Uncertainty-Aware Deep Learning: A Promising Tool for Trustworthy Fault Diagnosis

Jiaxin Ren, Jingcheng Wen, Zhibin Zhao, *Member, IEEE*, Ruqiang Yan, *Fellow, IEEE*, Xuefeng Chen, *Member, IEEE*, and Asoke K. Nandi, *Fellow, IEEE*

***Abstract*—Recently, intelligent fault diagnosis based on deep learning has been extensively investigated, exhibiting state-of-the-art performance. However, the deep learning model is often not truly trusted by users due to the lack of interpretability of "black box", which limits its deployment in safety-critical applications. A trusted fault diagnosis system requires that the faults can be accurately diagnosed in most cases, and the human in the decision-making loop can be found to deal with the abnormal situation when the models fail. In this paper, we explore a simplified method for quantifying both aleatoric and epistemic uncertainty in deterministic networks, called SAEU. In SAEU, Multivariate Gaussian distribution is employed in the deep architecture to compensate for the shortcomings of complexity and applicability of Bayesian neural networks. Based on the SAEU, we propose a unified uncertainty-aware deep learning framework (UU-DLF) to realize the grand vision of trustworthy fault diagnosis. Moreover, our UU-DLF effectively embodies the idea of "humans in the loop", which not only allows for manual intervention in abnormal situations of diagnostic models, but also makes corresponding improvements on existing models based on traceability analysis. Finally, two experiments conducted on the gearbox and aero-engine bevel gears are used to demonstrate the effectiveness of UU-DLF and explore the effective reasons behind.**

***Index Terms*—Out-of-distribution detection, traceability analysis, trustworthy fault diagnosis, uncertainty quantification.**

I. Introduction

ADVANCED engineering equipment, such as gas turbine, helicopter and nuclear power generator, is built to be the safety-critical, time-critical, and cost-critical missions under the premise of ensuring performance [1]−[3]. When these advanced engineering equipment breaks down and causes major accidents, it would bring huge economic losses and personal casualties. Therefore, it is necessary to develop prognostics and health management (PHM) technologies for diagnosing possible faults in equipment [4]−[7]. With the extensive applications of sensor networks and computing systems, the manufacturing industry has fully embraced the big data revolution [8]. However, facing with heterogeneous massive data, the manual feature extraction methods are time-consuming and empirical [9].

Due to its powerful representation ability, deep learning automatically extracts features from massive data, effectively solves the above problems, and is widely used in the field of intelligent fault diagnosis [10]−[15]. And the input of the deep learning model gradually expands from vibration signals to thermal signals, acoustic signals and encoder signals to meet the usage needs in different scenarios [16]−[25]. However, the deep learning models are often not truly trusted due to the lack of interpretability of "black box". Users do not know why the model predicts this result rather than other results and when to trust the prediction of the model. Unfortunately, the reality is more unfavorable for people to establish trust in intelligent fault diagnosis. The diagnosis models are commonly trained on limited data, and thus are only valid for independent identically distributed data based on the close-set assumption [26]. When facing with different working conditions, locations, and machines data [27]−[30], the accuracy of models decreases due to domain shift. There is a more extreme situation, that is, facing new faults that the model has never seen. At this time, deep learning models often blindly predict some previously trained fault types. This is because the distributions of training data and test data are different, and the distribution of new fault data is not within the consideration range of the diagnostic model, so it is called out-of-distribution (OOD) [31]−[33]. In the past, this situation was often classified as novelty detection. This awkward circumstance limits the further deployment of deep learning models in safety-critical applications.

Generally speaking, trustworthy fault diagnosis requires the ability to accurately diagnose faults in most cases, and when the model fails, experts will be warned to handle untrustworthy predictions given by the model. At this point, the "fault-maintenance" decision-making loop composed entirely of deep learning can shift towards the decision-making loop that combines manual intervention and deep learning. Furthermore, industrial users can make corresponding improvements on the existing model based on the type of model failure. In order to realize this grand vision, more and more studies have focused on uncertainty in deep learning, which is a promising way to solve the trust problem of the deep learning models

Manuscript received August 31, 2023; accepted January 30, 2024. This work was supported in part by the National Natural Science Foundation of China (52105116), Science Center for gas turbine project (P2022-DC-I-003-001), and the Royal Society award (IEC\NSFC\223294) to Professor Asoke K. Nandi. Recommended by Associate Editor Xin Luo. *(Corresponding author: Zhibin Zhao.)*

Citation: J. Ren, J. Wen, Z. Zhao, R. Yan, X. Chen, and A. Nandi, "Uncertainty-aware deep learning: A promising tool for trustworthy fault diagnosis," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 6, pp. 1317–1330, Jun. 2024.

J. Ren, J. Wen, Z. Zhao, R. Yan, and X. Chen are with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: r19981005@stu.xjtu.edu.cn; alphawjc@stu.xjtu.edu.cn; zhaozhibin@xjtu.edu.cn; yanruqiang@xjtu.edu.cn; chenxf@xjtu.edu.cn).

A. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, UK (e-mail: asoke.nandi@brunel.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2024.124290

[34]. Uncertainty can help the diagnosis models identify its own knowledge boundary, that is, assessing to what extent the model knows what it knows. An ideal trustworthy fault diagnosis model should not only achieve high-precision prediction, but also evaluate the uncertainty of the results to reflect the confidence in the prediction results.

It is exciting that the fault diagnosis community has paid growing attention to uncertainty in deep learning and regarded it as one of promising directions for deep learning [35], [36]. Zhou *et al.* [37] proposed the uncertainty-informed framework using Bayesian neural network (BNN) for trustworthy fault diagnosis. However, this framework remains at the stage of predictive uncertainty quantification and has not further refined the uncertainty to determine the source, so the impact of signal noise on diagnostic uncertainty has not been considered. Xiao *et al.* [38] proposed the trustworthy rotating machinery fault diagnosis method via attention uncertainty in transformer. It still does not break away from the framework of BNN and is not suitable for most widely used deterministic neural networks, such as convolutional neural network (CNN), recursive neural network (RNN) and long short-term memory (LSTM). In short, although BNN-based fault diagnosis has natural advantages in quantifying uncertainty, it has obvious disadvantage in terms of computational complexity. At the same time, there is still a significant gap in performance and scalability compared to mainstream deep learning models.

Some scholars have noticed the limitations of BNN and started exploring uncertainty quantification methods applicable to widely-used deterministic neural networks. Many research achievements have emerged in fields such as computer vision, digital histopathology and civil engineering [39]−[42]. In the field of fault diagnosis, Han and Li [43] proposed OOD detection assisted trustworthy fault diagnosis approach with uncertainty-aware deep ensembles. This is the most relevant literature to our work, which attempts to quantify uncertainty in the most widely used deterministic networks and utilize uncertainty to achieve OOD detection. Unfortunately, it only quantifies the epistemic uncertainty and does not consider the impact of aleatoric uncertainty. Therefore, in order to further improve the safety of intelligent fault diagnosis, it is necessary to consider aleatoric uncertainty caused by signal noise in deterministic networks.

Other scholars [44] have attempted to quantify uncertainty from the perspective of evidence theory, which models network output distribution as the Dirichlet distribution and changes the classic deep architecture, mainly reflected in losses. Zhou *et al.* proposed evidential deep neural networks with uncertainty estimation for fault diagnosis, one based on VGG [45] and another based on CNN [46]. The problem with evidential deep learning is that the output only has predictive uncertainty and cannot be decomposed based on the source of uncertainty. This is actually facing the same dilemma as [43], which requires deterministic neural networks to quantify both epistemic uncertainty and aleatoric uncertainty. And recently, someone showed that evidential deep learning in regression task is a heuristic rather than an exact uncertainty quantification from a theoretical perspective [47].

This forces us to regress from evidence theory to Bayesian theory in search of method for simultaneously quantifying and decomposing uncertainty in deterministic networks. In fact, evidence deep learning has inspired us by changing the output form of the model. The existing deterministic networks are unable to quantify and decompose uncertainty simultaneously due to insufficient representation ability of outputs. Therefore, assuming that the distribution of the model output follows a multivariate Gaussian distribution (MGD), where covariance is used to approximate aleatoric uncertainty, and then quantifying the epistemic uncertainty through the diversity of network structures is a feasible solution. In addition, the existing works focus on detecting the OOD situation of the model to partially achieve trustworthy fault diagnosis. However, there is still a lack of research on how users can utilize uncertainty to improve existing models to fully achieve trustworthy fault diagnosis we previously defined.

Based on existing work findings, the quantification and decomposition of uncertainty in deterministic neural networks have not been adequately addressed. And there is also a lack of a unified research framework for existing model improvement after uncertainty decomposition. So, in this study, a novel unified uncertainty-aware deep learning framework (UU-DLF) is proposed for trustworthy fault diagnosis. To the best of our knowledge, this is the first exploration to quantify and decompose uncertainty for fault diagnosis in the deterministic networks (i.e., CNNs or RNNs) and attempt to utilize uncertainty to feed back the modification of models according to different sources. And with only minor modifications to the existing network, it can be integrated into UU-DLF. Our main contributions are as follows:

1) We explore a simplified method for quantifying both aleatoric and epistemic uncertainties in deterministic networks. Specifically, we model the network output distribution as MGD in order to directly capture the aleatoric uncertainty, which is compatible with various epistemic uncertainty representation methods and effectively reduces the computational complexity of quantifying aleatoric and epistemic uncertainties.

2) We establish a unified trustworthy fault diagnosis framework based on the aleatoric and epistemic uncertainties, which is easy to be compatible with existing models. It helps to realize model failure warning and model improvement from the perspective of uncertainty decomposition, thereby providing a promising way for deep learning models to gain users' trust.

The rest of this paper is organized as follows. In Section II, we briefly elaborate the theoretical background of existing works and their shortcomings. In Section III, we introduce a simplified algorithm for simultaneously quantifying both aleatoric and epistemic uncertainties in deterministic networks. And we propose the unified uncertainty-aware deep learning framework and give experimental verifications in Sections IV and V. Finally, Section VI concludes this article.

