This is a peer-reviewed, accepted author manuscript of the following research article: Xiao, Y., Shao, H., Feng, M., Han, T., Wan, J., & Liu, B. (2023). Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer. *Journal of Manufacturing Systems, 70*, 186-201. https://doi.org/10.1016/j.jmsy.2023.07.012

Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in Transformer

Yiming Xiao[1], Haidong Shao[1*], Minjie Feng[1], Te Han[2], Jiafu Wan[3], Bin Liu[4]

1. College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, China

2. School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

3. Provincial Key Laboratory of Technique and Equipment for Macromolecular Advanced Manufacturing, South China University of Technology, Guangzhou 510641, China

4. Department of Management Science, University of Strathclyde, Glasgow G1 1XQ, UK

Corresponding author: Haidong Shao (hdshao@hnu.edu.cn)

Abstract: To enable researchers to fully trust the decisions made by deep diagnostic models, interpretable rotating machinery fault diagnosis (RMFD) research has emerged. Existing interpretable RMFD research focuses on developing interpretable modules embedded in deep models to assign physical meaning to results, or on inferring the logic of the model to make decisions based on results. However, there is limited work on how to quantify uncertainty in results and explain its sources and composition. Uncertainty quantification and decomposition not only provide the confidence of the results, but also identify the source of unknown factors in the data, and consequently guide to enhance the interpretability and trustworthiness of models. Therefore, this paper proposes to use Bayesian variational learning to introduce uncertainty into the attention weights of Transformer to construct a probabilistic Bayesian Transformer for trustworthy RMFD. A probabilistic attention is designed and the corresponding optimization objective is defined, which can infer the prior and variational posterior distributions of attention weights, thus empowering the model to perceive uncertainty. An uncertainty quantification and decomposition scheme is developed to achieve confidence characterization of results and separation of epistemic and aleatoric uncertainty. The effectiveness of the proposed method is fully verified in three out-of-distribution generalization scenarios.

Keywords: trustworthy rotating machinery fault diagnosis; probabilistic attention; Bayesian deep learning; Transformer; uncertainty quantification and decomposition.

1. Introduction

Rotating machinery plays an indispensable role in modern manufacturing. Bearing and gearbox are the key components of rotating machinery and their failure can cause serious economic losses or even endanger lives. Therefore, rotating machinery fault diagnosis (RMFD) for bearing and gearbox is of great importance [1].

In recent years, RMFD research based on deep learning (DL) has continuously received attention from scholars worldwide. In 2023, Chen *et al*. [2] designed a dual adversarial guided unsupervised multi-domain adaptation network that achieves multi-domain collaborative RMFD by fully extracting domain invariant features from multiple domains. In 2023, a generalized model-agnostic meta-learning model is proposed by Lin *et al*. [3] to enable few-shot RMFD cross various operating conditions driven by

heterogeneous signals. Although these advanced deep diagnostic models have shown superior performance, they inherit the black-box nature of DL, which makes it difficult for researchers to fully trust the diagnostic results they provide and limits the promotion and application of intelligent diagnostic methods [4]-[7].

To explain the rationale behind the diagnostic decisions made by deep models so that researchers can trust them more, interpretable RMFD studies have emerged. In 2022, Xiao *et al*. [8] proposed an unsupervised domain adaptation RMFD method from the simulation domain to the experimental domain, exploring a new way of "data-physical" coupling-driven fault diagnosis. In 2022, Li *et al*. [9] developed a continuous wavelet convolution layer and used it to improve the traditional convolutional neural networks (CNN), which gave CNN interpretability through the physical meaning of wavelet transform. In 2023, Shang *et al*. [10] designed a denoising fault-aware wavelet network using the interpretability and noise reduction properties of signal processing methods, which enabled efficient RMFD in a noisy background. Despite the popularity of the interpretable RMFD research, the existing methods mainly focus on developing interpretable modules and embedding them into deep models to assign some physical meaning to the diagnosis results, or on inferring the deep logic of the model to make decisions based on the results. There has been limited work on how to quantify uncertainty in diagnostic results and explain its sources and composition, which is essential for constructing trustworthy diagnostic models and establishing a dependency relationship between researchers and deep models.

Uncertainty quantification is a powerful tool that can be utilized to determine the confidence level of diagnostic results. High uncertainty indicates low confidence in the results, while low uncertainty implies high confidence in the results. In practical engineering, complicated mechanical structures and failure mechanisms can cause unknown failures, and noisy backgrounds can create unknown data acquisition environments. In addition, variable equipment speeds and loads can constitute unknown operating conditions, causing the distribution of test data unknown and significantly different from the known distribution of training data [11], [12]. When dealing with unknown out-of-distribution samples, deep models usually make untrustworthy diagnostic decisions without warning the researchers. However, if the uncertainty in the diagnostic results can be quantified, the researchers can clarify the confidence of such results and thus adjust the equipment operation and maintenance strategy to avoid potential failure risks. In addition, by explaining the source and composition of uncertainty in the results, the dependency relationship between the researchers and the deep models can be established, enhancing the transparency and interpretability of the diagnostic process [13], [14].

Uncertainty can be decomposed into epistemic uncertainty (also known as model uncertainty) and aleatoric uncertainty (also known as data uncertainty). (1) Epistemic uncertainty refers to the uncertainty in the model parameters caused by insufficient diagnostic knowledge, such as limited training data or imbalanced data that fail to cover all possible operating conditions, fault types, etc. Epistemic uncertainty can be described by the posterior distribution $p(\theta \mid D)$ of the parameters θ learned by the model given a training dataset D, where a flat posterior reflects high epistemic uncertainty and a peaked posterior characterizes low epistemic uncertainty. To reduce epistemic uncertainty, additional simulations are needed to collect richer data to help the model acquire knowledge for understanding the behavior of mechanical systems under unknown failure modes or operating conditions [15], [16]. (2) Aleatoric

uncertainty refers to the inherent randomness of data influenced by unobserved factors, such as noise interference, sensor hardware damage, etc. Aleatoric uncertainty can be described by the probability distribution of predicted label output by the model when given a set of deterministic model parameters and an input. A flat distribution indicates that the model cannot confidently assign input to a class and therefore has high aleatoric uncertainty, while a peaked distribution indicates that the model has sufficient confidence to predict the label of the input and therefore has low aleatoric uncertainty. Randomness is an inherent part of data, and adding more training data does not have an impact on aleatoric uncertainty. However, the use of more reliable and efficient instrumentation can help to capture crucial unknown variables that are hidden within the data, thereby reducing aleatoric uncertainty [17], [18]. Therefore, uncertainty quantification and decomposition can not only provide the confidence of diagnostic results, but also aid to analyze the source of unknown factors in test data and identify the measures to improve the credibility of the model, which is significant to achieve trustworthy RMFD.

However, the existing deep diagnostic models usually are not able to correctly express the uncertainty in the diagnostic results. The reason is that their model parameters are generally fixed values, making them only give overconfident point estimate predictions [19]. In contrast, Bayesian DL treats model parameters as random variables obeying some probability distribution rather than fixed values, which is a powerful tool for quantifying uncertainty [20]. In 2022, Zhou *et al*. [21] constructed a Bayesian CNN model for trustworthy fault diagnosis, which is the first work in the field of RMFD to explore and explain the sources and composition of uncertainty in diagnostic results obtained from deep models. It is worth clarifying that although relevant studies [22], [23] on how to consider the uncertainty in diagnostic results have been conducted prior to [21], these studies are still limited to applying uncertainty to improve diagnostic accuracy, without exploring the sources and composition of uncertainty in a deeper way. Therefore, this type of study is not appropriate for trustworthy RMFD. Considering the application prospects shown by the model Transformer based on self-attention mechanism in recent years [24], this paper proposes to use Bayesian variational learning to introduce uncertainty into the attention weights of Transformer to construct a **pro**babilistic Bayesian Trans**former** (ProFormer) for trustworthy RMFD. In the proposed model, the attention weights are no longer deterministic values directly obtained by calculation as in the traditional Transformer, but random variables sampled from the learned probability distributions. The main innovations and contributions of the proposed method are as follows:

(1) A new method for trustworthy RMFD is proposed, which improves the trustworthiness of the model by analyzing and explaining the sources and composition of uncertainty in the diagnosis results. The effectiveness of the proposed method is fully verified in out-of-distribution generalization scenarios with test data containing samples of unknown fault classes, unknown noise levels, or unknown operating conditions. This study is an important exploration in the current field of interpretable RMFD.

(2) A probabilistic attention is designed and the corresponding optimal objective function is defined, which as the core of ProFormer can model the prior and variational posterior distributions of attention weights, thus empowering the model to perceive uncertainty. This is a pioneering work to build an attention mechanism in a Bayesian DL framework.

(3) An uncertainty quantification and decomposition scheme is developed to achieve confidence characterization of diagnostic results and separation of epistemic and aleatoric uncertainty.

The remainder of this article is organized as follows: Section 2 briefly reviews the multi-head self-attention mechanism and Bayesian variational learning. Section 3 presents the proposed ProFormer model in detail. Section 4 describes the experimental setup and analyzes the diagnostic results. The conclusions and future research prospects are discussed in Section 5.

2. Preliminaries

2.1 Multi-head self-attention mechanism

The multi-head self-attention mechanism is the core of Transformer [25], which aims to learn an alignment that causes each token in token embeddings to aggregate information contained in other tokens. Given token embeddings $X \in \mathbb{R}^{m\times d_m}$, queries $Q = XW^Q \in \mathbb{R}^{m\times d_k}$, keys $K = XW^K \in \mathbb{R}^{m\times d_k}$, and values $V = XW^V \in \mathbb{R}^{m\times d_v}$ can be obtained through a group of linear projections, in which m is the number of tokens, and $W^Q \in \mathbb{R}^{d_m\times d_k}$, $W^K \in \mathbb{R}^{d_m\times d_k}$ and $W^V \in \mathbb{R}^{d_m\times d_v}$ are parameter matrices that the model needs to learn. As shown in the left figure of **Fig. 1**, Q, K, and V are the input for the scaled dot-product attention. In this attention, Q and K first conduct the dot-product operation and then are divided by a scaling constant $(d_k)^{1/2}$ to obtain unnormalized attention weights φ:

$$\varphi = f_{dot}(Q,K) = QK^T \big/ (d_k)^{1/2} \in \mathbb{R}^{m\times m} \tag{1}$$

Subsequently, the normalized attention weights can be acquired by regularizing φ across the key dimension using a softmax function:

$$A_{i,j} = \mathrm{softmax}(\varphi_{i,j}) = \exp(\varphi_{i,j}) \Big/ \sum_{j'=1}^{m} \varphi_{i,j'} \tag{2}$$

where i=1, …, m, j=1, …, m. Finally, the output of the scaled dot-production attention $O = AV \in \mathbb{R}^{m\times d_v}$ can be obtained by the dot-production operation of A and V.

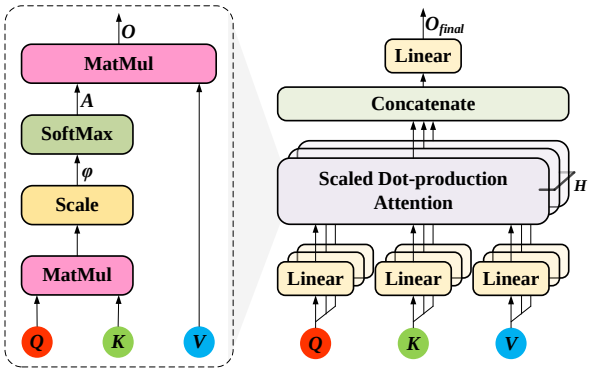

Fig. 1. Multi-head self-attention mechanism.

