi-Scale convolutional neural network with feature attention mechanism, ISA Trans. 110 (2021) 379–393, https://doi.org/10.1016/j.isatra.2020.10.054.
[20] Z. Xu, M. Bashir, W. Zhang, Y. Yang, X. Wang, C. Li, An intelligent fault diagnosis for machine maintenance using weighted soft-voting rule based Multi-attention module with Multi-Scale information Fusion1, Inform. Fusion 86–87 (2022) 17–29, https://doi.org/10.1016/j.inffus.2022.06.005.
[21] Bashir M., Xu Z., Wang J. Data-driven damage quantification of floating offshore wind turbine platforms based on multi-scale encoder – Decoder with self-attention mechanism 2022.
[22] W. Li, R. Huang, J. Li, Y. Liao, Z. Chen, G. He, et al., A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: theories, applications and challenges, Mech. Syst. Signal. Process. 167 (2022) 108487, https://doi.org/10.1016/j.ymssp.2021.108487.
[23] R. Zhang, H. Tao, L. Wu, Y. Guan, Transfer learning with neural networks for bearing fault diagnosis in changing working conditions, IEEe Access. 5 (2017) 14347–14357, https://doi.org/10.1109/ACCESS.2017.2720965.
[24] Z. Wang, X. He, B. Yang, N. Li, Subdomain adaptation transfer learning network for fault diagnosis of roller bearings, IEEE Transac. Industr. Electr. 69 (2022) 8430–8439, https://doi.org/10.1109/TIE.2021.3108726.
[25] X. Cao, B. Chen, N. Zeng, A deep domain adaption model with multi-task networks for planetary gearbox fault diagnosis, Neurocomputing. 409 (2020) 173–190, https://doi.org/10.1016/j.neucom.2020.05.064.
[26] Y. Song, Y. Li, L. Jia, M. Qiu, Retraining strategy-based domain adaption Network for intelligent fault diagnosis, IEEe Trans. Industr. Inform. 16 (2020) 6163–6171, https://doi.org/10.1109/TII.2019.2950667.
[27] L. Meng, M. Zhao, Z. Cui, X. Zhang, S. Zhong, Empirical mode reconstruction: preserving intrinsic components in data augmentation for intelligent fault diagnosis of civil aviation hydraulic pumps, Comput. Ind. 134 (2022) 103557, https://doi.org/10.1016/j.compind.2021.103557.
[28] A. Ragab, H. Ghezzaz, M. Amazouz, Decision fusion for reliable fault classification in energy-intensive process industries, Comput. Ind. 138 (2022) 103640, https://doi.org/10.1016/j.compind.2022.103640.

[29] Lee H.B., Lee H., Na D., Kim S., Park M., Yang E., et al. Learning to balance: bayesian meta-Learning for imbalanced and out-of-distribution tasks 2019:1–15.
[30] Z. Wang, Z. Xu, C. Cai, X. Wang, J. Xu, K. Shi, et al., Rolling bearing fault diagnosis method using time-frequency information integration and multi-scale TransFusion network, Knowl. Based. Syst. 284 (2024), https://doi.org/10.1016/j.knosys.2023.111344.
[31] H. Bhatt, N.K. Jadav, A. Kumari, R. Gupta, S. Tanwar, Z. Polkowski, et al., Artificial neural network-driven federated learning for heart stroke prediction in healthcare 4.0 underlying 5G, Concurr. Comput. 36 (2024), https://doi.org/10.1002/cpe.7911.
[32] K. Feng, J.C. Ji, Q. Ni, M. Beer, A review of vibration-based gear wear monitoring and prediction techniques, Mech. Syst. Signal. Process. 182 (2023), https://doi.org/10.1016/j.ymssp.2022.109605.
[33] Q. Ni, J. Ji, K. Feng, Data-driven prognostic scheme for bearings based on a novel health indicator and gated recurrent unit network, IEEe Trans. Industr. Inform. 3203 (2022), https://doi.org/10.1109/TII.2022.3169465.
[34] Z. Wang, X. Jiang, Z. Xu, C. Cai, X. Wang, J. Xu, et al., Early anomaly detection of wind turbine gearbox based on SLFormer neural network, Ocean Eng. 311 (2024), https://doi.org/10.1016/j.oceaneng.2024.118925.