II. Theoretical Backgrounds

As stated in Section I, as a potentially promising tool to achieve trustworthy fault diagnosis, the aleatoric and epistemic uncertainties represented by BNN will be extrapolated

to all deterministic neural networks (NNs) in this study. Before introducing our proposed method, we need to understand the mechanism how BNN represents uncertainty. Blundell *et al.* [48] conducted probabilistic modeling for the model weights and further proposed BNN. Differently from the point estimation in deterministic NN weights, BNN attempts to learn the posterior distribution over the weights $p(\boldsymbol{\omega}|\boldsymbol{D})$, which captures the set of plausible model weights given by the dataset $\boldsymbol{D}$. The posterior distribution of the trainable weights based on Bayesian theorem can be written as (1).

$$p(\boldsymbol{\omega}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\omega})\,p(\boldsymbol{\omega})}{p(\boldsymbol{D})} \tag{1}$$

where $p(\boldsymbol{\omega})$ is the prior distribution over model weights, $p(\boldsymbol{D}|\boldsymbol{\omega})$ is the data likelihood for specific weights and $p(\boldsymbol{D})$ is the marginal likelihood, often a constant. Generally speaking, a prior distribution represents priori knowledge of parameters based on experience, while a posterior distribution is the recognition of weights under the constraint of data likelihood. However, the marginal likelihood is difficult to calculate since it requires the integral over all possible weights as (2).

$$p(\boldsymbol{D}) = \int p(\boldsymbol{D}|\boldsymbol{\omega})\,p(\boldsymbol{\omega})\,d\boldsymbol{\omega}. \tag{2}$$

For increasingly large deep NNs, the marginal likelihood estimation becomes intractable due to the large number of weights. One way to solve the above problem is variational inference, whose core idea is to find an approximate distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ that is as close as the true posterior distribution. In $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$, $\boldsymbol{\theta}$ is the parameterized parameter of the $q(\boldsymbol{\omega})$ distribution. For example, $\boldsymbol{\theta}$ represents the mean and variance for Gaussian distribution. Thus, the learning process of BNN involves optimizing a composite loss function (3), one of which is to find the optimal model weights with respect to the training data via relying on the approximate distribution, and the other is to minimize the difference between the approximate distribution and the true posterior distribution by Kullback-Leibler divergence (KL divergence).

$$\mathcal{L}_{VI}(\boldsymbol{\theta}) = -\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{\omega})}(\log p(\boldsymbol{D}|\boldsymbol{\omega})) + \mathrm{KL}(q_{\boldsymbol{\theta}}(\omega)||p(\boldsymbol{\omega})). \tag{3}$$

To reduce the computational complexity, the loss function is approximated by an unbiased Monte Carlo estimation. Here, the tractable loss function becomes

$$\mathcal{L}_{VI}(\boldsymbol{\theta}) \approx -\sum_{k=1}^{K} \log p\left(\boldsymbol{D}|\boldsymbol{\omega}^k\right) + \log q_{\boldsymbol{\theta}}\left(\boldsymbol{\omega}^k\right) - \log p\left(\boldsymbol{\omega}^k\right) \tag{4}$$

where K denotes the number of Monte Carlo sampling from the approximate weight distribution.

Traditionally, uncertainty in deep learning is decomposed into aleatoric and epistemic parts according to different sources, also known as data and model uncertainties [30]. Aleatoric uncertainty, reflecting the inherent uncertainty in the input data, cannot be reduced with more data (the noise of data will not decrease with the increase of data volume). High aleatoric uncertainty means that the input data is too noisy to make accurate predictions. Epistemic uncertainty, reflecting the uncertainty in the different model weights, can be reduced with more data (the knowledge of data will increase with the increase of data volume). High epistemic uncertainty means that the model only learns limited knowledge and the test data exceeds the boundary of model knowledge.

In order to quantify aleatoric and epistemic uncertainties in BNN, it is necessary to conduct an additional Monte Carlo sampling to obtain different outputs $\hat{\boldsymbol{y}}^m$ under each fixed weight $\boldsymbol{\omega}^k$. Aleatoric uncertainty $\mathcal{U}_a$ can be mathematically represented by the mean entropy of outputs as (5).

$$\begin{aligned}\mathcal{U}_a &= E\left[\mathcal{H}\left(\hat{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{\omega}^k\right)\right] \\ &\approx -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\left(\frac{1}{M}\sum_{m=1}^{M}\hat{y}^{c,m,k}\right)\log\left(\frac{1}{M}\sum_{m=1}^{M}\hat{y}^{c,m,k}\right)\end{aligned} \tag{5}$$

where $\mathbb{E}(\cdot)$ denotes the mathematical expectations, $\mathcal{H}(\cdot)$ denotes the entropy, M denotes the number of Monte Carlo sampling from the output distribution and C denotes the class number. Correspondingly, the epistemic uncertainty $\mathcal{U}_e$ can be represented by the difference between the entropy of mean outputs $\mathcal{H}(\mathbb{E}(\hat{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{\omega}^k))$ and the mean entropy of outputs $\mathbb{E}[\mathcal{H}\left(\hat{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{\omega}^k\right)]$.

$$\begin{aligned}\mathcal{U}_e &= \mathcal{H}\left(\mathbb{E}(\hat{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{\omega}^k)\right) - \mathbb{E}\left[\mathcal{H}\left(\hat{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{\omega}^k\right)\right] \\ &\approx -\sum_{c=1}^{C}\left(\frac{1}{K}\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\hat{y}^{c,m,k}\right)\log\left(\frac{1}{K}\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\hat{y}^{c,m,k}\right) \\ &\quad + \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\left(\frac{1}{M}\sum_{m=1}^{M}\hat{y}^{c,m,k}\right)\log\left(\frac{1}{M}\sum_{m=1}^{M}\hat{y}^{c,m,k}\right).\end{aligned} \tag{6}$$

It is not difficult to find that uncertainty quantification and decomposition based on BNN are currently complex and high in calculation cost, among which the algorithm complexity for uncertainty quantification is $O(n^3)$. More importantly, they are difficult to be compatible with most deterministic NNs.

In order to quantify uncertainty in deterministic NNs, the entropy of softmax probability is calculated as the predictive uncertainty estimates [43], written as (7).

$$\begin{aligned}\mathcal{U}_p &= \mathcal{H}\left(\mathbb{E}(\hat{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{\omega}^k)\right) \\ &\approx -\sum_{c=1}^{C}\left(\frac{1}{K}\sum_{k=1}^{K}\hat{y}^{c,k}\right)\log\left(\frac{1}{K}\sum_{k=1}^{K}\hat{y}^{c,k}\right).\end{aligned} \tag{7}$$

Although the uncertainty in deterministic NNs has been quantified, the lack of an additional Monte Carlo sampling makes it ineffective for uncertainty decomposition.

Other methods for quantifying uncertainty in deterministic NNs, such as evidential deep learning, have significant gaps compared to existing diagnostic models in terms of representing and learning uncertainty. Zhou *et al.* [46] replaced the *softmax* function of the network output layer with Dirichlet distribution, where the probability density function is denoted as (8).

$$\boldsymbol{D}(\mathbf{p}|\boldsymbol{\beta}) = \begin{cases}\dfrac{1}{B(\boldsymbol{\beta})}\prod_{i=1}^{M} p_i^{\beta_i - 1}, & \text{for } \mathbf{p} \in V_M \\ 0, & \text{otherwise}\end{cases} \tag{8}$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_M]$ is the Dirichlet parameter and V_M is

the M-dimensional unit simplex. For learning uncertainty in evidential deep learning, the loss function can be given by Bayes risk, as shown in (9).

$$\mathcal{L}_{\mathrm{E}}(\boldsymbol{\theta}) = \int \left[\sum_{m=1}^{M} -y_{im} \log(p_{im}) \right] \prod_{m=1}^{M} p_{im}^{\beta_{im}-1} d\mathbf{p}_i. \tag{9}$$

Finally, the mean and variance of the Dirichlet distribution are used for prediction and its uncertainty in evidential learning, as calculated in (10) and (11) respectively.

$$\hat{p}_m = \mathbb{E}[p_m] = \frac{\beta_m}{V} \tag{10}$$

$$\mathcal{U}_p = \mathbb{V}[p_m] = \frac{\beta_m(V-\beta_m)}{V^2(V+1)} \tag{11}$$

where $V = \sum_{m=1}^{M} \beta_m$ is the Dirichlet strength. Whether in terms of model structure or model training, evidential deep learning requires significant changes to existing networks. From the perspective of quantifying uncertainty, it still only quantifies the predictive uncertainty and cannot be further decomposed.

In summary of the above work, our research goal is to find a simplified aleatoric and epistemic uncertainties representation, learning, and quantification method that is compatible with existing models. And due to the lack of research on uncertainty decomposition, there is no clear explanation on the important role of aleatoric and epistemic uncertainties in intelligent fault diagnosis. Another goal of our work is attempting to establish a reliable connection between uncertainty decomposition and model improvement through traceability analysis.

III. Methodology

In this section, the architecture for representing, learning, and quantifying aleatoric and epistemic uncertainties in widely used deterministic NNs is elaborated. The motivation comes from our understanding and promotion of BNN [48], [49], and the specific implementation is inspired by the evidential deep learning to model the network output as a Dirichlet distribution.

A. Representing Uncertainty in Deterministic NNs

Based on the above theoretical backgrounds of BNN, the mechanism of uncertainty comes from the probabilistic modeling of the model output and weights. Two-fold Monte Carlo sampling realized the effective approximation of difficult weight distributions and output distributions which are intractable to solve. In order to simplify the operations and make it suitable for deterministic NNs, the representation of uncertainty needs to be re-designed from both the model output and weights.

For representing the aleatoric uncertainty, the MGD is employed as the model output. For the C-class fault diagnosis, the model output $\hat{\boldsymbol{y}}$ can be decoupled into different distribution parameters of MGD as (12).

$$\mathcal{N}(\hat{\boldsymbol{y}}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{C/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\right) \tag{12}$$

where $\boldsymbol{\delta} = \hat{\boldsymbol{y}} - \boldsymbol{\mu}$ represents the biases, $\boldsymbol{\mu}$ is the true mean vector of MGD and $\boldsymbol{\Sigma}$ is the true covariance matrix of MGD. Since the class is independent of each other, the non-diagonal elements in the true covariance matrix $\boldsymbol{\Sigma}$ are all zero. The advantage of output distribution parameterization is that the model output can directly represent the distribution without Monte Carlo sampling. To embed the MGD into in deterministic NNs, the core design is to use a common encoder to extract the rich features in the input signal, and use different decoders to map these common features to different parameters of MGD, as shown in the Fig. 1. Then, the approximate output mean vector $\hat{\boldsymbol{\mu}}$ can be regarded as the classification probability in the sense of MGD and the approximate output diagonal elements $\mathrm{diag}(\hat{\boldsymbol{\sigma}})$ in $\boldsymbol{\Sigma}$ can represent the aleatoric uncertainty.