In practice, a single attention calculation for Q, K, and V often results in poor model performance, thus motivating the multi-head self-attention mechanism shown in the right figure of **Fig. 1**, which consists mainly of multiple attention heads running in parallel. The multi-head self-attention mechanism firstly processes the original Q, K, and V through H different groups of linear projections to obtain H different versions of Q, K, and V, then performs the above scaled dot-product attention operation on each

version of the *Q*, *K*, and *V* to obtain *H* outputs. Finally, these *H* outputs are concatenated and projected linearly again to obtain the final output O_{final}:

$$\begin{aligned} O_{final} &= \text{Concat}(O_1, O_2, ..., O_H)W^O \in \mathbb{R}^{m \times d_m} \\ \text{where } O_h &= \text{softmax}(f_{dot}(Q_h, K_h))V_h \\ Q_h &= QW_h^Q, K_h = KW_h^K, V_h = VW_h^V \end{aligned} \tag{3}$$

where $\text{Concat}(\cdot)$ is the concatenate function, the subscript h ($1 \le h \le H$) indicates the hth attention head, and $W_h^Q \in \mathbb{R}^{d_m \times d_k}$, $W_h^K \in \mathbb{R}^{d_m \times d_k}$, $W_h^V \in \mathbb{R}^{d_m \times d_v}$, and $W^O \in \mathbb{R}^{Hd_v \times d_m}$ are parameter matrices that the model needs to learn. Generally, $d_k = d_v = d_m/H$ in Transformer model.

2.2 Bayesian variational learning

The training goal of deep neural networks (DNN) is to find the optimal model parameters with respect to training data, and each optimal parameter is only a point estimate of that parameter. Thus, the model parameters of a trained DNN are deterministic, and only one fixed output can be provided for a given input. Instead of giving point estimates of these parameters, Bayesian neural networks (BNN) [19] provides the probability distribution for all parameters, i.e., the posterior distribution $p(\theta \mid D)$. Generally, Bayes' rule can be used to solve the posterior distribution:

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)} = \frac{p(D \mid \theta)\, p(\theta)}{\int_\theta p(D \mid \theta)\, p(\theta) d\theta} \tag{4}$$

where $p(D \mid \theta)$ is the likelihood, $p(D)$ is the marginal, and the prior $p(\theta)$ is usually chosen as a Gaussian distribution. However, since neural networks usually contain a large number of parameters, making the calculation of $\int_\theta p(D \mid \theta) p(\theta) d\theta$ tricky, therefore, variational inference is required to capture an approximate distribution of this posterior distribution [26].

The goal of variational inference is to find a distribution in a family of distributions q_ϕ controlled by parameters ϕ that approximates the true posterior distribution as closely as possible, thus transforming the above posterior inference problem into an optimization problem, i.e., minimizing the discrepancy between the variational distribution $q_\phi(\theta)$ and the true posterior distribution $p(\theta \mid D)$. Typically, Kullback-Leibler (KL) divergence is chosen as a measure:

$$\begin{aligned} KL(q_\phi(\theta) \,||\, p(\theta \mid D)) &= \int q_\phi(\theta) \log \frac{q_\phi(\theta)}{p(\theta \mid D)} d\theta \\ &= \int q_\phi(\theta) \log \frac{q_\phi(\theta) p(D)}{p(D \mid \theta) p(\theta)} d\theta \\ &= \log p(D) - \underbrace{\int q_\phi(\theta) \log \frac{p(D \mid \theta) p(\theta)}{q_\phi(\theta)} d\theta}_{L(D,\theta)} \end{aligned} \tag{5}$$

where $\log p(D)$ and $L(D, \theta)$ are referred as evidence and evidence lower bound (ELBO),

respectively.

Since $\log p(D)$ is a computationally difficult constant, it is worthwhile to transform the above problem of minimizing $KL(q_\phi(\theta) \| p(\theta \mid D))$ into maximizing ELBO, which can be further defined as:

$$\begin{aligned} L(D,\theta) &= \int q_\phi(\theta) \log p(D \mid \theta) d\theta - \int q_\phi(\theta) \log \frac{q_\phi(\theta)}{p(\theta)} d\theta \\ &= \underbrace{\mathbb{E}_{q_\phi(\theta)}\left[\log p(D \mid \theta)\right]}_{L_1} - \underbrace{KL(q_\phi(\theta) \| p(\theta))}_{L_2} \end{aligned} \tag{6}$$

where L_1 is the likelihood cost, which represents how well the model fits the data, and L_2 is the complexity cost, which represents the similarity between the variational posterior and the given prior. According to unbiased Monte Carlo estimation, ELBO can be approximate by drawing T times the model parameters from the variational posterior distribution $p(\theta \mid D)$:

$$\begin{aligned} L(D,\theta) &= \mathbb{E}_{q_\phi(\theta)}[\log p(D \mid \theta)] - \mathbb{E}_{q_\phi(\theta)}[\log q_\phi(\theta) - \log p(\theta)] \\ &\approx \sum_{t=1}^{T} \log p(D \mid \theta^{(t)}) - \log q_\phi(\theta^{(t)}) + \log p(\theta^{(t)}) \end{aligned} \tag{7}$$

where $\theta^{(t)}$ is the model parameters for the tth sampling. BNN is trained with $-L(D,\theta)$ as the optimization objective to obtain a variational posterior that is close to the true posterior.

3. Proposed probabilistic Bayesian Transformer

3.1 Model architecture

As shown in **Fig. 2**, the proposed ProFormer model is composed of a convolutional layer, a ProFormer encoder stacked by multiple ProFormer blocks, and a classifier stacked by multiple fully connected (FC) layers. Specifically, ProFormer block consists of the designed probabilistic attention, a multilayer perceptron (MLP), two Layer normalization layers, and two residual connections.

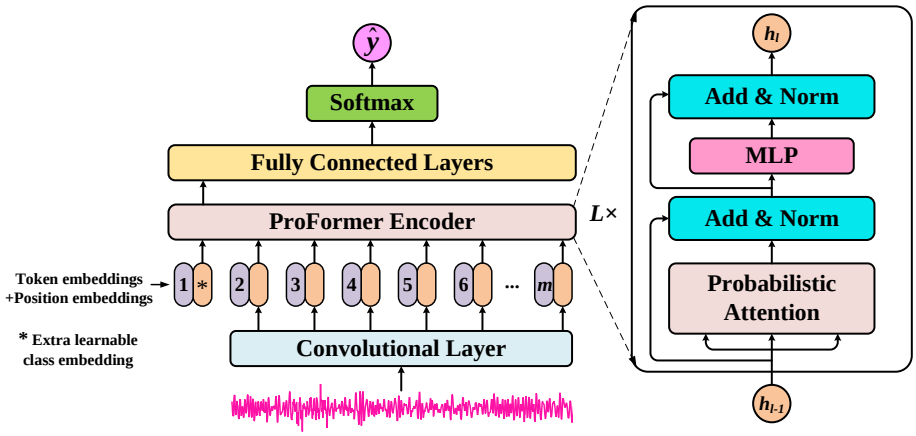

Fig. 2. Model architecture of ProFormer.

Given a one-dimensional vibration signal dataset $D := \{x_i, y_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{1\times s}$ represents the

*i*th sample labeled y_i, *s* is the sample length, and *N* is the number of samples. For ease of presentation, the data index *i* is omitted next. Each sample is first divided into $(m-1)$ signal segments, each segment is called a token, and then each token is linearly projected to an embedding with dimension d_m to obtain the token embeddings. This process can be implemented by the convolutional layer:

$$h_0'' = [\mathrm{Conv}(x, 1, d_m, \mathrm{kernel} = a, \mathrm{stride} = b)]^T \tag{8}$$

where $[\cdot]^T$ denotes transpose operation, $h_0'' \in \mathbb{R}^{(m-1)\times d_m}$ is the token embeddings, 1 and d_m represents the numbers of input and output channels, respectively, and *a* and *b* are the size and shift step of the convolutional kernel, respectively. Subsequently, a learnable embedding $E^{class} \in \mathbb{R}^{1\times d_m}$, called class token [27], is concatenated to the beginning of h_0'' to acquire token embeddings $h_0' \in \mathbb{R}^{m\times d_m}$. In addition, a learnable position embedding $E^{pos} \in \mathbb{R}^{m\times d_m}$ is required to be added to h_0':

$$h_0 = h_0' + E^{pos} \in \mathbb{R}^{m\times d_m} \tag{9}$$

where h_0 is the token embeddings with position information.

Subsequently, h_0 is fed to the encoder to acquire the hidden features of the sample:

$$\begin{aligned} h_l' &= \mathrm{LayerNorm}(\mathrm{Attention}(h_{l-1})) + h_{l-1} \\ h_l &= \mathrm{LayerNorm}(\mathrm{MLP}(h_l')) + h_l' \end{aligned} \tag{10}$$

where *l* ($1 \le l \le L$) represents the *l*th ProFormer block, and the transformed class token E_L^{class} in h_L can be extracted to be used as the hidden features for classification. The final classification process can be described as:

$$\hat{y} = \mathrm{softmax}(\mathrm{FC}(E_L^{class})) \in \mathbb{R}^{1\times C} \tag{11}$$

where $\hat{y}$ is the probabilistic distribution of the predicted label, and *C* is the number of fault classes.

3.2 Probabilistic attention design and optimization objective definition

The designed probabilistic attention is the core of the proposed ProFormer, which can replace the scaled dot-product attention to give Transformer similar properties to BNN. As can be observed in **Fig. 3**, in the proposed probabilistic attention, the attention weights are no longer deterministic values obtained by computation, but latent random variables sampled from the posterior distribution of the attention weights. However, we note that the attention weights are not directly equivalent to the model parameters; the attention weights should depend on the input, while the model parameters can be shared across all inputs. Thus, for dataset $D := \{x_i, y_i\}_{i=1}^N$, what needs to be modeled is the posterior distribution of attention weights for each input $p(A \mid x, y)$, in which $A = \{A_l\}_{l=1}^L$ is the normalized attention

weights for sample x from all ProFormer blocks. Furthermore, considering that using the family of Gaussian distributions as the variational posterior distribution $q_\phi(A \mid x, y)$ does not satisfy the constraints $A_{l,i,j} > 0$ and $\sum_j A_{l,i,j} = 1$, it is useful to meet this constraint by modelling the variational posterior distribution $q_\phi(\varphi \mid x, y)$ of the unnormalized weights $\varphi = \{\varphi_l\}_{l=1}^{L}$ using the family of Gaussian distributions and regularizing the sampled random unnormalized weights $\varphi \sim q_\phi(\varphi \mid x, y)$ using **Eq. (2)**.

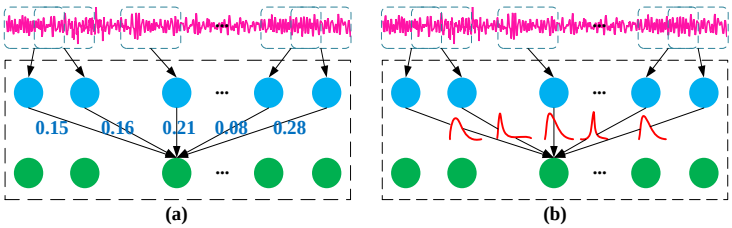

Fig. 3. The difference between scaled dot-production attention and probabilistic attention: (a) scaled dot-product attention; (b) probabilistic attention.