[35] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, et al., A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inform. Fusion 76 (2021) 243–297, https://doi.org/10.1016/j.inffus.2021.05.008.
[36] R. Rajesh, S. Hemalatha, S.M. Nagarajan, G.G. Devarajan, M. Omar, A.K. Bashir, Threat detection and mitigation for tactile internet driven consumer IoT-Healthcare system, IEEE Transac. Consumer Electr. 70 (2024) 4249–4257, https://doi.org/10.1109/TCE.2024.3370193.
[37] S.M. Nagarajan, G.G. Devarajan, T.V. R, M. AJ, A.K. Bashir, Y.D. Al-Otaibi, Adversarial Deep learning based Dampster–Shafer data fusion model for intelligent transportation system, Information Fusion 102 (2024), https://doi.org/10.1016/j.inffus.2023.102050.
[38] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, 110, Springer US, 2021, https://doi.org/10.1007/s10994-021-05946-3.
[39] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, I. Kompatsiaris, Thesis_ Uncertainty in deep learning, Online Informat. Rev. 43 (2019) 72–88.
[40] G.G. Devarajan, U. Kumaran, G. Chandran, R.P. Mahapatra, A. Alkhayyat, Next-generation imaging methodology: an intelligent transportation system for consumer industry, IEEE Transac. Consumer Electronics 70 (2024) 3680–3687, https://doi.org/10.1109/TCE.2024.3372906.
[41] S.M. Nagarajan, G.G. Devarajan, M. ST, V. RT, A.K. Bashir, A.A. AlZubi, Artificial intelligence based zero trust security approach for consumer industry, IEEE Transac. Consumer Electr. (2024), https://doi.org/10.1109/TCE.2024.3412772.
[42] G.G. Devarajan, S.M. Nagarajan, A. Daniel, T. Vignesh, R. Kaluri, Consumer product recommendation system using adapted PSO with federated learning method, IEEE Transact. Consumer Electr. 70 (2024) 2708–2715, https://doi.org/10.1109/TCE.2023.3319374.
[43] R. Zhu, W. Peng, D. Wang, C.G. Huang, Bayesian transfer learning with active querying for intelligent cross-machine fault prognosis under limited data, Mech. Syst. Signal. Process. 183 (2022) 109628, https://doi.org/10.1016/j.ymssp.2022.109628.
[44] H. Li, T. Wang, G. Wu, A Bayesian deep learning approach for random vibration analysis of bridges subjected to vehicle dynamic interaction, Mech. Syst. Signal. Process. 170 (2022) 108799, https://doi.org/10.1016/j.ymssp.2021.108799.
[45] G. Chen, M. Liu, J. Chen, Frequency-temporal-logic-based bearing fault diagnosis and fault interpretation using bayesian optimization with bayesian neural networks, Mech. Syst. Signal. Process. 145 (2020) 106951, https://doi.org/10.1016/j.ymssp.2020.106951.
[46] Z. Xu, M. Bashir, Q. Liu, Z. Miao, X. Wang, J. Wang, et al., A novel health indicator for intelligent prediction of rolling bearing remaining useful life based on unsupervised learning model, Comput. Ind. Eng. 176 (2023) 108999, https://doi.org/10.1016/j.cie.2023.108999.
[47] M. Soualhi, K.T.P. Nguyen, K. Medjaher, F. Nejjari, V. Puig, J. Blesa, et al., Dealing with prognostics uncertainties: combination of direct and recursive remaining useful life estimations, Comput. Ind. 144 (2023) 103766, https://doi.org/10.1016/j.compind.2022.103766.
[48] Y. Lin, M. Xiao, H. Liu, Z. Li, S. Zhou, X. Xu, et al., Gear fault diagnosis based on CS-improved variational mode decomposition and probabilistic neural network, Measurement. (Lond) 192 (2022) 110913, https://doi.org/10.1016/j.measurement.2022.110913.
[49] Y. Yao, N. Wang, Fault diagnosis model of adaptive miniature circuit breaker based on fractal theory and probabilistic neural network, Mech. Syst. Signal. Process. 142 (2020) 106772, https://doi.org/10.1016/j.ymssp.2020.106772.
[50] Q. Fang, G. Xiong, X. Shang, S. Liu, B. Hu, Z. Shen, An enhanced fault diagnosis method with uncertainty quantification using Bayesian convolutional neural network, in: IEEE International Conference on Automation Science and Engineering, 2020, pp. 588–593, https://doi.org/10.1109/CASE48305.2020.9216773, 2020-Augus.