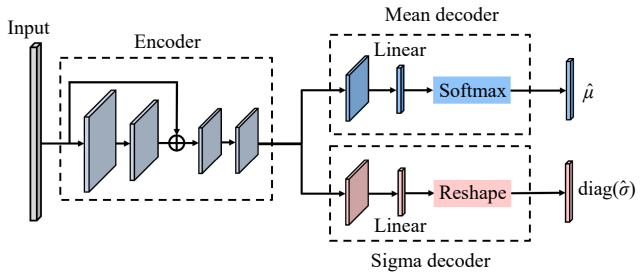


Fig. 1. The embedding process of MGD in deterministic NN.

For representing the epistemic uncertainty, the key is to construct multiple different weights of the same deterministic NN. According to the BNN backgrounds, the construction process of multiple different weights is equivalent to Monte Carlo sampling from the approximate weight distribution. At present, there are different mature methods to represent the epistemic uncertainty, namely Monte Carlo-dropout (MCD) , deep ensembling (DE) [50] and its variant, bootstrapping (BS). These methods are already good enough to represent model uncertainty, and there is no need for us to continue proposing new methods. The innovation of our work lies in the ability to represent aleatoric uncertainty based on existing mature epistemic uncertainty representation methods. And by embedding MGD distribution, it is possible to quantify both aleatoric uncertainty and epistemic uncertainty simultaneously in deterministic NNs. At the same time, the representation of epistemic uncertainty is not limited to deep ensembling [43], which allows users to make flexible choices according to their needs.

According to the mechanism of uncertainty generation in BNN, epistemic uncertainty mainly comes from the model weights distribution. According to our understanding, the three methods mentioned above attempt to construct the diversity of model weights in different ways. Fig. 2 shows a schematic comparison of diversity embodied by MCD, DE and BS methods. They make use of the diversity to generate the approximate weights distribution similar to BNN, in which MCD utilizes the diversity of network structures, DE utilizes the diversity of model initializations, and BS utilizes the diversity of training datasets. According to different diversities, the epistemic uncertainty representation ability of models may also vary.

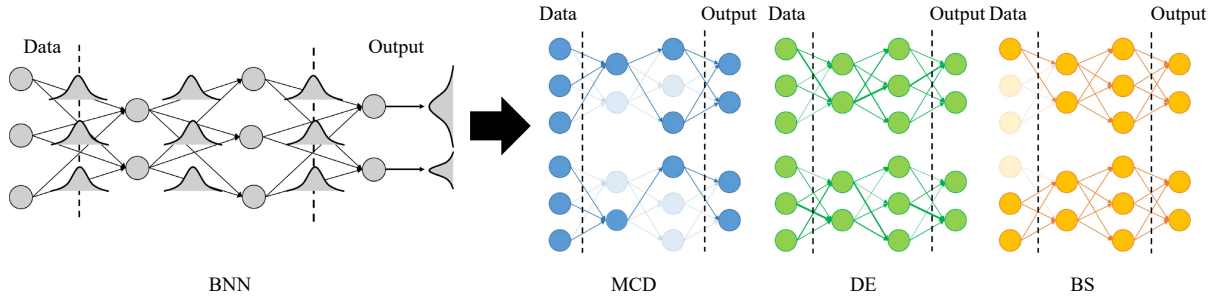


Fig. 2. A schematic comparison of diversity embodied by different methods. In order to obtain a weight distribution similar to BNN, MCD utilizes the diversity of network structure, while DE utilizes the diversity of model initialization and BS utilizes the diversity of the training dataset.

B. Learning Uncertainty in Deterministic NNs

Since there is no weight distribution in the deterministic NNs, naturally the variational inference loss function in (4) based on approximate distribution is no longer applicable. However, the cross-entropy loss commonly used in the training procedure cannot deal with the uncertainty under the MGD. In order to learn uncertainty in deep NNs, an appropriate loss function needs to be explored.

For learning the mean vector and covariance matrix of the MGD at the same time, the re-parameterization technique is used to reorganize the widely used cross-entropy loss function. It implicitly constrains the output approximate mean vector and covariance matrix by sampling from the parameter-free noise $\boldsymbol{\epsilon}$ and ensure that the derivative can be back-propagated. For the ith sample in the training dataset, the classification probability after re-parameterization can be written as (13).

$$\hat{p}^{i,c} = \frac{\hat{\mu}^{i,c} + \hat{\sigma}^{i,c}\epsilon^{i,c}}{\sum_{c}\left(\hat{\mu}^{i,c} + \hat{\sigma}^{i,c}\epsilon^{i,c}\right)} \text{ with } \epsilon^{i,c} \sim \mathcal{N}(0,1) \tag{13}$$

where c means cth class in C, $\hat{\mu}^{i,c}$ means the approximate output mean of the ith sample's c class logits, $\hat{\sigma}^{i,c}$ means the approximate output diagonal elements of the ith sample's c class logits and $\epsilon^{i,c}$ means the noise of the ith sample's c class. It is essentially an extension of Softmax probability in the deterministic NN, taking into account the diagonal elements representing the aleatoric uncertainty. Naturally, the cross-entropy loss can continue to be used after slight modification of the classification probability as (14), marked as U-CE loss. Compared to the loss function of evidential deep learning, our U-CE loss is clearly better optimized and more compatible with existing models.

$$\mathcal{L}_{\text{U-CE}}(\boldsymbol{\theta}) \approx -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y^{i,c}\log\left(\hat{p}^{i,c}\right). \tag{14}$$

C. Quantifying Uncertainty in Deterministic NNs

Due to the embedding of MGD in deterministic NNs, it is not possible to quantify aleatoric uncertainty by multiple Monte Carlo sampling under the same weight. Meanwhile, for different outputs of each sample, the epistemic uncertainty quantified by calculating entropy is computing-consuming like (5) and (6). In order to realize successfully simpler uncertainty quantification in the deterministic NN, a new way to quantify the aleatoric and epistemic uncertainties is bound to find.

After performing K stochastic forward passes to obtain different weights, all those approximate output mean vectors are averaged to compute the final prediction $\sum_{k=1}^{K}\hat{\boldsymbol{\mu}}_k/K$. Differently from the entropy, a convenient way to quantify the uncertainty is to directly quantify the "variance" in MGD output. The determinant of approximate output covariance matrix $\left|\hat{\boldsymbol{\Sigma}}_k\right|$ is used to quantify aleatoric uncertainty, which describes the fluctuation of classification probability on all categories under a fixed weight. The form of continuous multiplication will amplify the weak fluctuations, so that the uncertainty of ID sample is smaller, and the uncertainty of OOD sample is greater. Similarly, the variance of approximate output mean vectors is used to quantify epistemic uncertainty, which describes the fluctuation of different classification probabilities caused by the variability of model weights. In summary, our aleatoric and epistemic uncertainties could be calculated respectively, shown in (15) and (16).

$$\mathcal{U}_a = \frac{1}{K}\sum_{k=1}^{K}\left|\hat{\boldsymbol{\Sigma}}_k\right| = \frac{1}{K}\sum_{k=1}^{K}\prod_{c=1}^{C}\hat{\sigma}^{i,c^2} \tag{15}$$

$$\mathcal{U}_e = \frac{1}{K}\sum_{k=1}^{K}(\hat{\boldsymbol{\mu}}_k - \bar{\boldsymbol{\mu}})^2 = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\boldsymbol{\mu}}_k - \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\mu}}_k\right)^2. \tag{16}$$

At this point, our proposed methodology for representing, learning and quantifying uncertainty can be summarized as the simplified aleatoric-epistemic uncertainties (SAEU) algorithm, and its flowchart is shown in Fig. 3.

Unlike the existing methods, our SAEU model not only outputs an integrated diagnostic result, but also provides the aleatoric and epistemic uncertainties. After training all the data, we can collect a set of diagnostic uncertainties for the training dataset. Then, we use kernel density estimation to infer joint and marginal distributions of decomposed uncertainties based on training data, and based on this, we can draw an uncertainty decomposition graph as shown in Fig. 4. Specifically, we recorded the aleatoric uncertainty and epistemic uncertainty logarithmically for each input, as shown in the middle scatter. The vertical and horizontal axes represent the histogram of aleatoric and epistemic uncertainties. Finally, kernel density estimation is used to fit the marginal distributions and joint distribution of aleatoric and epistemic uncertainties separately.

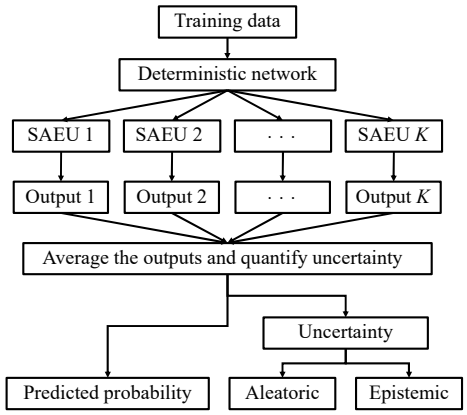


Fig. 3. The flowchart of SAEU algorithm.

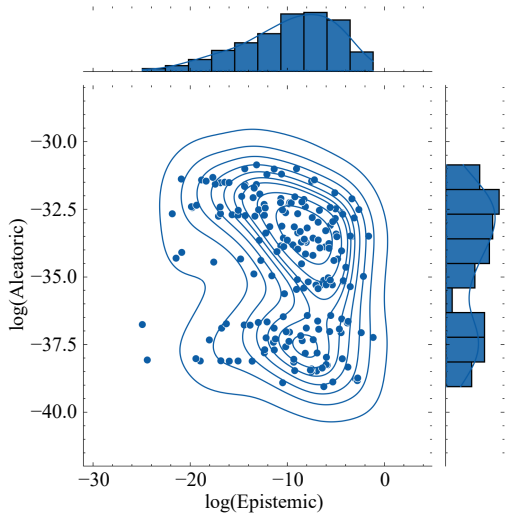


Fig. 4. Schematic diagram of uncertainty decomposition graph for training set.