***1) Inference network for variational posterior distribution*:** According to mean-filed theory, $q_\phi(\varphi \mid x, y)$ can be decomposed into a product of L Gaussian distributions:

$$q_\phi(\varphi \mid x, y) = \prod_{l=1}^{L} \mathcal{N}(\varphi_l \mid \mu_l^0, (\sigma_l^0)^2) \tag{12}$$

where $\mu_l^0 \in \mathbb{R}^{m \times m}$ and $\sigma_l^0 \in \mathbb{R}^{m \times m}$ are the matrices consisting of the means and standard deviations of m^2 Gaussian distributions, respectively, and this indicates that each value $\varphi_{l,i,j}$ of φ_l is sampled from a separate Gaussian distribution $\mathcal{N}(\mu_{l,i,j}, (\sigma_{l,i,j})^2)$, as shown in **Fig. 3(b)**.

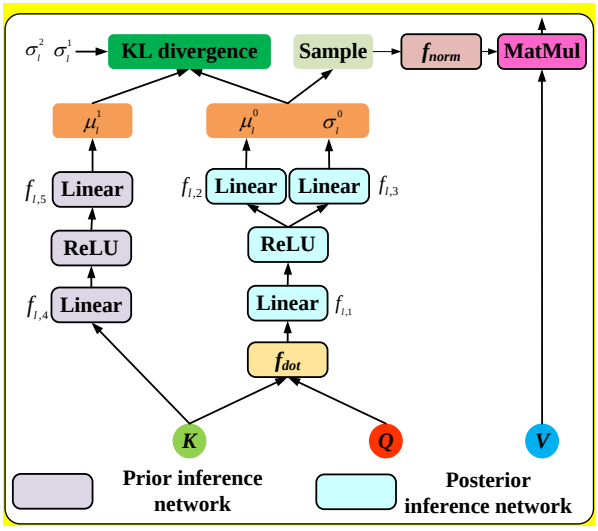

Fig. 4. Probabilistic attention in the *l*th ProFormer block.

To derive the two distribution parameters μ_l^0 and σ_l^0 of $\mathcal{N}(\varphi_l \mid \mu_l^0, (\sigma_l^0)^2)$, the probabilistic attention as shown in **Fig. 4** is designed. Similar to the scaled dot-product attention, the probabilistic attention of the *l*th ProFormer block also takes Q_l, K_l, V_l as inputs, and first obtains the unnormalized weights by $\varphi_l = f_{dot}(Q_l, K_l)$. Then, the posterior inference network consists of two MLPs in probabilistic attention is used to derive μ_l^0 and σ_l^0:

$$\begin{aligned} \mu_l^0 &= f_{l,2}(\mathrm{ReLU}(f_{l,1}(\varphi_l))) \\ \sigma_l^0 &= f_{l,3}(\mathrm{ReLU}(f_{l,1}(\varphi_l))) \end{aligned} \tag{13}$$

where $f_{l,1}(\cdot)$, $f_{l,2}(\cdot)$, $f_{l,3}(\cdot)$ represent the linear projections, and $\mathrm{ReLU}(\cdot)$ represents the activation function.

Considering that the action of sampling random weight $\varphi_{l,i,j}$ from $\mathcal{N}(\mu_{l,i,j}, (\sigma_{l,i,j})^2)$ is not differentiable and that the standard deviation must be a non-negative value, this sampling action is equated using the reparameterization trick in [28] as:

$$\varphi_{l,i,j} = g_\phi(\varepsilon_{l,i,j}) = \mu_{l,i,j} + \varepsilon_{l,i,j}(\ln(1+\exp(\sigma_{l,i,j}))) \tag{14}$$

where ε_l is a matric with the same shape as φ_l, whose elements are drawn from $\mathcal{N}(0,1)$.

Finally, the random normalized attention weights can be acquired by $A_l = \mathrm{softmax}(\varphi_l)$, and the output of the *l*th probabilistic attention is $O_l = A_l V_l$

***2) Inference network for prior distribution*:** To avoid overfitting, instead of directly giving a deterministic prior distribution $p(\varphi)$ of φ, we construct inference network with keys K_l as input to model $p(\varphi)$ as in [29], such that $p(\varphi)$ depends on the input *x*. However, if the mean and standard deviation of the Gaussian prior are obtained entirely using the inference network, it may not yield the prior that we expect. Therefore, this paper chooses to use the prior inference network in **Fig. 4** to derive the mean of the Gaussian prior and treat the standard deviation as a hyperparameter and give it directly. In addition, we construct a Gaussian mixture prior with a scaled mixture of two Gaussian distributions, in which the two Gaussian distributions share the derived mean but have different standard deviations. According to mean-filed theory, this prior can be defined as:

$$p(\varphi \mid x) = \prod_{l=1}^{L} (\alpha \mathcal{N}(\varphi_l \mid \mu_l^1, (\sigma_l^1)^2) + (1-\alpha)\mathcal{N}(\varphi_l \mid \mu_l^1, (\sigma_l^2)^2)) \tag{15}$$

where α is a regularization coefficient, $\mu_l^1 \in \mathbb{R}^{m \times m}$ is the matric consisting of m^2 derived shared means, and $\sigma_l^1 \in \mathbb{R}^{m \times m}$ and $\sigma_l^2 \in \mathbb{R}^{m \times m}$ are matrices whose elements are all the given standard deviations. The inference process of μ_l^1 can be described as:

$$\mu_l^1 = f_{l,5}(\mathrm{ReLU}(f_{l,4}(K_l))) \tag{16}$$

where $f_{l,4}(\cdot)$, $f_{l,5}(\cdot)$ represent the linear projections.

***3) Optimization objective definition*:** The model parameters of ProFormer are updated by minimizing $KL(q_\phi(\varphi) \| p(\varphi \mid x, y))$, which is equivalent to maximizing ELBO:

$$L(x, y) = \mathbb{E}_{q_\phi(\varphi)}[\log p(x, y \mid \varphi)] - KL(q_\phi(\varphi) \| p(\varphi)) \tag{17}$$

To weight the complexity cost relative to the likelihood cost, **Eq. (17)** can be rewritten as:

$$L(x, y) = \mathbb{E}_{q_\phi(\varphi)}[\log p(x, y \mid \varphi)] - \lambda KL(q_\phi(\varphi) \| p(\varphi)) \tag{18}$$

where λ is a regularization coefficient, and $KL(q_\phi(\varphi) \| p(\varphi))$ can be further expressed as:

$$\begin{aligned} KL(q_\phi(\varphi) \| p(\varphi)) &= \mathbb{E}_{q_\phi(\varphi)}[\log q_\phi(\varphi) - \log p(\varphi)] \\ &= \mathbb{E}_{q_\phi(\varphi)}[\log \prod_{l=1}^{L} q_\phi(\varphi_l) - \log \prod_{l=1}^{L} p(\varphi_l)] \\ &= \mathbb{E}_{q_\phi(\varphi)}[\sum_{l=1}^{L} \log q_\phi(\varphi_l) - \log p(\varphi_l)] \end{aligned} \tag{19}$$

Combining **Eq. (14)**, **Eq. (18)**, and **Eq. (19)**, the ELBO objective can be rewritten as:

$$\begin{aligned} L(x, y) &= \mathbb{E}_{q_\phi(\varphi)}[\log p(x, y \mid \varphi) - \lambda \sum_{l=1}^{L} \log q_\phi(\varphi_l) - \log p(\varphi_l)] \\ &= \mathbb{E}_{\varepsilon}[\log p(x, y \mid g_\phi(\varepsilon)) - \lambda \sum_{l=1}^{L} \log q_\phi(\varphi_l \mid g_\phi(\varepsilon_{1:l-1})) - \log p(\varphi_l \mid g_\phi(\varepsilon_{1:l-1}))] \end{aligned} \tag{20}$$

where $\varepsilon = \{\varepsilon_l\}_{l=1}^{L}$. According to unbiased Monte Carlo estimation, $L(x, y)$ can be approximated by sampling ε from $\mathcal{N}(0,1)$ for T times:

$$L(x, y) \approx \frac{1}{T} \sum_{t=1}^{T} [\log p(x, y \mid g_\phi(\varepsilon^{(t)})) - \lambda \sum_{l=1}^{L} \log q_\phi(\varphi_l \mid g_\phi(\varepsilon_{1:l-1}^{(t)})) - \log p(\varphi_l \mid g_\phi(\varepsilon_{1:l-1}^{(t)}))] \tag{21}$$

where $\varepsilon^{(t)}$ is ε for the tth sampling. In general, $-L(x, y)$ is used as the optimization objective for model training, where $-\log p(x, y \mid g_\phi(\varepsilon^{(t)}))$ can be equivalent to the cross-entropy loss. The detailed training procedure of the proposed ProFormer model is given in **Algorithm 1**. From **Algorithm 1**, it can be seen that in each epoch, the proposed method will construct T networks for sample x by sampling T attention weights from $q_\phi(\varphi \mid x, y)$. At this time, the attention weights of each network are determined and therefore are equivalent to the traditional Transformer. Furthermore, the proposed method will use the average of the losses of these T networks as the training loss of this sample. Notably, only the single-head form of the probabilistic attention is presented above, but it can be easily extended to the multi-head mechanism.

Algorithm 1: Training procedure of ProFormer

Require: The initialization parameters θ, the training dataset $D := \{x_i, y_i\}_{i=1}^{N}$, the batch size B, and the learning rate η, the regularization coefficient λ.

1: **while** θ has not converged **do**

2: $\{x_i, y_i\}_{i=1}^{B} \leftarrow$ Random minibatch of B samples
3: **for** i=1, 2, …, B **do**
4 $h_0'' \leftarrow [\mathrm{Conv}(x, 1, d_m, a, b)]^T$
5: $h_0' \leftarrow \mathrm{Concat}(E^{class}, h_0'')$
6 $h_0 \leftarrow h_0' + E^{pos}$
7: **for** t=1, 2, …, T **do**
8: **for** l=1, 2, … L **do**
9: $h_l' = \mathrm{LayerNorm}(\mathrm{Attention}(h_{l-1})) + h_{l-1}$
10: **compute** $\log q_\phi(\varphi_l \mid g_\phi(\varepsilon_{1:l-1})) - \log p(\varphi_l \mid g_\phi(\varepsilon_{1:l-1}))$
11: $h_l = \mathrm{LayerNorm}(\mathrm{MLP}(h_l')) + h_l'$
12: **end for**
13: $\hat{y} \leftarrow \mathrm{softmax}(\mathrm{FC}(E_L^{class}))$
14: $-\log p(x, y \mid g_\phi(\varepsilon^{(t)})) \leftarrow \mathrm{crossentropy}(\hat{y}, y)$
15: **end for**
16: **compute** $-L_i(x, y) = \frac{1}{T}\sum_{t=1}^{T}[-\log p(x, y \mid g_\phi(\varepsilon^{(t)})) + \lambda \sum_{l=1}^{L} \log q_\phi(\varphi_l \mid g_\phi(\varepsilon_{1:l-1}^{(t)})) - \log p(\varphi_l \mid g_\phi(\varepsilon_{1:l-1}^{(t)}))]$
17: **end for**
18: $\theta \leftarrow \mathrm{Adam}(\nabla_\theta \frac{1}{B}\sum_{i=1}^{B} -L_i(x, y), \theta, \eta)$
19: **end while**

3.3 Uncertainty quantification and decomposition

In the testing phase, given a test sample x, the proposed method also draws T attention weights from $q_\phi(\varphi \mid x, y)$ to construct T networks. Let the output of the tth network be $\hat{y}^{(t)}$. Taking the average of the outputs $\overline{y}$ of all networks as the probability distribution of the predicted label, the index of the largest element in $\overline{y}$ is the predicted label of this test sample, and the entropy of $\overline{y}$ can be used to approximate the total uncertainty in this result:

$$\mathbb{H}[\overline{y} \mid x] \approx -\sum_{c=1}^{C} \overline{y}_c \log \overline{y}_c \tag{22}$$

where $\overline{y}_c$ is the element at position c in $\overline{y}$. It is worth mentioning that a deterministic network can only provide overconfident point estimate prediction and therefore cannot correctly characterize the uncertainty in the diagnostic results. In contrast, the proposed method constructs different networks by sampling multiple weights from the posterior distribution of the attention weights, which can lead to disagreement in the diagnostic decisions of these networks, thus enabling uncertainty quantification [30].