[51] W. Peng, Z.S. Ye, N. Chen, Bayesian Deep-learning-based health prognostics toward prognostics uncertainty, IEEE Transac. Industr. Electr. 67 (2020) 2283–2293, https://doi.org/10.1109/TIE.2019.2907440.
[52] T. Zhou, T. Han, E.L. Droguett, Towards trustworthy machine fault diagnosis: a probabilistic Bayesian deep learning framework, Reliab. Eng. Syst. Saf. 224 (2022) 108525, https://doi.org/10.1016/j.ress.2022.108525.
[53] T. Han, Y.F. Li, Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles, Reliab. Eng. Syst. Saf. 226 (2022), https://doi.org/10.1016/j.ress.2022.108648.
[54] S. Lee, H. Kim, J. Lee, GradDiv: adversarial robustness of randomized neural networks via gradient diversity regularization, IEEe Trans. Pattern. Anal. Mach. Intell. 45 (2022) 2645–2651, https://doi.org/10.1109/TPAMI.2022.3169217.
[55] Shridhar K., Laumann F., Liwicki M. A comprehensive guide to Bayesian Convolutional Neural Network with Variational Inference 2019:1–38.
[56] J.R. Hershey, P.A. Olsen, Approximating the Kullback Leibler divergence between Gaussian mixture models, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 4, 2007, pp. 317–320, https://doi.org/10.1109/ICASSP.2007.366913.
[57] P.T. De Boer, D.P. Kroese, S. Mannor, R.Y. Rubinstein, A tutorial on the cross-entropy method, Ann. Oper. Res. 134 (2005) 19–67, https://doi.org/10.1007/s10479-005-5724-z.
[58] Z. Xu, C. Li, Y. Yang, Fault diagnosis of rolling bearings using an improved Multi-Scale convolutional neural network with feature attention mechanism, ISA Trans. 110 (2021), https://doi.org/10.1016/j.isatra.2020.10.054.
[59] Z. Xu, X. Mei, X. Wang, M. Yue, J. Jin, Y. Yang, et al., Fault diagnosis of wind turbine bearing using a multi-scale convolutional neural network with bidirectional long short term memory and weighted majority voting for multi-sensors, Renew. Energy 182 (2022), https://doi.org/10.1016/j.renene.2021.10.024.
[60] Z. Xu, M. Bashir, W. Zhang, Y. Yang, X. Wang, C. Li, An intelligent fault diagnosis for machine maintenance using weighted soft-voting rule based multi-attention module with multi-scale information fusion, Inform. Fusion (2022) 86–87, https://doi.org/10.1016/j.inffus.2022.06.005.
[61] Arif T.M. Deep transfer learning 2020;15:37–9. https://doi.org/10.1007/978-3-031-79665-4_5.
[62] S. Sheng, Investigation of various condition monitoring techniques based on a damaged wind turbine gearbox, in: Structural Health Monitoring 2011: Condition-Based Maintenance and Intelligent Structures - Proceedings of the 8th International Workshop on Structural Health Monitoring 2, 2011, pp. 1664–1671.
[63] P. Tamilselvan, P. Wang, S. Sheng, J.M. Twomey, A two-stage diagnosis framework for wind turbine gearbox condition monitoring, Int. J. Progn. Health Manage 4 (2013), https://doi.org/10.36001/ijphm.2013.v4i3.2140.
[64] Clark K. Computing neural network gradients 2017;1:1–7.
[65] D. Mazza, M. Pagani, Automatic differentiation in PCF, Proc. ACM. Program. Lang. 5 (2021) 1–4, https://doi.org/10.1145/3434309.
[66] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., PyTorch: an imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[67] M. Zhao, S. Zhong, X. Fu, B. Tang, M. Pecht, Deep residual shrinkage networks for fault diagnosis, IEEe Trans. Industr. Inform. 16 (2020) 4681–4690, https://doi.org/10.1109/TII.2019.2943898.
[68] Yazdi M., Zarei E., Adumene S., Abbassi R., Rahnamayiezekavat P. Uncertainty modeling in risk assessment of digitalized process systems, 2022, p. 389–416. https://doi.org/10.1016/bs.mcps.2022.04.005.