IV. Unified Uncertainty-Aware Deep Learning Framework

Typically, deep learning models are evaluated based on a testing dataset, which is sampled from the same distribution as the training dataset. However, for security reasons, there is a strong need for industrial equipment to evaluate against OOD/domain shift inputs sampled from the different distributions. Among them, OOD inputs refer to new fault data unknown to models, while domain shift inputs generally refer to variable speed, variable load, and noisy samples affected by the different environments [51], [52]. This situation is quite common in real safety-critical applications such as gas turbine, helicopter and nuclear power generator. Therefore, we establish a unified uncertainty-aware deep learning framework (UU-DLF) based on our SAEU algorithm, in order to improve the interpretability of existing models from post perspective, as shown in Fig. 5.

A. Human Intervention in UU-DLF

By estimating the aleatoric and epistemic uncertainties to understand the knowledge boundary of the "black box" model, users can trust the decision-makings with low uncertainty and intervene the decision-makings with high uncertainty, as shown in the human intervention part in Fig. 5. We use uncertainty decomposition graph to detect potential OOD/domain shift samples by threshold determined on the training dataset. When there is new input, SAEU will provide corresponding aleatoric and epistemic uncertainties. If one of the uncertainties exceeds the 0.9 quantile of the uncertainty distribution in the training set, the input will be detected as anomaly. Based on existing threshold rule, the detection area in uncertainty decomposition graph is a rectangle. Obviously, the detection threshold can be determined using composite rules, meaning that the detection area can be any two-dimensional shape in the uncertainty decomposition graph.

For the prediction with low uncertainty, it can be regarded as trustworthy and diagnostic results can be directly output. In contrast, the prediction with high uncertainty would be regarded as untrustworthy and alert human experts, which to some extent realizes the assumption of "human in the loop". Then the untrustworthy prediction is output together with expert opinions judged by other effective methods (i.e., expert experience or signal processing results). Specifically, human experts can combine the analysis results of classical signal processing methods, such as wavelet analysis, with the model predictions as trustworthy expert opinions. Taking the new fault input as an example, its epistemic uncertainty is much larger than the known fault sample, so it can be detected by setting a threshold on the epistemic uncertainty.

B. Traceability Analysis in UU-DLF

By analyzing the sources of uncertainty, whether it is aleatoric or epistemic, users can conduct traceability analysis to make corresponding improvements on existing models, as shown in the traceability analysis part in Fig. 5. Specifically, for an unknown testing set, another uncertainty joint distribution can be obtained using SAEU method. Taking OOD testing dataset as an example, the joint distributions of uncertainty for the training and testing datasets are shown in blue and red in Fig. 6(a), respectively. Traceability analysis in our framework divides it into four regions based on the relative position of the probability density centers of two uncertainty joint distributions to determine the cause: Trustworthy, OOD, NOISY, OOD and NOISY, as shown in Fig. 6(b). The threshold rule still follows the 0.9 quantile of the uncertainty distribution in the training set. In fact, from the relative position in Fig. 6(b), it can be preliminarily observed that the reason for the shift of the testing data distribution. For easier usage, we quantitatively provide the criteria for determining the causes for model failure in the traceability analysis part in Fig. 5. Compared to the threshold, it is the value of the joint distribution center. It is worth noting that human intervention mainly considers the trustworthiness of the model for a single input sample, while traceability analysis is the trustworthiness of the model for the entire testing dataset.

For samples with high aleatoric uncertainty, they cannot be effectively predicted due to data noise or sensor failure. At this time, the model gives the warning of "NOISY" to remind experts to gather more information to refine the current input data. In this case, installing better sensors or improving data

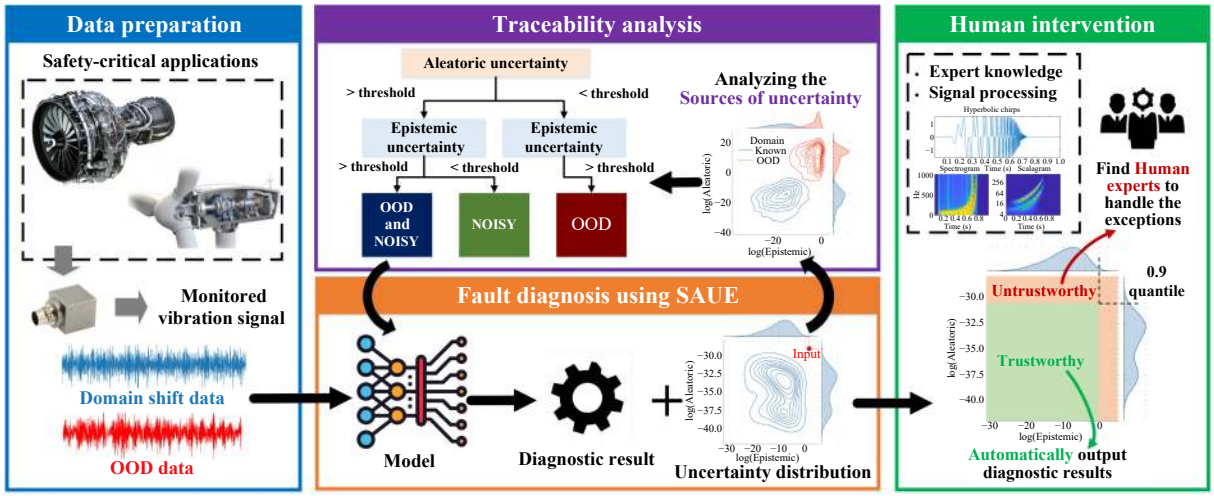


Fig. 5. The overall architectures for our proposed UU-DLF including human intervention and traceability analysis.

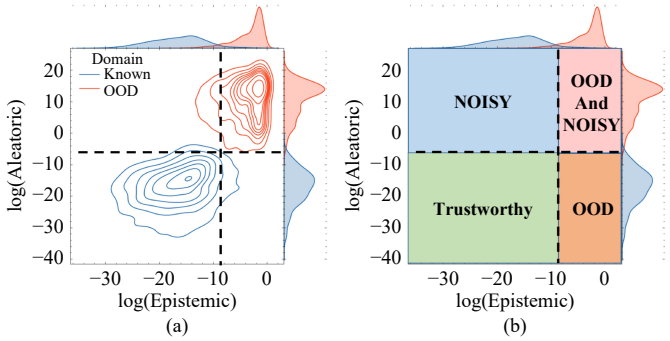


Fig. 6. Four regions (Trustworthy, OOD, NOISY, OOD and NOISY) demonstrated in the uncertainty decomposition graph after traceability analysis. (a) Uncertainty decomposition graph; (b) Four regions after traceability analysis.

preprocessing procedures can effectively improve the model performance. For samples with high epistemic uncertainty, they also cannot be effectively predicted because the deployed model contains only limited knowledge. At this time, the model gives the warning of "OOD" to remind human experts to improve insufficient model capability. In this case, expanding the amount of training dataset or enhancing the presentation ability can also effectively improve the model performance. For samples with high aleatoric and epistemic uncertainties, the model gives the joint warning of "OOD and NOISY" and reminds human experts to take corresponding measures.

In short, our proposed unified trustworthy fault diagnosis framework can not only handle the abnormal situations that existing deep learning models cannot handle through human intervention, but also conduct traceability analysis based on uncertainty decomposition, identify the causes of uncertainty, and further improve existing models. This makes it possible for the "black box" model to have post-hoc interpretability.

V. Experimental Result

For the sake of robust evaluation in the field of trustworthy fault diagnosis, the experiments are verified on both open source and real-world datasets respectively. The proposed UU-DLF was implemented in Python 3.8 and Pytorch 1.8 and experiments were conducted on a NVIDIA GeForce GTX 2060. Since our UU-DLF is applicable to any deterministic NN, the ResNet18 [53] is used as the backbone model in this study. In our UU-DLF, DE and BS methods both use unmodified ResNet18 and the MCD method adds a dropout layer ($p = 0.5$) between each weight layer in unmodified ResNet18. And Bayesian ResNet18 is compared as a baseline method for uncertainty estimation, which is similar to ResNet18 except that its parameters are set to Gaussian distributions. The optimizer is Adam with the learning rate of 0.001. The batch size is 32, and the weight decay is 10^{-5}.

Due to the fact that our experiment needs to reflect both the diagnosis performance in the original domain and the detection performance in the new domain, we use multiple indicators to quantitatively evaluate the model performance including diagnosis accuracy (ACC), OOD detection accuracy (ODA) and false alarm rate (FAR). Among them, ACC is the most basic indicators in fault diagnosis, which can be written as (17).

$$\mathrm{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

where TP, FN, TN and FP are the numbers of true positive, false positive, true negative and false negative samples in the confusion matrix, respectively. Note that OOD samples are also involved in the confusion matrix. ODA reflects the performance of OOD detection, which can be written as (18). It mainly counts how many samples are detected in all OOD samples.

$$\mathrm{ODA} = \frac{M_d}{N_{\mathrm{OOD}}} \tag{18}$$

where M_d is the number of correctly detected OOD samples; N_{OOD} is the number of OOD samples in the mixed testing set. FAR reflects the ratio of OOD sample detection errors, which can be written as (19). It mainly counts how many samples that require fault diagnosis are incorrectly detected as OOD samples.

$$\mathrm{FAR} = \frac{M_e}{N_{\mathrm{total}} - N_{\mathrm{OOD}}} \tag{19}$$

where M_e is the number of incorrectly detected OOD samples; N_{total} is the number of all samples in the mixed testing set.

A. Fault Diagnosis Experiment of Gearbox

The mechanical gearbox datasets from Southeast university are used as an open-source dataset to validate the effectiveness of the proposed framework for diagnosis [54]. These vibration data are collected from the drivetrain dynamic simulator (DDS), and the test bench is shown in the Fig. 7. After data collection, the original vibration signals are divided into segments (each with 1024 data points) that do not overlap. The model is trained to identify five different gear health states, namely, healthy (H), chipped fault (CF), miss fault (MF), root fault (RF), and surface fault (SF) under 20Hz-0V. However, in practical applications, the gearbox will not fail in a given way under a fixed working condition as in model training.