Total uncertainty can be further decomposed into epistemic uncertainty and aleatoric uncertainty:

$$\mathbb{H}[\overline{y} \mid x] = \mathbb{I}[\overline{y}, \varphi \mid x] + E_{q_\phi(\varphi \mid x)}[\mathbb{H}[\overline{y} \mid x, \varphi]] \tag{23}$$

As shown in **Eq. (23)**, the epistemic uncertainty $\mathbb{I}[\overline{y}, \varphi \mid x]$ refers to the discrepancy between the

results obtained by the model when processing the test sample x with different attention weights. The aleatoric uncertainty $E_{q_\varphi(\varphi|x)}[\mathbb{H}[\overline{y} \mid x, \varphi]]$ is defined as the expectation of the entropy of the results obtained by the model with the input x and the determined attention weights. According to Monte Carlo estimation, the aleatoric uncertainty can be approximated as:

$$E_{q_\varphi(\varphi|x)}[\mathbb{H}[\overline{y} \mid x, \varphi]] \approx -\frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C}\hat{y}_c^{(t)} \log \hat{y}_c^{(t)} \tag{24}$$

Combing **Eq. (22)**, **Eq. (23)**, and **Eq. (24)**, the epistemic uncertainty can be approximated as:

$$\begin{aligned}\mathbb{I}[\overline{y}, \varphi \mid x] &= \mathbb{H}[\overline{y} \mid x] - E_{q_\varphi(\varphi|x)}[\mathbb{H}[\overline{y} \mid x, \varphi]] \\ &\approx -\sum_{c=1}^{C}\overline{y}_c \log \overline{y}_c + \frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C}\hat{y}_c^{(t)} \log \hat{y}_c^{(t)}\end{aligned} \tag{25}$$

3.4 Running flow of the proposed method

A trustworthy RMFD framework is established based on the above analysis as shown in **Fig. 5**, and its running flow can be further summarized as follows:

1) Data acquisition: The vibration signals are collected to construct the training dataset (known domain) when the machine is under known failure modes, noise levels, and operating conditions. In addition, the vibration signals are collected to construct the test dataset (unknown domain) when the machine is under unknown failure modes, noise levels, or operating conditions.

2) Model training: The proposed ProFormer model is trained using the known domain.

3) Model testing: The trained ProFormer model is used to give diagnostic results on all test samples in the unknown domain.

4) Uncertainty Analysis: The uncertainty in these diagnostic results is analyzed through the developed uncertainty quantification and decomposition scheme.

5) Researcher intervention: Based on the outcome of the uncertainty analysis, the researchers determine the confidence of the diagnostic results and expose the unknown factors hidden in the test data to find the key to improve the performance of the model.

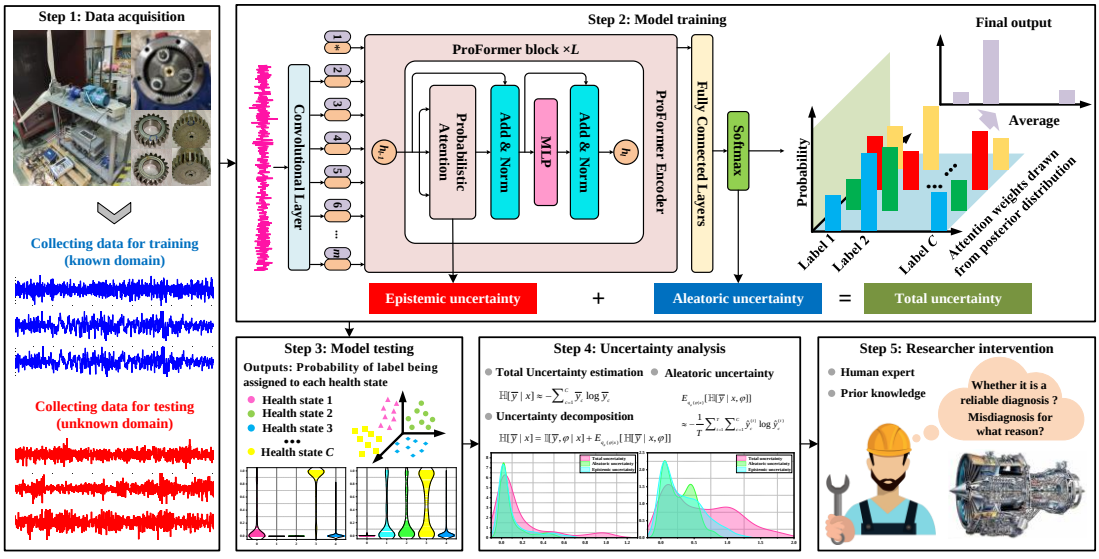

Fig. 5. Running flow of the proposed method.

4. Experimental study

In this section, two experimental cases are constructed to verify the effectiveness of the proposed ProFormer model, using the fault diagnosis of bearing and gearbox as examples respectively. In addition, we consider three experimental scenarios in each experimental case, i.e., there are samples with unknown failure modes, unknown noise levels or unknown operating conditions in the test data.

The main hyperparameters of the proposed ProFormer model is listed in **Table 1**. Moreover, the proposed method is compared with two benchmark methods, namely Resnet18 [32] and Vision Transformer (VIT) [33]. It is worth noting that the parameter settings of VIT and ProFormer are the same. The only difference between them is that VIT uses scaled dot-product attention to calculate the deterministic attention weights, while ProFormer uses probabilistic attention to solve the posterior distribution of attention weights and draw random weights from it.

Table 1

Setting of main hyperparameters of ProFormer

Hyperparameter	Value	Hyperparameter	Value
d_m	128	L	3
a	64	η	5×10^{-4}
b	32	number of epochs	40
number of heads	4	T	128
α	0.5	dropout ratio	0.1
σ^1	0.5	σ^2	20

4.1 Case1: trustworthy fault diagnosis for gearbox

***1) Dataset description*:** The data used for Case 1 is taken from a dataset constructed by Tsinghua University on planetary gearbox faults for wind turbine [31]. The test rig used for the data acquisition experiments is shown in **Fig. 6(a)**, where the input is driven by a motor and the output is connected to a wind wheel. In addition, two accelerometers are deployed on the gearbox casing to collect vibration signals from the X and Y directions with a sampling frequency of 20kHz. A total of 9 gears with different health states are used in the experiments, including normal gear, 4 failure sun gears and 4 failure planetary gears. The details of 8 failure gears are listed in **Table 2**, and **Fig. 6(b)** shows some of the failure gears. In addition, the rotational speed of the input end varies between 15Hz and 40Hz at 1Hz interval, so the signals are collected under 26 different operating conditions.

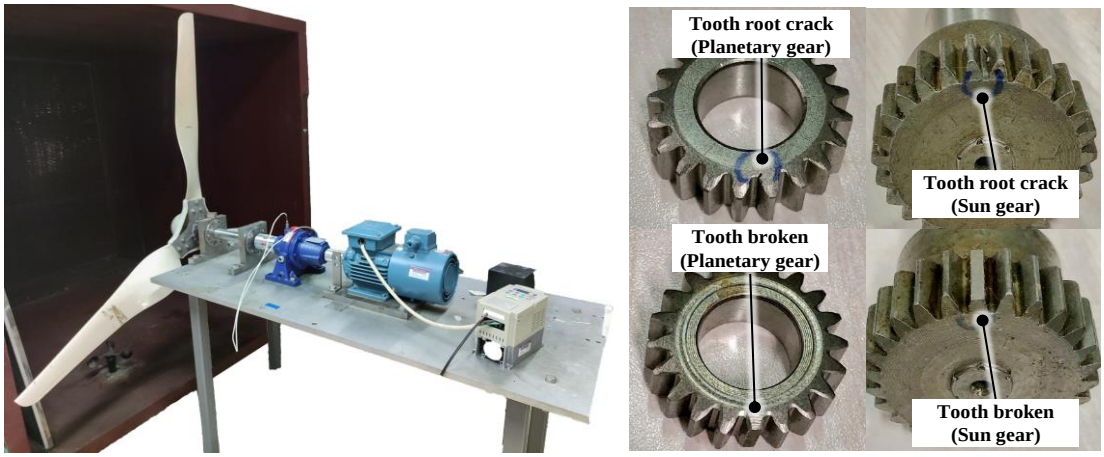

Fig. 6. Tsinghua University wind turbine test rig. (a) Test rig; (b) Failure gears

Table 2

Description of 9 gears with different health states

Gear type	Health states	Description	Label
	Normal		0
Sun gear	Tooth crack	Crack depth is 1/8 dedendum	1
Sun gear	Tooth crack	Crack depth is 1/4 dedendum	2
Sun gear	Tooth crack	Crack depth is 1/2 dedendum	3
Sun gear	Tooth broken	Cut 1/3 of the tooth depth	4
Planetary gear	Tooth crack	Crack depth is 1/8 dedendum	5
Planetary gear	Tooth crack	Crack depth is 1/4 dedendum	6
Planetary gear	Tooth crack	Crack depth is 1/2 dedendum	7
Planetary gear	Tooth broken	Cut 1/3 of the tooth depth	8

***2) Experimental scenario setting*:** Case 1 uses the vibration signals from the Y-direction for the validation experiments. In the experiments, the three diagnostic models are first trained and tested using samples with operating conditions of 32Hz, 36Hz and 40Hz and labels of 0, 1, 2, 3 and 4, which are referred to as the known domain. The purpose of this experiment is to validate the diagnostic performance of the proposed method in the general scenario where the test data does not contain unknown samples. Subsequently, these trained models are applied to three experimental scenarios, where the composition of the unknown domain used as the test dataset differs in each scenario. In Scenario 1, the unknown domain consists of the fault samples with the same operating conditions as the known domain, but with labels 5, 6, 7 and 8. In Scenario 2, the unknown domain consists of the fault samples with the same operating conditions and labels as the known domain, but with a certain signal-to-noise ratios (SNR) of Gaussian white noise imposed, where 4 different SNR of 2dB, 1dB, 0dB, and -1dB are considered. In Scenario 3, the unknown domain consists of fault samples with labels consistent with the known domain, but with operating conditions of 16Hz, 20Hz, 24Hz and 28Hz. The detailed settings for the known domain and the three unknown domains are listed in **Table 3**, where the number of training and test samples are evenly allocated to each operating condition, each sample is 1024 in length, and each sample is processed using the zero-mean normalization method.

Table 3

Detailed settings for known and unknown domains in Case 1

Domain	Label	Operating condition	Number of training samples	Number of test samples
Known domain	0	32Hz, 36Hz, 40Hz	300	75
	1		300	75
	2		300	75
	3		300	75
	4		300	75
Unknown failure mode	5	32Hz, 36Hz, 40Hz		75
	6			75
	7			75
	8			75
Unknown noise level	0	32Hz, 36Hz, 40Hz		75
	1			75
	2			75
	3			75
	4			75
Unknown operating condition	0	16Hz, 20Hz, 24Hz, 28Hz		100
	1			100
	2			100
	3			100
	4			100

***3) Analysis of experimental results for Scenario 1*:** When processing the test data from the known domain, the diagnostic accuracies of ProFormer, VIT, and Resnet18 are 98.93%, 97.07%, and 99.73%, respectively. This suggests that the proposed method has comparable or better diagnostic performance than the comparison methods in the general scenario where the test data does not contain unknown samples. However, when processing samples with unknown fault types, the comparison methods can blindly misdiagnose them as known fault types, which demonstrates the necessity and advantages of the proposed method.

Fig. 7 shows the probability distributions of the predicted labels given by Resnet18, VIT and ProFormer when processing 1 known sample and 4 samples with unknown fault types, where the abscissa represents the fault label and the ordinate denotes the probability value. As shown in **Fig. 7**, Resnet18 and VIT can only give overconfident diagnostic results, while ProFormer can provide uncertainty in these results. This is because the proposed method can construct T different networks by Monte Carlo sampling, and these T networks give different prediction probability distributions. All three models give correct diagnostic result when processing the known sample with a true label of 1. Specifically, in ProFormer, T networks give similar prediction probability distributions, reflecting the low uncertainty and high confidence of this result. However, when processing an unknown sample with a true label of 5, VIT and Resnet18 confidently predict its label as 4 and 0 without warning to the researchers. In contrast, in ProFormer, the large differences between the diagnostic decisions given by T networks make the probability of the predicted label being assigned to each health state exhibit high uncertainty, which indicates that the result is less reliable and requires further investigation by the researchers. In processing the unknown samples with true labels of 6, 7 and 8, all three models show experimental phenomena

similar to those described above, indicating that ProFormer not only identifies known fault samples with sufficient confidence, but also improves diagnostic reliability by conveying to the researchers the high uncertainty in the results to indicate that the equipment may be in an unknown failure mode.

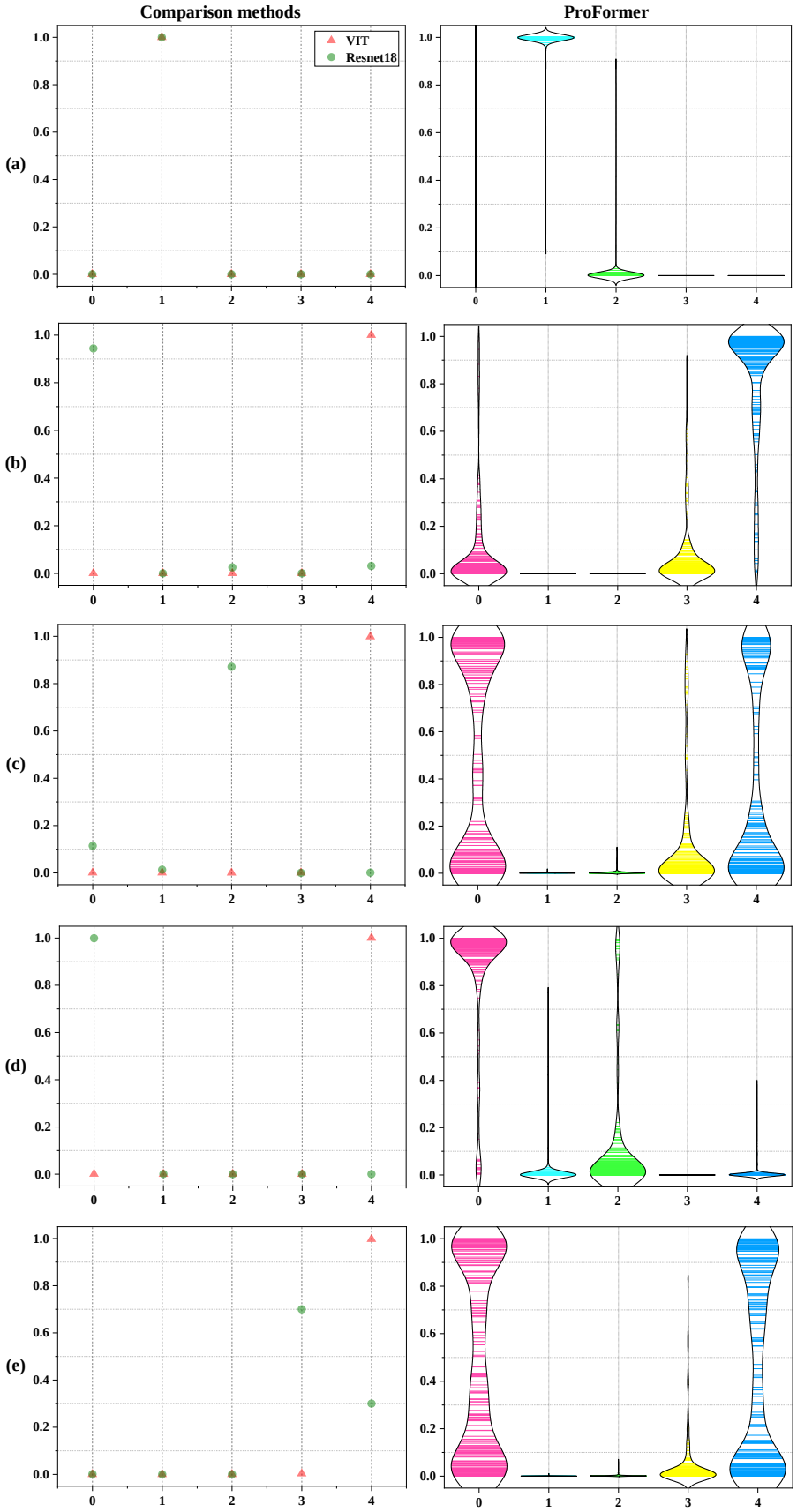

Fig. 7 Diagnostic results (Scenario 1, Case 1) of each method for samples with different fault labels: (a) label 1; (b) label 5; (c) label 6; (d) label 7; (e) label 8.

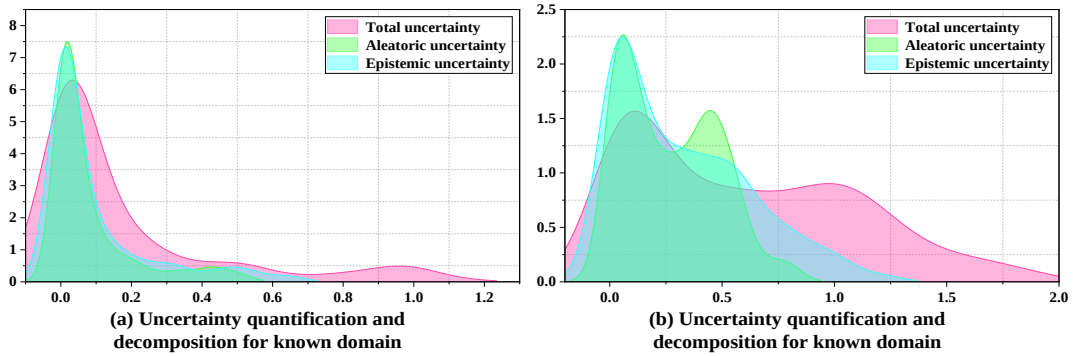

Fig. 8. Uncertainty estimations (Scenario 1, Case 1) of diagnostic results for test samples in known and unknown domains by the proposed method.

To validate the uncertainty quantification and decomposition capability of the proposed method and its interpretability, we explain the sources and composition of the uncertainty obtained by ProFormer for all test samples in known and unknown domains. **Fig. 8** shows the distributions of the three types of uncertainty, where the abscissa represents the probability value and the ordinate represents the density. As shown in **Fig. 8**, in the known domain, the three distributions show peaked patterns with values concentrated around 0, indicating low uncertainty in the results, while in the unknown domain, the shapes of the three distributions are flatter, indicating high uncertainty in the results. This experimental phenomenon is consistent with the expectations. Furthermore, the contribution of epistemic uncertainty to the total uncertainty is significant, which can be interpreted as a lack of diagnostic knowledge used by the model to identify unknown samples. Considering that the test samples in Scenario 1 is precisely the samples with unknown fault types, the interpretability of the proposed method is then demonstrated to some extent.

***4) Analysis of experimental results for Scenario 2*:** In practical engineering, the acquired vibration signals are usually accompanied by severe noise, which may be caused by the noisy equipment working environment, or by the sensor measurement errors. Therefore, the unknown domain is constructed by adding a certain SNR of Gaussian white noise to the test samples in the known domain. As shown in **Fig. 9**, the diagnostic accuracy of each method shows a significant decline as the SNR decreases, with Resnet18 showing the most significant decline, while ProFormer demonstrates a strong stability. There are two main reasons for this experimental phenomenon: (1) The vibration signals with low SNR lose their periodicity, making CNN no longer able to identify signals by capturing valid local information such as the pulse band of signals. Under the strong noise interference, it is more important to use the self-attention mechanism of Transformer to mine the global information within signals. (2) The proposed method does not train a single network, but an ensemble of networks, and this training method is similar to ensemble learning, naturally enhances the generalization of the model.

In addition, despite the large number of misdiagnoses at low SNR, the comparison methods do not alert the researchers, while the proposed method conveys to the researchers the confidence of these diagnosis results. As shown in **Fig. 10**, when processing a fault sample with a true label of 2 and a SNR of 0 dB, both Resnet18 and VIT identify it with high confidence as 4. For the proposed method, although

it incorrectly identifies this sample as 1, the probability of the predicted label being assigned to each health state show high uncertainty, which suggests that the model does not have sufficient confidence to give diagnostic result, requiring the intervention of the researchers.

Fig. 11 shows the uncertainty estimates of the proposed method for the diagnostic results of all test samples at different noise levels. As the SNR continues to decrease, all three types of uncertainty gradually increase, and the contribution of aleatoric uncertainty to total uncertainty tends to be significant. Since aleatoric uncertainty characterizes the inherent randomness hidden in the data, this experimental phenomenon can be interpreted by that the valid information in the data is drowned in noise, making the model unable to make diagnostic decisions with certainty. Considering that the test samples in Scenario 2 is precisely the samples accompanied with noise, the interpretability of the proposed method is further validated. As the inherent randomness of the data is not eliminated, using more data for training does not improve the performance of the model, but requires the researcher to use more efficient instruments or techniques to resist the interference of noise and collect cleaner signals. In this way, the dependency between the researcher and the diagnostic model is established.

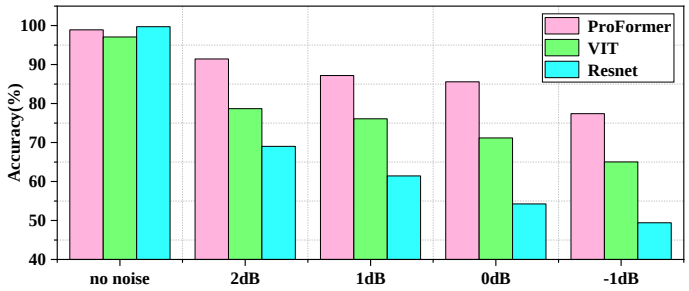

Fig. 9. Diagnostic accuracy of each method in Scenario 2 of Case 1.

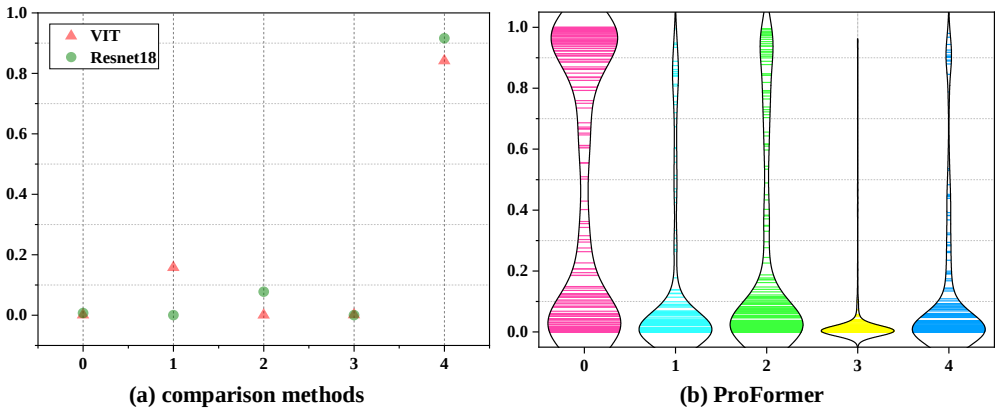

Fig. 10. Diagnostic result (Scenario 2, Case 1) of each method for a fault sample with a true label of 2 and a SNR of 0dB.