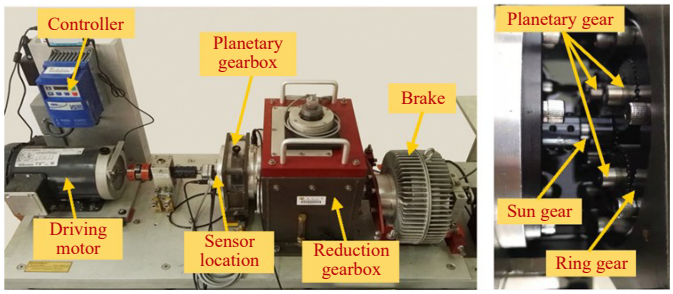


Fig. 7. DDS gearbox test bench [54].

In this experiment, we test the OOD detection performance of our proposed UU-DLF when encountering unknown new fault. The bearing inner race fault (IRF) is used to simulate the OOD fault, since the model is only trained on gear fault data. Additional bearing fault samples for testing maintain a 1 : 1 balance ratio with the entire original testing dataset. The experimental settings are shown in Table I.

TABLE I
DETAILS OF THE MECHANICAL GEARBOX DATASET

Domain	Health state	Label	Training samples	Testing samples
ID	H	0	800	200
ID	CF	1	800	200
ID	MF	2	800	200
ID	RF	3	800	200
ID	SF	4	800	200
OOD	IRF	5	/	1000

After using our SAEU algorithm in mixed testing dataset, three different uncertainty decomposition graphs are provided based on MCD, DE and BS respectively, which represent the joint distributions composed of aleatoric and epistemic uncertainties. The comparison with our SAEU algorithm is the BNN, and we have also drawn an uncertainty decomposition graph. These uncertainty decomposition graphs are shown in Fig. 8. It is not difficult to find that the joint distribution of uncertainty in the testing set is located in the upper right corner of the uncertainty distribution in the training set. According to the region division of UU-DLF, it can be roughly classified as "OOD and NOISY". Starting from the aspect of uncertainty representation ability, our SAEU algorithm outperforms the baseline method (BNN) in distinguishing ID and OOD data, whether based on aleatoric uncertainty distribution or epistemic uncertainty distribution, laying the foundation for better OOD detection performance.

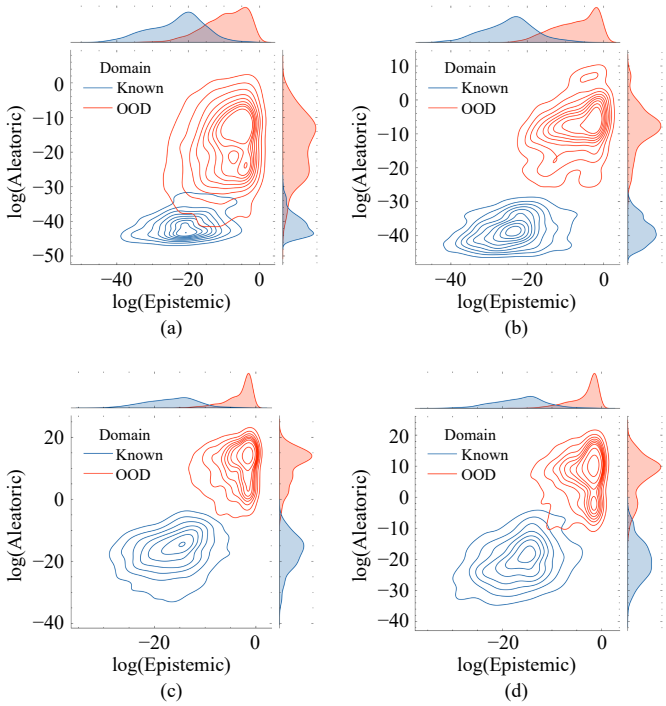


Fig. 8. Uncertainty decomposition graphs of different methods for the mixed testing dataset. BNN is used as a baseline for comparison. (a) BNN (baseline); (b) SAEU-MCD; (c) SAEU-DE; (d) SAEU-BS.

As shown in Fig. 8, it is clear that the aleatoric and epistemic uncertainties of unknown domain are much higher than those of the known domain in all methods. This result is consistent with our previous analysis of the sources of uncertainties and is due to the insufficient diagnostic knowledge. After obtaining the uncertainty decomposition graph, we can use the threshold to achieve the OOD detection and handle abnormal samples that the model cannot effectively recognize to experts for judgment based on human intervention in UU-DLF. Further analysis of Fig. 8 reveals that there is a distribution differentiation of ID and OOD uncertainty, with little overlap. Thus, we use the 0.9 quantile of epistemic uncertainty within the training dataset as a threshold to achieve the uncertainty-guided OOD detection. The diagnostic and OOD detection results are summarized in Table II and shown in the confusion matrices of Fig. 9.

According to the traceability analysis in UU-DLF, we find that the OOD data contains more noise than the ID data. Based on the quantitative criteria for determining the causes for model failure, after comparison of the joint distribution probability density centers, it is judged as "OOD and NOISY". Therefore, we can provide two suggestions for

TABLE II
THE OOD DETECTION RESULTS OF THE MECHANICAL GEARBOX DATASET

		CNN		BNN		SAEU-MCD		SAEU- DE		SAEU- BS	
		Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best
ID dataset	ACC	98.70	100.00	96.70	99.90	98.50	100.00	99.10	100.00	98.70	100.00
Mixed dataset	ACC	49.35	50.00	48.35	49.95	49.25	50.00	49.55	50.00	49.35	50.00
Mixed dataset using UU-DLF	ACC	N	N	84.70	90.75	87.80	91.55	93.35	**93.40**	93.05	92.80
	ODA	N	N	79.30	91.50	85.50	93.10	96.60	**96.80**	94.80	95.50
	FAR	N	N	10.02	9.89	10.01	10.00	9.95	**9.81**	10.20	10.01

* N represents that model has no OOD detection ability.

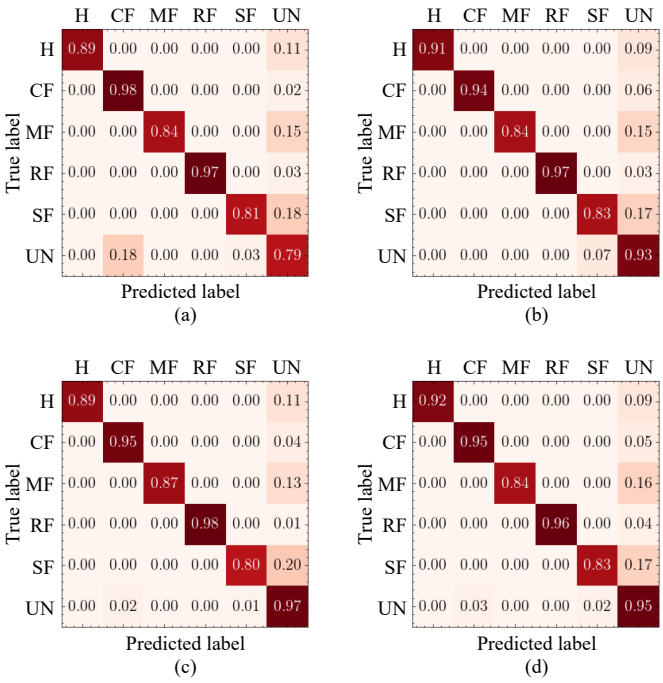


Fig. 9. Confusion matrixes of different methods in gearbox mixed testing dataset. BNN is used as a baseline for comparison. (a) BNN (baseline); (b) SAEU-MCD; (c) SAEU-DE; (d) SAEU-BS.

industry users from the perspective of model improvement: on the one hand, in order to expand the knowledge boundary of the model, users can add new fault samples in model training, and on the other hand, check the sensor installation and other aspects to reduce noise, thereby further improving the safety and reliability of the existing fault diagnosis system.

B. Fault Diagnosis Experiment of Aero-Engine Bevel Gear

The aero-engine bevel gear datasets are collected from aero-engine lubricating oil accessory control system, whose test bench and sensors installation are shown in Fig. 10. The sampling rate of the testing system is set to 20 000 Hz. The diagnostic model is trained to identify four different gear health states, namely, healthy (H), surface wear (SW), broken tooth (BT), and small end collapse (SEC) under 1000 r/min. Four different bevel gear health states are shown as shown in Fig. 11. Same as the settings in the previous section, the bearing inner race fault (IRF) is used to simulate the possible OOD samples in reality and additional OOD samples for mixed testing dataset maintain a 1 : 1 balance ratio with the

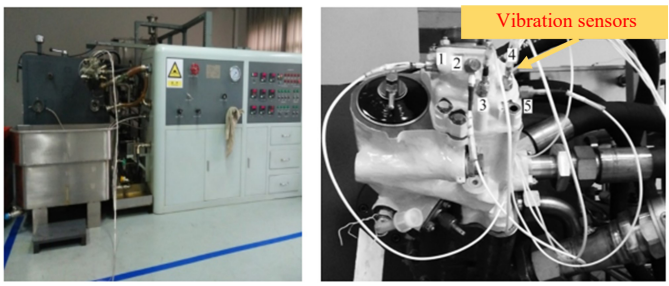


Fig. 10. Experimental bench of lubricating oil accessory control system and its sensor measuring points.

Fig. 11. Four different aero-engine bevel gear health states. (a) H; (b) SW; (c) BT; (d) SEC.

entire original testing set. The experimental settings are shown in Table III.

TABLE III
DETAILS OF THE NOISY AERO-ENGINE BEVEL GEAR DATASET

Domain	Health state	Label	Training samples	Testing samples
ID	H	0	400	100
ID	SW	1	400	100
ID	BT	2	400	100
ID	SEC	3	400	100
Noise	H	0	/	100
Noise	SW	1	/	100
Noise	BT	2	/	100
Noise	SEC	3	/	100

After using our SAEU algorithm in mixed testing dataset, we obtained the uncertainty decomposition graphs similar to Fig. 8. This is because we still use the bearing inner race fault samples as OOD samples, resulting in the higher aleatoric and epistemic uncertainties than ID samples. In the same way, we can use the 0.9 quantile of epistemic uncertainty within the

training dataset as threshold to achieve the OOD detection in UU-DLF. The results are shown in the confusion matrices in Fig. 12, and we have summarized the results in Table IV. After OOD detection, it is obvious that the model performances of different methods in the mixed testing dataset are improved. In the five repeated experiments, the best ACC, ODA and FAR in our UU-DLF can achieve 95.10%, 100.00% and 9.92%, respectively. From the perspective of traceability analysis, as the OOD sample remains unchanged for bearing faults, it is still classified as "OOD and NOISY". The existing model improvement suggestions are consistent with the previous section.

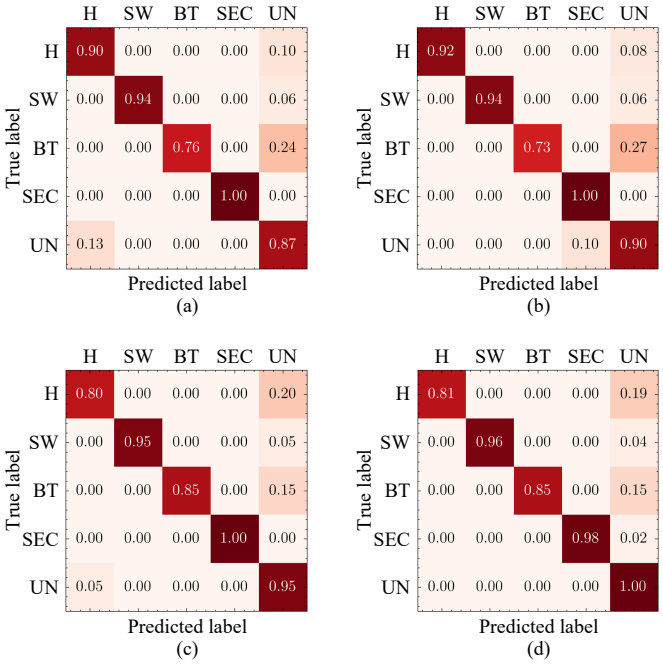


Fig. 12. Confusion matrixes of different methods in mixed testing dataset. BNN is used as a baseline for comparison. (a) BNN (baseline); (b) SAEU-MCD; (c) SAEU-DE; (d) SAEU-BS.

In order to further explore the reason why our proposed UU-DLF is effective, t-distribution stochastic neighbor embedding (t-SNE) is adopted to realize visualization. Differently from traditional feature visualization, the uncertainty is introduced as the background. Fig. 13 illustrates the visual comparison of the BNN and our proposed SAEU-BS. Note that each class of samples is balanced after sampling. It can be seen that our methods have better diagnostic separability due to the larger uncertainty intervals. The uncertainty provided by BNN ranges from 0 to 0.02, which can be confused with the training set. However, the uncertainty given by SAEU-BS ranges from 0 and 0.24. This significant difference between the training and testing sets can lead to better OOD detection performance.

C. Fault Diagnosis Experiment Using Human Knowledge

In order to demonstrate the effectiveness of UU-DLF using human knowledge to improve performance, we conducted fault diagnosis experiments in noisy environments. We still use aero engine level gear datasets in previous section, but the difference is that no additional bearing fault samples have been added. Random variance Gaussian noise is added into all samples included in the original testing dataset to form a new mixed testing dataset. The noise samples in the mixed testing dataset still maintain a 1 : 1 balance ratio with the original testing dataset. The specific experimental settings are shown in Table III.

The uncertainty decomposition graphs of testing dataset can be obtained by our SAEU algorithm, as shown in Fig. 14. It can be observed that the joint distribution of noise samples tends to be overall upward, that is, only increasing aleatoric uncertainty. However, for our SAEU algorithm, epistemic uncertainty also increases. We believed that this is due to the noise affecting the impact characteristics of gear faults, resulting in insufficient information on fault characteristics in the signal. The signal in time domain of SEC sample is shown in Fig. 15(a), and after adding noise, the time-domain signal is shown in the Fig. 15(b). It can be observed that the impact characteristics after adding noise are no longer obvious.

After the traceability analysis in our UU-DLF, we found that there were differences on the sources of uncertainty according to the rules. The SAEU-MCD and SAEU-BS determine it as "OOD and NOISY", while SAEU-DE determines it as only "NOISY". However, in any case, the analysis shows that the testing set contains too much noise. Thus, reducing the noise is undoubtedly the important directions of model improvement. Then, we use the Gaussian filtering to reduce the Gaussian noise according to the characteristics of the signal. Certainly, other methods can also be used, such as wavelet denoising, empirical mode decomposition denoising, etc. At this point, it is equivalent to embedding some human knowledge into the preprocessing of the existing model input under the guidance of uncertainty.

The uncertainty decomposition graphs after denoising can be obtained as shown in Fig. 16. For aleatoric uncertainty, it can be found that the marginal distributions of denoised samples and original samples have more overlapping areas, and the difference between the two epistemic uncertainty distributions is significantly reduced. As for the joint uncertainty distribution, the probability density centers of the denoised samples and the original samples are closer, compared with the noise samples. That is to say, the diagnosis of testing set is more trustworthy.

The average results of repeated experiments are summarized in Table V. Noise1 represents the accuracy of the mixed testing set, and Noise2 represents the accuracy after denoising. It can be seen that the model performance has indeed been improved after the introduction of human knowledge according to UU-DLF.

In this case, the traceability analysis of UU-DLF helps users to identify the noise in the environment and guide experts to remove noise for data preprocessing, thus establishing more trustworthy fault diagnosis. Through this experiment in noisy environment, we found an interesting thing: for the samples with domain shift, after embedding human knowledge, the uncertainty joint distribution is successfully aligned. Before our work, alignment in transfer learning was mostly embod-

TABLE IV
THE OOD DETECTION RESULTS OF THE AERO-ENGINE BEVEL GEAR DATASET

		CNN		BNN		SAEU-MCD		SAEU- DE		SAEU- BS	
		Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best
ID dataset	ACC	98.21	100.00	97.70	100.00	97.95	100.00	98.21	100.00	98.21	100.00
Mixed dataset	ACC	49.10	50.00	48.85	50.00	48.98	50.00	49.11	50.00	49.11	50.00
Mixed dataset using UU-DLF	ACC	N	N	87.34	88.23	89.76	89.89	92.09	92.45	92.71	**95.01**
	ODA	N	N	84.91	86.70	87.72	90.02	94.37	94.88	94.37	**100.00**
	FAR	N	N	10.01	9.94	10.25	10.00	10.02	9.89	10.00	**9.92**

* N represents that model has no OOD detection ability.

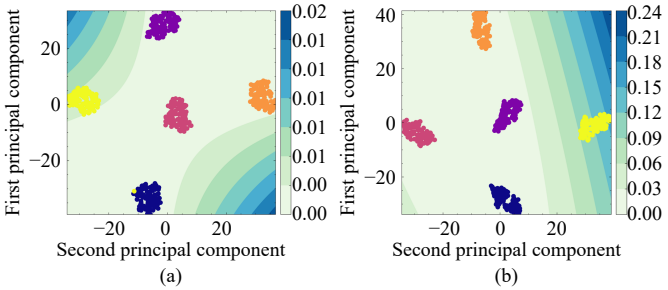


Fig. 13. Uncertainty visualization via t-SNE, where yellow points represent the OOD samples and are distributed in areas with larger uncertainty. (a) BNN; (b) SAEU-BS.

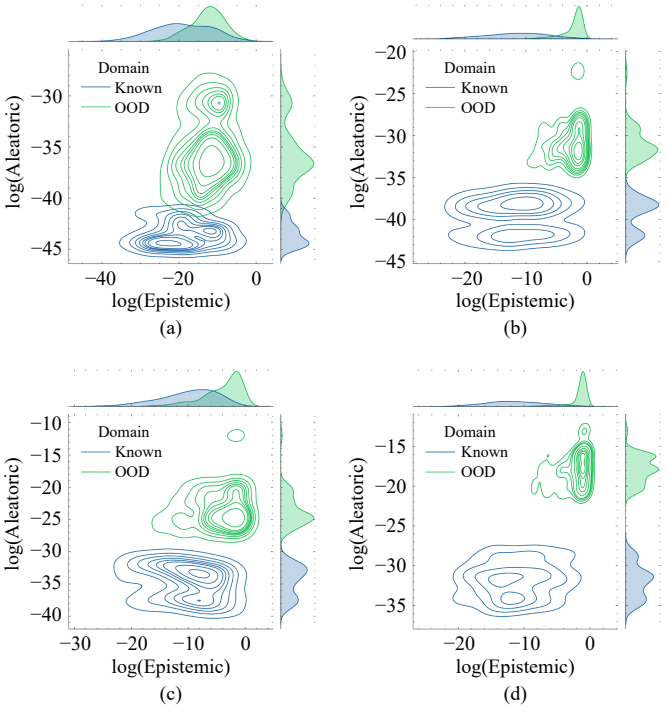


Fig. 14. Uncertainty decomposition graphs of different methods for the noise mixed testing dataset. BNN is used as a baseline for comparison. (a) BNN (baseline); (b) SAEU-MCD; (c) SAEU-DE; (d) SAEU-BS.

ied by feature visualization [55]–[59]. Now we provide an additional visualization method based on uncertainty decomposition graph, and this visualization has certain guidance for improving the model. And another interesting thing is that the

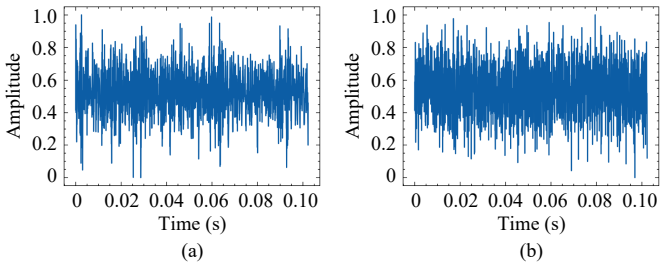


Fig. 15. Time-domain signals of SEC sample and SEC sample adding random Gaussian noise. (a) SEC sample; (b) SEC Sample with Gaussian noise.