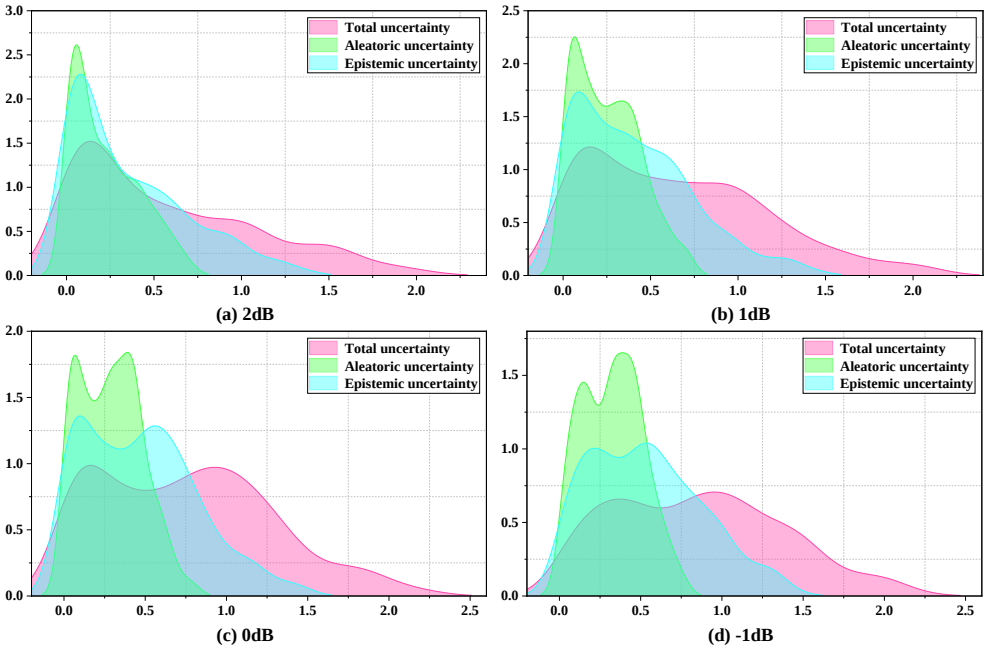

Fig. 11. Uncertainty estimations (Scenario 2, Case 1) of diagnostic results for test samples in unknown domain by the proposed method.

***5) Analysis of experimental results for Scenario 3*:** Mechanical equipment often operates under unpredictable operating conditions, making the distribution of the collected test data unknown and causing degradation in model performance. Therefore, the unknown domain is constructed by varying the rotational speed of the equipment. As shown in **Fig. 12**, there is an overall decreasing trend in the diagnostic accuracy of each method as the speed difference between the unknown and known domains gradually increases, with Resnet18 showing the most stable performance, probably because the local receptive field of CNN is better at capturing domain invariant features. Although the accuracy of ProFormer is low under the unknown operating conditions, it can still ensure the confidence of the results by conveying the uncertainty of these results. As shown in **Fig. 13**, when processing a fault sample with an operating condition of 16Hz and a true label of 0, even though ProFormer incorrectly identifies it as 3, it is evident from the diagnostic result that the probability of the predicted label being assigned to both 0 and 3 exhibits high uncertainty, which indicates a low confidence of this result. Similarly, experimental phenomena similar to those described above emerge when processing a fault sample with a true label of 1, fully validating the ability of the proposed method to evaluate the confidence of the results.

Fig. 14 shows the uncertainty estimations of the diagnostic results for all test samples by the proposed method under different unknown operating conditions. As the speed difference between the unknown and known domains gradually increases, all three types of uncertainty show an overall upward trend. Furthermore, as can be seen in **Fig. 15**, the proportion of epistemic uncertainty in the total uncertainty is also gradually increasing as the speed difference increases. This experimental phenomenon can be interpreted as the exacerbation of the model's lack of diagnostic knowledge. As the test samples in Scenario 3 are precisely the samples with unknown operating conditions, the interpretability of the proposed method is again demonstrated. To reduce the epistemic uncertainty, richer and more diverse data need to be collected for model training.

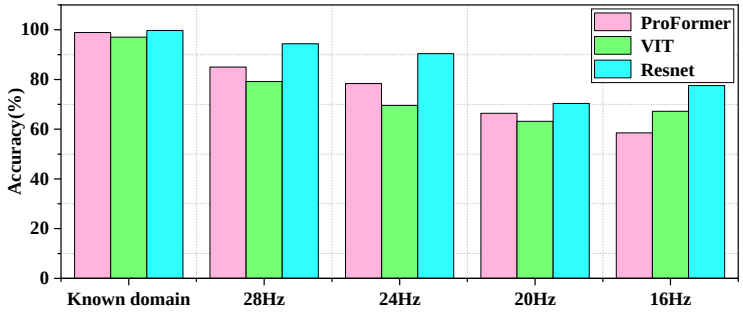

Fig. 12. Diagnostic accuracy of each method in Scenario 3 of Case 1.

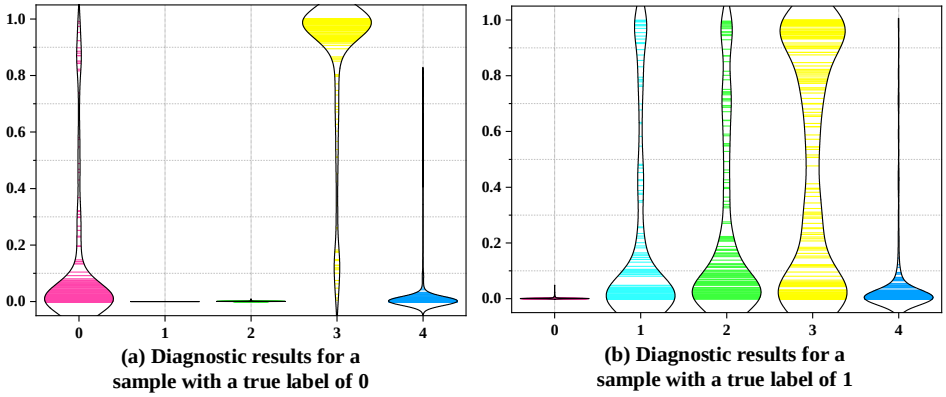

Fig. 13. Diagnostic results (Scenario 3, Case 1) of the proposed method under operating conditions of 16Hz.

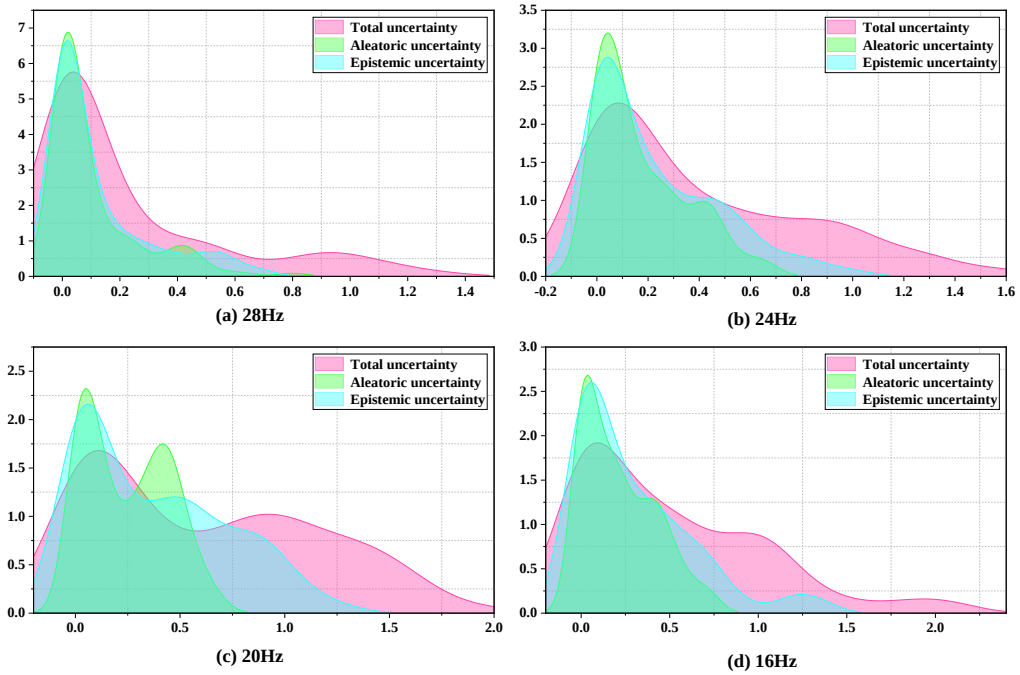

Fig. 14. Uncertainty estimations (Scenario 3, Case 1) of diagnostic results for test samples in unknown domain by the proposed method.

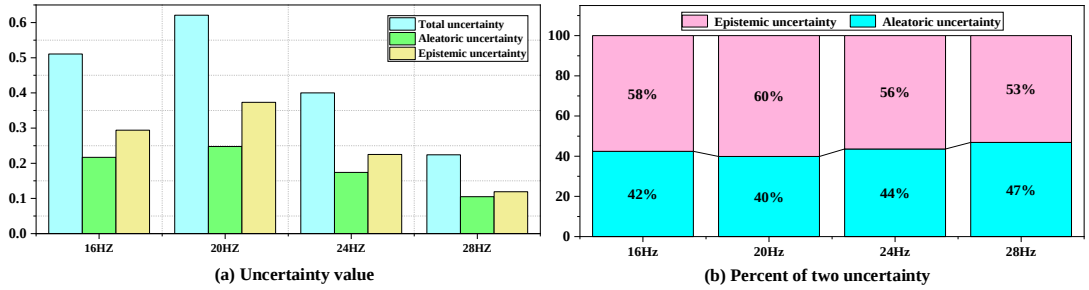

Fig. 15. Uncertainty composition (Scenario 3, Case 1) under different operating conditions.

4.2 Case2: trustworthy fault diagnosis for bearing

***1) Data preparation*:** The data used for Case 2 is collected using QPZZ-II bearing fault test rig [8] shown in **Fig. 16**. The data acquisition experiments consider 5 different health states of bearing, namely normal, inner ring failure, outer ring failure, ball failure and compound failure (ball failure + outer ring failure) with labels 0, 1, 2, 3, 4 in that order. **Fig. 17** shows the 3 failure bearings, where the failures are simulated by cutting square grooves of 0.5mm width and 0.5mm depth in the bearing surface. In addition, the motor speed of this test rig varies between 900rpm and 1500rpm at 100rpm intervals, so the signals are collected under 7 different operating conditions. The sampling frequency of the installed Piezo-electric accelerometer is 25.6kHz and the collected data has 6 different channels.

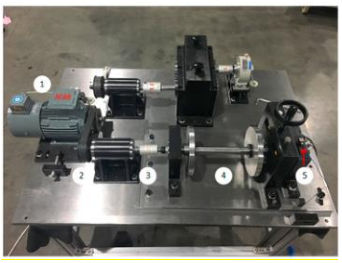

Fig.16. QPZZ-II bearing fault test rig.

Fig.17. Three failure bearings used in QPZZ-II bearing fault test rig.

***2) Experimental scenario setting*:** Case 2 uses the data from the first channel for the validation experiments. As in Case 1, the three diagnostic models are first trained and tested using samples in the known domain, and these trained models are then applied to the three experimental scenarios described above. The detailed settings for the known domain and the three unknown domains in Case 2 are listed in **Table 4**. It should be added that in Scenario 2, the SNR of the fault sample with Gaussian noise in the unknown domain is also 5dB, 4dB, 3dB or 2dB. In addition, each sample has a length of 1024 and is pre-processed by the zero-mean normalization method.

Table 4

Detailed settings for known and unknown domains in Case 2

Domain	Label	Operating condition	Number of training samples	Number of test samples
Known domain	0	1500rpm	350	50
	1		350	50
	2		350	50
	3		350	50
Unknown failure mode	4	1500rpm		50
Unknown noise level	0	1500rpm		50
	1			50
	2			50
	3			50
Unknown operating condition	0	1100rpm, 1300rpm		100
	1			100
	2			100
	3			100

***3) Analysis of experimental results for Scenario 1*:** When processing the test data from the known domain, the diagnostic accuracies of ProFormer, VIT, and Resnet18 are 100.0%, 98.5%, and 100.0%, respectively. This again validates that the proposed method has comparable or higher diagnostic accuracy than the comparison methods even in the general scenario. However, when processing samples with unknown fault types, the necessity and advantages of the proposed method is demonstrated.

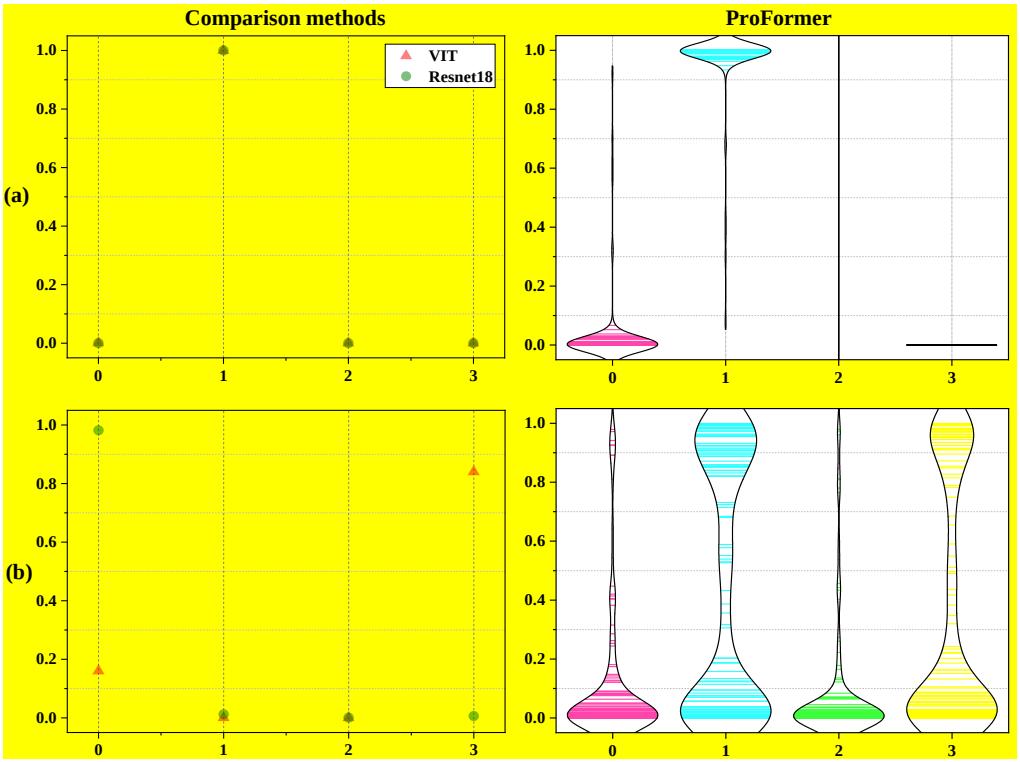

Fig. 18 Diagnostic results (Scenario 1, Case 2) of each method for samples with different fault labels: (a) label 1; (b) label 4.

Fig. 18 shows the probability distributions of the predicted labels given by Resnet18, VIT and ProFormer when processing a known sample with a true label of 1 and an unknown sample with a true label of 4. As shown in **Fig. 18**, all three methods give correct diagnostic result for the known sample. However, for the unknown sample, VIT and Resnet18 incorrectly predict its label as 3 and 0, respectively, without warning the researcher. In contrast, ProFormer considers that this sample has a high uncertainty in the probability of being assigned to each known health state, suggesting that this sample may be a sample from an unknown fault class that requires careful handling by the researcher intervention. This demonstrates again that ProFormer can improve diagnostic reliability by conveying to researcher the uncertainty of the results.

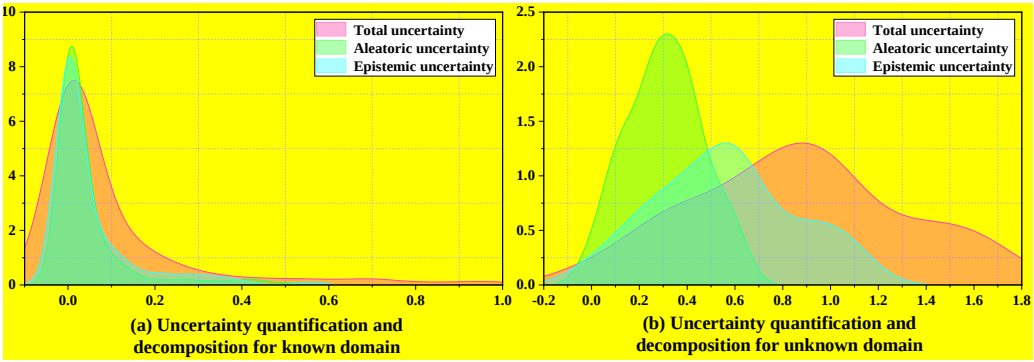

Fig. 19. Uncertainty estimations (Scenario 1, Case 2) of diagnostic results for test samples in known and unknown domains by the proposed method.

Fig. 19 shows the sources and composition of the uncertainty obtained by ProFormer for all test samples in the known and unknown domains. Consistent with Case 1, the distributions of the three types of uncertainty show a peak pattern in the known domain and a flat pattern in the unknown domain. In addition, the proportion of epistemic uncertainty in the total uncertainty is also significant, which again demonstrates the interpretability of the proposed method.

***4) Analysis of experimental results for Scenario 2*: Fig. 20** shows the diagnostic results provided by VIT, Resnet18 and ProFormer when processing a fault sample with a true label of 3 and a SNR of 2dB. For all three methods, they both incorrectly assume that the label of this sample is 0 due to the interference of noise. Meanwhile, both VIT and Resnet18 have a high confidence for their diagnoses. ProFormer, however, considers that the probability of the label of this sample being assigned to 3 also has high uncertainty, which implies that it is not confident enough to give the predicted label of this sample and requires the intervention of the researcher.

Fig. 21 shows the uncertainty estimates of the proposed method for the diagnostic results of all test samples at different noise levels. Consistent with Case 1, all three types of uncertainty gradually increase as the SNR continues to decrease, and the contribution of aleatoric uncertainty to the total uncertainty tends to be significant. In particular, the proportion of aleatoric uncertainty in total uncertainty in Scenario 2 is significantly higher that in Scenario 1, precisely because of the increased randomness hidden in the samples.

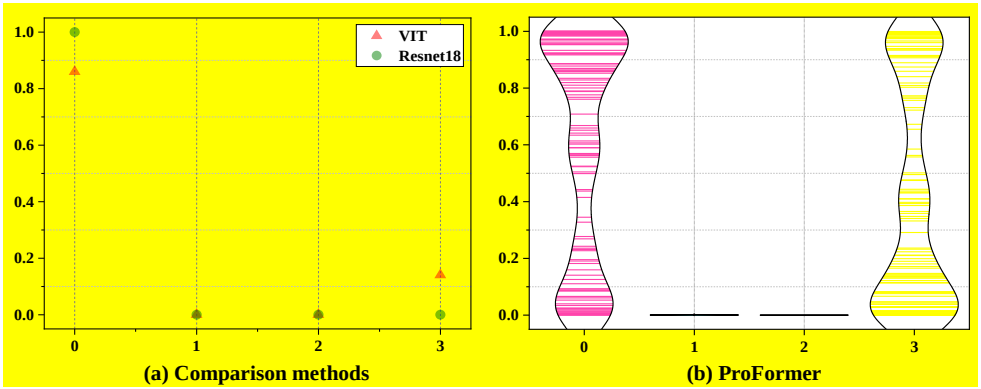

Fig. 20. Diagnostic result (Scenario 2, Case 2) of each method for a fault sample with a true label of 3 and a SNR of 2dB.

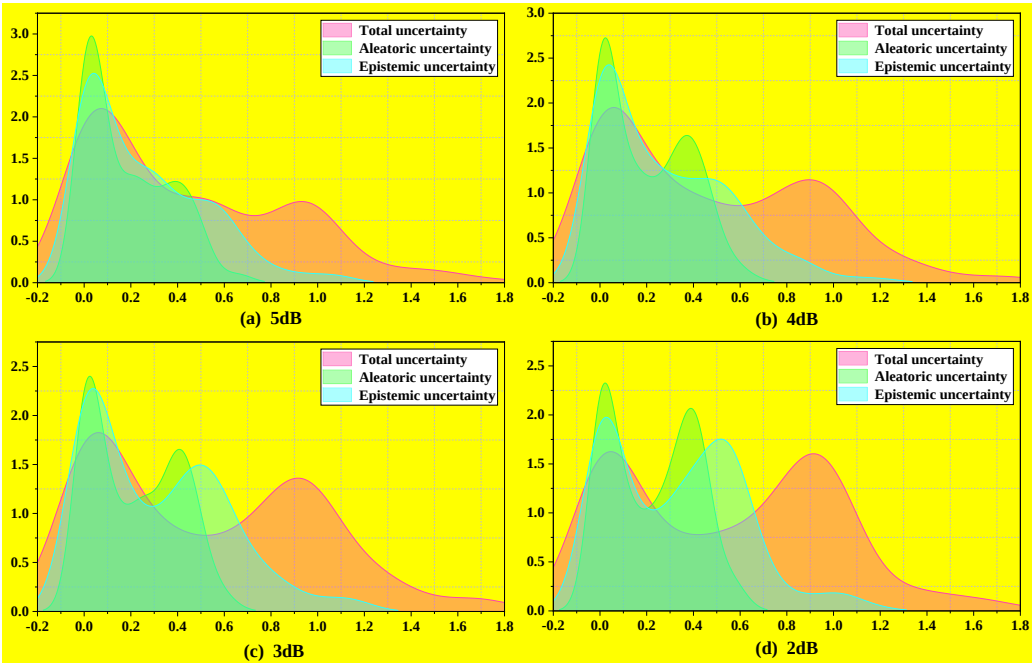

Fig. 21. Uncertainty estimations (Scenario 2, Case 2) of diagnostic results for test samples in unknown domain by the proposed method.

***5) Analysis of experimental results for Scenario 3*: Fig 22** shows the diagnostic results provided by ProFormer when processing two samples with unknown operating conditions. For a sample with a true label of 1 and an operating condition of 1300rpm, although ProFormer incorrectly predicts its label as 0 due to the domain shift, it can still warn the researcher by showing uncertainty in the results, thus avoiding misdiagnosis. A similar experimental phenomenon can also be observed when ProFormer processes a sample with a true label of 3 and an operating condition of 1300rpm.

Fig. 23 shows the uncertainty estimations of the diagnostic results for all test samples by the proposed method under different unknown operating conditions. As expected, the uncertainty obtained by ProFormer for samples with operating conditions of 1100 rpm is significantly higher than that obtained for samples with operating conditions of 1300 rpm due to the increased speed difference.