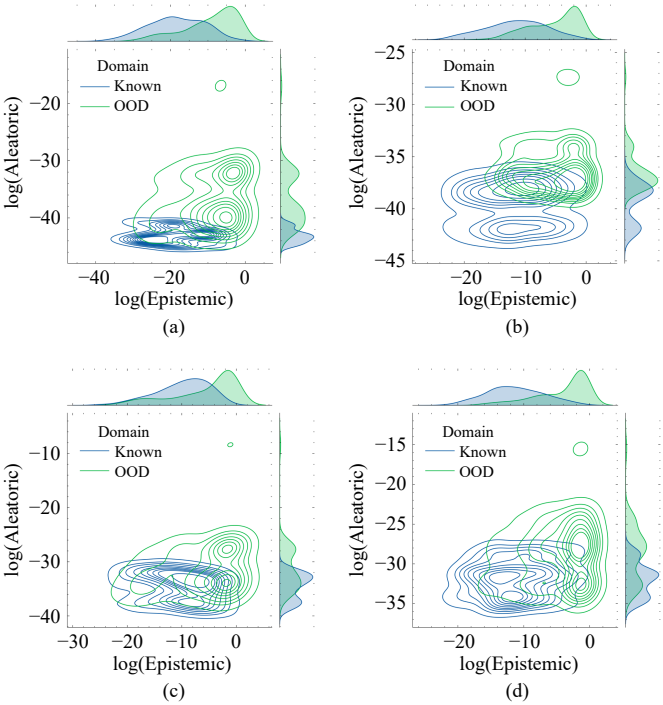


Fig. 16. Uncertainty decomposition graphs of different methods after denoising for the noise mixed testing dataset. BNN is used as a baseline for comparison. (a) BNN (baseline); (b) SAEU-MCD; (c) SAEU-DE; (d) SAEU-BS.

processing of OOD samples and domain shift samples is in the opposite direction on the uncertainty decomposition graph. That is, for OOD detection, the greater difference between the uncertainty distribution of the testing set and the training set, the better; for domain shift diagnosis, the more consistent distribution of two sets, the better.

TABLE V
THE RESULTS OF THE NOISY AERO-ENGINE BEVEL GEAR DATASET

	BNN	SAEU-MCD	SAEU-DE	SAEU-BS
ID	97.70	97.95	98.21	98.21
Noise1	62.50	64.03	63.01	**68.87**
Noise2	66.84	75.26	73.72	**80.86**

VI. CONCLUSION

This paper first explored a simplified method for quantifying uncertainty in deterministic networks. We employed MGD into the deep architecture and combined it with some methods of representing diversity to quantify both aleatoric and epistemic uncertainties simultaneously. Compared with the BNN and other methods, our SAEU algorithm effectively reduces the computational complexity of uncertainty quantification and is compatible with existing mainstream models. And on the basis of uncertainty decomposition graph given by SAEU algorithm, we proposed a unified trustworthy fault diagnosis framework, named as UU-DLF. It gives the "black box" models a certain degree of post-hoc interpretability from two aspects. On the on hand, it can handle the abnormal situations that existing deep learning models cannot handle through human intervention; on the other hand, it can conduct traceability analysis based on uncertainty decomposition, identify the causes of uncertainty, and further improve existing models. At last, the effectiveness of UU-DLF has been validated through experiments on both open source and real-world datasets. Based on the powerful performance and promising prospects shown by experiments, it may provide a promising way for deep learning models to gain industrial users' trust.

However, limited on the current representation methods of model diversity, additional storage resources are required to store multiple model parameters. So, it is necessary to study methods that only require one forward propagation of the network to achieve diversity and quantify uncertainty. And due to the relatively directness of existing threshold rules, it is possible to consider using composite threshold rules to further improve detection performance in the future. Furthermore, even if OOD samples are detected, existing models are unable to incorporate new faults into the network, meaning that the model cannot dynamically update the detected OOD samples. Further research should also focus on incremental learning in safety-critical applications. Finally, our current work mainly focuses on vibration signals, but we expect to validate the effectiveness of our methodology in more diverse research topics, such as thermal and acoustic signals.

REFERENCES

[1] S. Sankararaman, "Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction," *Mech. Syst. Signal Process.*, vol. 52–53, pp. 228–247, Feb. 2015.

[2] A. White, A. Karimoddini, and M. Karimadini, "Resilient fault diagnosis under imperfect observations–A need for Industry 4.0 era," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 5, pp. 1279–1288, Sep. 2020.

[3] Z. Zhu, Y. Lei, G. Qi, Y. Chai, N. Mazur, Y. An, and X. Huang, "A review of the application of deep learning in intelligent fault diagnosis of rotating machinery," *Measurement*, vol. 206, p. 112346, Jan. 2023.

[4] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems–Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, no. 1-2, pp. 314–334, Jan. 2014.

[5] Y. Hu, X. Miao, Y. Si, E. Pan, and E. Zio, "Prognostics and health management: A review from the perspectives of design, development and decision," *Reliab. Eng. Syst. Saf.*, vol. 217, p. 108063, Jan. 2022.

[6] R. Flage, T. Aven, E. Zio, and P. Baraldi, "Concerns, challenges, and directions of development for the issue of representing uncertainty in risk assessment," *Risk Anal.*, vol. 34, no. 7, pp. 1196–1207, Jul. 2014.

[7] J. Long, H. Wang, P. Li, and H. Fan, "Applications of fractional lower order time-frequency representation to machine bearing fault diagnosis," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 4, pp. 734–750, 2017.

[8] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.

[9] Z. Zhao, T. Li, J. Wu, C. Sun, S. Wang, R. Yan, and X. Chen, "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," *ISA Trans.*, vol. 107, pp. 224–255, Dec. 2020.

[10] A. Kumar, A. Glowacz, H. Tang, and J. Xiang, "Knowledge addition for improving the transfer learning from the laboratory to identify defects of hydraulic machinery," *Eng. Appl. Artif. Intell.*, vol. 126, p. 106756, Nov. 2023.

[11] Z. He, H. Shao, Z. Ding, H. Jiang, and J. Cheng, "Modified deep autoencoder driven by multisource parameters for fault transfer prognosis of aeroengine," *IEEE Trans. Ind. Electron.*, vol. 69, no. 1, pp. 845–855, Jan. 2022.

[12] D.-T. Hoang and H.-J. Kang, "A survey on deep learning based bearing fault diagnosis," *Neurocomputing*, vol. 335, pp. 327–335, Mar. 2019.

[13] M. Ma and Z. Mao, "Deep-convolution-based LSTM network for remaining useful life prediction," *IEEE Trans. Ind. Inf.*, vol. 17, no. 3, pp. 1658–1667, Mar. 2021.

[14] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, "Deep learning for smart manufacturing: Methods and applications," *J. Manuf. Syst.*, vol. 48, pp. 144–156, Jul. 2018.

[15] X. Wang, X. Liu, and Y. Li, "An incremental model transfer method for complex process fault diagnosis," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 5, pp. 1268–1280, Sep. 2019.

[16] Y. Zhang, K. Yu, Z. Lei, J. Ge, Y. Xu, Z. Li, Z. Ren, and K. Feng, "Integrated intelligent fault diagnosis approach of offshore wind turbine bearing based on information stream fusion and semi-supervised learning," *Expert Syst. Appl.*, vol. 232, p. 120854, Dec. 2023.

[17] Q. Qian, Y. Qin, J. Luo, Y. Wang, and F. Wu, "Deep discriminative transfer learning network for cross-machine fault diagnosis," *Mech. Syst. Signal Process.*, vol. 186, p. 109884, Mar. 2023.

[18] H. Shao, W. Li, B. Cai, J. Wan, Y. Xiao, and S. Yan, "Dual-threshold attention-guided GAN and limited infrared thermal images for rotating machinery fault diagnosis under speed fluctuation," *IEEE Trans. Ind. Inf.*, vol. 19, no. 9, pp. 9933–9942, Sep. 2023.

[19] P. Shi, S. Wu, X. Xu, B. Zhang, P. Liang, and Z. Qiao, "TSN: A novel intelligent fault diagnosis method for bearing with small samples under variable working conditions," *Reliab. Eng. Syst. Saf.*, vol. 240, p. 109575, Dec. 2023.

[20] S. Tang, Y. Zhu, and S. Yuan, "A novel adaptive convolutional neural network for fault diagnosis of hydraulic piston pump with acoustic images," *Adv. Eng. Inf.*, vol. 52, p. 101554, Apr. 2022.

[21] A. Glowacz, "Thermographic fault diagnosis of shaft of BLDC motor," *Sensors*, vol. 22, no. 21, p. 8537, Nov. 2022.

[22] A. Glowacz, "Thermographic fault diagnosis of electrical faults of commutator and induction motors," *Eng. Appl. Artif. Intell.*, vol. 121, p. 105962, May 2023.

[23] A. Choudhary, R. K. Mishra, S. Fatima, and B. K. Panigrahi, "Multi-input CNN based vibro-acoustic fusion for accurate fault diagnosis of induction motor," *Eng. Appl. Artif. Intell.*, vol. 120, p. 105872, Apr. 2023.

[24] Z. Feng, A. Gao, K. Li, and H. Ma, "Planetary gearbox fault diagnosis via rotary encoder signal analysis," *Mech. Syst. Signal Process.*, vol. 149, p. 107325, Feb. 2021.

[25] J. Jiao, M. Zhao, J. Lin, and J. Zhao, "A multivariate encoder information based convolutional neural network for intelligent fault diagnosis of planetary gearboxes," *Knowl. Based Syst.*, vol. 160,

pp. 237–250, Nov. 2018.

[26] T. Han, Y.-F. Li, and M. Qian, "A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, p. 3520011, Jun. 2021.

[27] G. Michau and O. Fink, "Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer," *Knowl. Based Syst.*, vol. 216, p. 106816, Mar. 2021.

[28] Z. Chen, G. He, J. Li, Y. Liao, K. Gryllias, and W. Li, "Domain adversarial transfer network for cross-domain fault diagnosis of rotary machinery," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 11, pp. 8702–8712, Nov. 2020.

[29] J. Jiao, M. Zhao, J. Lin, and K. Liang, "Residual joint adaptation adversarial network for intelligent transfer fault diagnosis," *Mech. Syst. Signal Process.*, vol. 145, p. 106962, Nov.–Dec. 2020.

[30] J. Huang, Z. Li, and Z. Zhou, "A simple framework to generalized zero-shot learning for fault diagnosis of industrial processes," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 6, pp. 1504–1506, Jun. 2023.

[31] B. Yang, S. Xu, Y. Lei, C.-G. Lee, E. Stewart, and C. Roberts, "Multi-source transfer learning network to complement knowledge for intelligent diagnosis of machines with unseen faults," *Mech. Syst. Signal Process.*, vol. 162, p. 108095, Jan. 2022.