However, contrary to expectations, the proportion of aleatoric uncertainty in total uncertainty is equal to or even higher than that of epistemic uncertainty. The reason for this experimental phenomenon could be that the prior distribution used in the proposed method is not comprehensive enough or that the data used in Case 2 itself contains a high level of randomness.

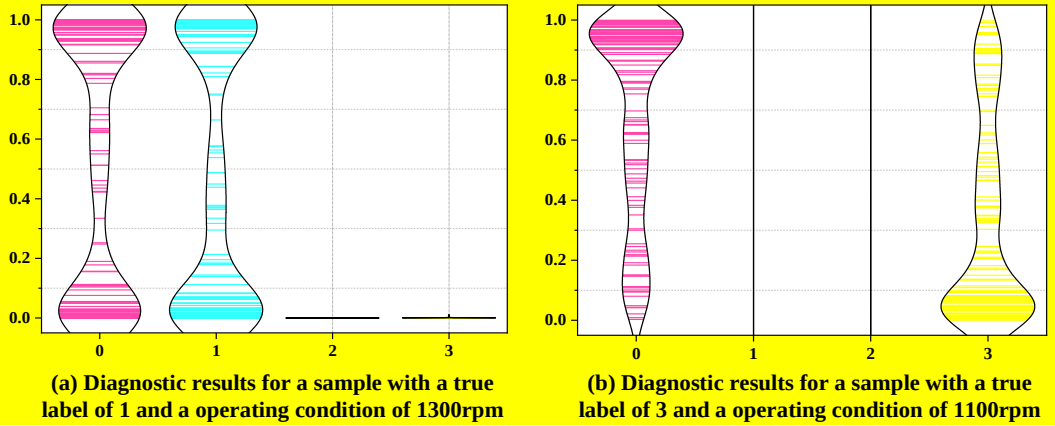

Fig. 22. Diagnostic results (Scenario 3, Case 2) of the proposed method for samples with different labels and operating conditions.

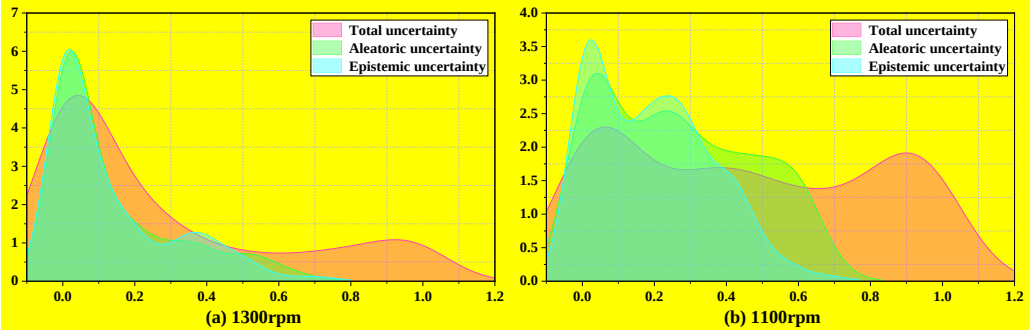

Fig. 23. Uncertainty estimations (Scenario 3, Case 2) of diagnostic results for test samples in unknown domain by the proposed method.

5. Conclusion

In this paper, a ProFormer model is proposed for trustworthy RMFD, with the following key conclusions: (1) Analyzing and explaining the sources and composition of uncertainty in the diagnostic results provided by DL models is beneficial for improving their interpretability and trustworthiness. (2) The designed probabilistic attention and the defined optimization objective function can be used to model the prior and variational posterior distributions of attention weights, thus empowering the model to perceive uncertainty. (3) The developed uncertainty quantification and decomposition scheme can be used to characterize the confidence in the diagnostic results and to decompose the total uncertainty in the results into epistemic uncertainty and aleatoric uncertainty.

Considering that the performance of models that capture uncertainty using Bayesian variational learning is largely dependent on the constructed prior distribution, in the future, we plan to use more comprehensive prior distributions to enhance the ability of the models to perceive uncertainty. In addition,

the mean-field theory used in this paper assumes that the attention weights from different blocks are independent of each other. We plan to investigate how this assumption can be relaxed to capture the dependence between the attention weights of different blocks.

Furthermore, to construct a more reliable mechanism for human-model interaction, we should not only consider explicit knowledge such as the uncertainty of diagnostic results, which has been emphasized in this paper, but also some implicit knowledge. The implicit knowledge can be obtained in the actual production process by operators for their machines, supervisors for their cells and mangers for their factories. A comprehensive self-improvement mechanism for deep diagnostic models should be based on both explicit and implicit knowledge, rather than only considering explicit knowledge as in this paper. This is the shortcoming of this paper, and we will explore the implicit knowledge deeply in future work to promote the achievement of the trustworthy fault diagnosis.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 52275104, No. 51905160) and the Natural Science Fund for Excellent Young Scholars of Hunan Province (No. 2021JJ20017).

Reference

[1] S. Yan, H. Shao, Z. Min, *et al*., "FGDAE: A new machinery anomaly detection method towards complex operating conditions," *Reliab. Eng. Syst. Saf.*, vol. 236, Art. no. 109319, Aug. 2023.

[2] X. Chen, H. Shao, Y. Xiao, *et al*., "Collaborative fault diagnosis of rotating machinery via dual adversarial guided unsupervised multi-domain adaptation network," *Mech. Syst. Signal Process*., vol. 198, Art. no. 110427, Sep. 2023.

[3] J. Lin, H. Shao, X. Zhou, *et al*., "Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals," *Expert Systems with Applications*, vol. 230, Art. no. 120696, Nov. 2023.

[4] Y. Li, Z. Zhou, C. Sun, *et al*., "Variational Attention-Based Interpretable Transformer Network for Rotary Machine Fault Diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, doi: 10.1109/TNNLS.2022.3202234.

[5] D. Wang, Y. Chen, C. Shen, *et al*., "Fully Interpretable Neural Networks for Machine Health Monitoring," *Mech. Syst. Signal Process*., vol. 168, Art. no. 108673, Apr. 2022.

[6] H. Wang, Z. Liu, D. Peng, *et al*., "Understanding and learning discriminant features based on multiattention 1DCNN for wheelset bearing fault diagnosis," *IEEE Trans. Ind. Informat*., vol. 16, no. 9, pp. 5735-5745, Sep. 2020.

[7] B. An, S. Wang, Z. Zhao, *et al*., "Interpretable Neural Network via Algorithm Unrolling for Mechanical Fault Diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, Art no. 3517011, Jul. 2022.

[8] Y. Xiao, H. Shao, S. Han, *et al*., "Novel Joint Transfer Network for Unsupervised Bearing Fault Diagnosis From Simulation Domain to Experimental Domain," *IEEE-ASME Trans. Mech*., vol. 27, no. 6, pp. 5254-5263, Jun. 2022.

[9] T. Li, Z. Zhao, C. Sun, *et al*., "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE Trans. Syst. Man Cybern*., vol. 52, no. 4, pp. 2302-2312, Apr. 2022.

[10] Z. Shang, Z. Zhao, R. Yan, *et al*. "Denoising Fault-Aware Wavelet Network: A Signal Processing Informed Neural Network for Fault Diagnosis," *Chin. J. Mech. Eng.*, vol. 36, Art. no. 9, Jan. 2023.

[11] Y. Xiao, H. Shao, Z. Min, *et al*., "Multiscale dilated convolutional subdomain adaptation network with attention for unsupervised fault diagnosis of rotating machinery cross operating conditions," *Measurement*, vol. 204, Art. no. 112146, Nov. 2022.

[12] J. Luo, H. Shao, H. Cao, *et al*., "Modified DSAN for unsupervised cross-domain fault diagnosis of bearing under speed fluctuation," *J. Manuf. Syst*., vol. 65, pp. 180-191, Oct. 2022.

[13] T. Han, Y. Li, "Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles," *Reliab. Eng. Syst. Saf.*, vol. 226, Art. no. 108648, Oct. 2022.

[14] T. Zhou, L. Zhang, T. Han, *et al*. "An uncertainty-informed framework for trustworthy fault diagnosis in safety-critical applications," *Reliab. Eng. Syst. Saf.*, vol. 229, Art. no. 108865, Jan. 2023.

[15] J. Gawlikowski, C. Tassi, M. Ali, *et al*., "A Survey of Uncertainty in Deep Neural Networks," arXiv:2107.03342, 2021.

[16] A. Malinin, M. Gales, "Predictive Uncertainty Estimation via Prior Networks," *Adv. Neural Inf. Process. Syst.* (*NIPS*), vol. 31, 2018.

[17] M. Abdar, F. Pourpanah, S. Hussain, *et al*., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243-297, Jan. 2021.

[18] J. Arco, A. Ortiz, J. Ramírez, *et al*., "Uncertainty-driven ensembles of multi-scale deep architectures for image classification," *Inf. Fusion*, vol. 89, pp. 53-65, Jan. 2023.

[19] C. Blundell, J. Cornebise, K. Kavukcuoglu, *et al*., "Weight uncertainty in neural network," arXiv :1505.05424, 2015.

[20] B. Xue, S. Hu, J. Xu, *et al*., "Bayesian Neural Network Language Modeling for Speech Recognition," *IEEE/ACM Trans. Audio, Speech, Language Process*., vol. 30, pp. 2900-2917, Sep. 2022.

[21] T. Zhou, T. Han, and E. Droguettcd, "Towards trustworthy machine fault diagnosis: A probabilistic Bayesian deep learning framework," *Reliab. Eng. Syst. Saf.*, vol. 224, Art. no. 108525, Aug. 2022.

[22] A. Maged, M. Xie, "Uncertainty utilization in fault detection using Bayesian deep learning," *J. Manuf. Syst*., vol. 64, pp. 316-329, Jul. 2022.

[23] M. Liang, K. Zhou, "Probabilistic bearing fault diagnosis using Gaussian process with tailored feature extraction," *Int. J. Adv. Manuf. Technol*., vol. 119, pp. 2059-2076, Mar. 2022.

[24] Y. Ding, M. Jia, Q. Miao, *et al*., "A novel time-frequency transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings," *Mech. Syst. Signal Process.*, vol. 168, Art. no. 108616, Apr. 2022.

[25] A. Vaswani, S. Noam, P. Niki, e*t al*., "Attention is all you need," *Proc. Adv. Neural Inf. Process. Syst*., pp. 5998-6008, 2017.

[26] K. Shridhar, F. Laumann, M. Liwicki, "A comprehensive guide to Bayesian convolutional neural network with variational inference," arXiv: 1901.02731, 2019.

[27] J. Devlin, M. Chang, K. Lee, *et al*., "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, 2018.

[28] S. Zhang, X. Fan, B. Chen, *et al*., "Bayesian attention belief networks," arXiv:2106.05251, 2021.

[29] X. Fan, S. Zhang, B. Chen, *et al*., "Bayesian Attention Modules," arXiv: 2010.10604, 2020.

[30] L. Chai, "Uncertainty estimation in Bayesian neural networks and links to interpretability," University of Cambridge, 2018.

[31] T. Han, C. Liu, L. Wu, *et al*., "An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems," *Mech. Syst. Signal Process*., vol. 117, pp. 170-187, Feb. 2019.

[32] Z. Zhao, T. Li, J. Wu, *et al*., "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," *ISA Trans*., vol. 107, pp. 224-255, Dec. 2020.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al*., "An image is worth 16×16 words: Transformers for image recognition at scale," *Proc. Int. Conf. Learn. Represent*. (*ICLR*), pp. 1-22, 2020.