[32] W. Zhang, X. Li, H. Ma, Z. Luo, and X. Li, "Open-set domain adaptation in machinery fault diagnostics using instance-level weighted adversarial learning," *IEEE Trans. Ind. Inf.*, vol. 17, no. 11, pp. 7445–7455, Nov. 2021.

[33] X. Yu, Z. Zhao, X. Zhang, Q. Zhang, Y. Liu, C. Sun, and X. Chen, "Deep-learning-based open set fault diagnosis by extreme value theory," *IEEE Trans. Ind. Inf.*, vol. 18, no. 1, pp. 185–196, Jan. 2022.

[34] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, Dec. 2021.

[35] 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.*, vol. 218, p. 108119, Feb. 2022.

[36] O. Fink, Q. Wang, M. Svensén, P. Dersin, W.-J. Lee, and M. Ducoffe, "Potential, challenges and future directions for deep learning in prognostics and health management applications," *Eng. Appl. Artif. Intell.*, vol. 92, p. 103678, Jun. 2020.

[37] T. Zhou, L. Zhang, T. Han, E. L. Droguett, A. Mosleh, and F. T. S. Chan, "An uncertainty-informed framework for trustworthy fault diagnosis in safety-critical applications," *Reliab. Eng. Syst. Saf.*, vol. 229, p. 108865, Jan. 2023.

[38] Y. Xiao, H. Shao, M. Feng, T. Han, J. Wan, and B. Liu, "Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer," *J. Manuf. Syst.*, vol. 70, pp. 186–201, Oct. 2023.

[39] J. Xia, M. Xu, H. Zhang, J. Zhang, W. Huang, H. Cao, and S. Wen, "Robust face alignment via inherent relation learning and uncertainty estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10358–10375, Aug. 2023.

[40] J. M. Dolezal, A. Srisuwananukorn, D. Karpeyev, S. Ramesh, S. Kochanny, B. Cody, A. S. Mansfield, S. Rakshit, R. Bansal, M. C. Bois, A. O. Bungum, J. J. Schulte, E. E. Vokes, M. C. Garassino, A. N. Husain, and A. T. Pearson, "Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology," *Nat. Commun.*, vol. 13, no. 1, p. 6572, Nov. 2022.

[41] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3139–3153, Jun. 2022.

[42] X. Li, J. Liu, B. Liu, Q. Zhang, K. Li, Z. Dong, and L. Mou, "Impacts of data uncertainty on the performance of data-driven-based building fault diagnosis," *J. Build. Eng.*, vol. 43, p. 103153, Nov. 2021.

[43] T. Han and Y.-F. Li, "Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles," *Reliab. Eng. Syst. Saf.*, vol. 226, p. 108648, Oct. 2022.

[44] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. 32nd Int. Conf. Neural Information Processing Systems*, Montreal, Canada, 2018, pp. 3183–3193.

[45] H. Zhou, W. Chen, L. Cheng, D. Williams, C. W. De Silva, and M. Xia, "Reliable and intelligent fault diagnosis with evidential VGG neural networks," *IEEE Trans. Instrum. Meas.*, vol. 72, p. 3508612, Feb. 2023.

[46] H. Zhou, W. Chen, L. Cheng, J. Liu, and M. Xia, "Trustworthy fault diagnosis with uncertainty estimation through evidential convolutional neural networks," *IEEE Trans. Ind. Inf.*, vol. 19, no. 11, pp. 10842–10852, Nov. 2023.

[47] N. Meinert, J. Gawlikowski, and A. Lavin, "The unreasonable effectiveness of deep evidential regression," in *Proc. 37th AAAI Conf. Artificial Intelligence*, Washington, USA, 2023, pp. 9134–9142.

[48] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France, 2015, pp. 1613–1622.

[49] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, Univ. Cambridge, Cambridge, UK, 2016.

[50] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Machine Learning*, New York, USA, 2016, pp. 1050–1059.

[51] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. 31st Conf. Neural Information Processing Systems*, Long Beach, USA, 2017, pp. 6405–6416.

[52] W. Xie, T. Han, Z. Pei, and M. Xie, "A unified out-of-distribution detection framework for trustworthy prognostics and health management in renewable energy systems," *Eng. Appl. Artif. Intell.*, vol. 125, p. 106707, Oct. 2023.

[53] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, and C. W. Coley, "Uncertainty quantification using neural networks for molecular property prediction," *J. Chem. Inf. Model.*, vol. 60, no. 8, pp. 3770–3780, Jul. 2020.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016, pp. 770–778.

[55] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Inf.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019.

[56] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," *IEEE Trans. Artif. Intell.*, vol. 1, no. 2, pp. 151–166, Oct. 2020.

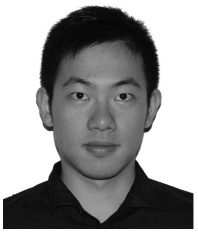
[57] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, Jan. 2021.

[58] W. Zhang and D. Wu, "Manifold embedded knowledge transfer for brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1117–1127, May 2020.

[59] X. Wang, R. Yang, and M. Huang, "An unsupervised deep-transfer-learning-based motor imagery EEG classification scheme for brain-computer interface," *Sensors*, vol. 22, no. 6, p. 2241, Mar. 2022.

Jiaxin Ren received the B.S. degree in Tsien Hsue-Shen Honor Class from Xi'an Jiaotong University in 2021, He is currently pursuing the Ph.D. degree with mechanical engineering in the School of Mechanical Engineering, Xi'an Jiaotong University. His fields of interests include trustworthy deep learning, transfer learning, and fault diagnosis.

Jingcheng Wen received the B.S. degree in mechanical engineering from Xi'an Jiao tong University in 2022. He is currently pursuing the Ph.D. degree in mechanical engineering in the School of Mechanical Engineering, Xi'an Jiaotong University. His current research is focused on deep learning and intelligent health monitoring for machines.



Zhibin Zhao (Member, IEEE) received the B.S. degree in Tsien Hsue-Shen Honor Class from Xi'an Jiaotong University in 2015, and the Ph.D. degree in mechanical engineering from Xi'an Jiaotong Universit in 2020. He was also a visiting Ph.D. student in AI for healthcare at the University of Manchester, UK, from 2019 to 2020. He is now a Lecturer in mechanical engineering in the School of Mechanical Engineering, Xi'an Jiaotong University. He is an Associate Editor of *IEEE Transactions on Instrumentation and Measurement*. His current research is focused on sparse signal processing and machine learning algorithms for machinery health monitoring and healthcare.

Ruqiang Yan (Fellow, IEEE) received the M.S. degree in precision instrument and machinery from the University of Science and Technology of China in 2002, and the Ph.D. degree in mechanical engineering from the University of Massachusetts at Amherst, USA, in 2007.

He was a Guest Researcher at the National Institute of Standards and Technology (NIST) in 2006–2008 and a Professor with the School of Instrument Science and Engineering, Southeast University, from 2009 to 2018. He joined the School of Mechanical Engineering, Xi'an Jiaotong University, in 2018. His research interests include data analytics, AI, and energy-efficient sensing and sensor networks for the condition monitoring and health diagnosis of large-scale, complex, dynamical systems.

Dr. Yan is a Fellow of ASME (2019). His honors and awards include the IEEE Instrumentation and Measurement Society Technical Award in 2019 and Outstanding Service Award in 2022, and multiple best paper awards, such as Andrew P. Sage Best Transactions Paper Award. Dr. Yan serves as the Editor-in-Chief of the *IEEE Transactions on Instrumentation and Measurement*. He is also an Associate Editor-in-Chief of *Chinese Journal of Mechanical Engineering*.

Xuefeng Chen (Member, IEEE) is a Full Professor and Dean of School of Mechanical Engineering in Xi'an Jiaotong University, where he received the Ph.D. degree in 2004. He works as the Executive Director of the Fault Diagnosis Branch in China Mechanical Engineering Society. Besides, he is also a Member of ASME and IEEE, and the Chair of IEEE the Xian and Chengdu Joint Section Instrumentation and Measurement Society Chapter.

He has authored over 100 SCI publications in areas of composite structure, aero-engine, wind power equipment, etc. He won National Excellent Doctoral Thesis Award in 2007, First Technological Invention Award of Ministry of Education in 2008, Second National Technological Invention Award in 2009, First Provincial Teaching Achievement Award in 2013, First Technological Invention Award of Ministry of Education in 2015, and he was awarded as Science & Technology Award for Chinese Youth in 2013. Additionally, he hosted a National Key 973 Research Program of China as principal scientist in 2015.

Asoke K. Nandi (Fellow, IEEE) received the Ph.D. degree in physics from the University of Cambridge (Trinity College). He held academic positions in several universities, including Oxford, Imperial College London, Strathclyde, and Liverpool as well as Finland Distinguished Professorship. In 2013 he moved to Brunel University London.

In 1983, Professor Nandi co-discovered the three fundamental particles known as W+, W− and Z0, providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Committee for Physics in 1984 awarded the prize to two of his team leaders for their decisive contributions. His current research interests lie in signal processing and machine learning, with applications to machine health monitoring, functional magnetic resonance data, gene expression data, communications, and biomedical data. He made fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has much expertise in "Big Data". Professor Nandi has authored over 650 technical publications, including 300 journal papers as well as six books, entitled Image Segmentation: Principles, Techniques, and Applications (Wiley, 2022), Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines (Wiley, 2020), Automatic Modulation Classification: Principles, Algorithms and Applications (Wiley, 2015), Integrative Cluster Analysis in Bioinformatics (Wiley, 2015), Blind Estimation Using Higher-Order Statistics (Springer, 1999), and Automatic Modulation Recognition of Communications Signals (Springer, 1996). The H-index of his publications is 85 (Google Scholar) and ERDOS number is 2. Professor Nandi is a Fellow of the Royal Academy of Engineering and a Fellow of six other institutions including the IEEE. In 2023, he has been honoured by the Academia Europaea and the Academia Scientiarum et Artium Europaea. He has received many awards, including the IEEE Heinrich Hertz Award in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers in 1999, and the Mountbatten Premium of the Institution of Electrical Engineers in 1998. Professor Nandi is an IEEE Distinguished Lecturer (EMBS, 2018–2019).