Trustworthy Fault Diagnosis with Uncertainty Estimation through Evidential Convolutional Neural Networks

***Abstract*—Deep neural networks (DNNs) have been widely used for intelligent fault diagnosis under the closed world assumption that any testing data is within classes of the training data. However, in reality, out-of-distribution (OOD) cases such as new fault conditions can happen after the original trained model is deployed. Most of the current DNNs are deterministic which can misclassify with high confidence in the open-world scenario. This overconfident behavior would not guarantee the reliability and robustness of fault diagnosis results in practice. Therefore, trustworthy intelligent fault diagnosis with uncertainty estimation is crucial for real applications. In this paper, we develop a novel convolutional neural network integrating evidence theory to achieve fault classifications with prediction uncertainty estimation. The estimated prediction uncertainty can identify potential OOD samples. This approach allows a minimal modification of the state-of-the-art DNN model by using a risk-calibrated evidential loss function and Dirichlet distribution that replaces the classification probabilities. The experimental results show that the proposed approach can not only achieve accurate classification of known classes but also detect unknown classes effectively. The proposed method shows significant potential in detecting OOD patterns and provides trustworthy fault diagnosis in open and non-stationary environments.**

Index Terms*— Trustworthy AI, Fault diagnosis, Open set recognition (OSR), Evidential convolutional neural networks, Uncertainty estimation.

I. Introduction

With the development of smart and digital manufacturing, condition monitoring of industrial machines has become increasingly important in guaranteeing production efficiency and safety. With more access to sensory data, data-driven approaches for fault diagnosis have gained extensive interest due to the outstanding performance of machine learning (ML), especially deep learning (DL). Most of the existing development is under the closed-world assumption that the testing data are drawn from the same distribution as the training data, known as the in-distribution (ID). However, the practical scenario can be open-world where the testing samples may be out-of-distribution (OOD) compared with training samples [1]. The existing fault diagnosis methods can misclassify the OOD samples with high confidence which means that they are incapable of detecting the unseen fault classes. This limitation prevents its application in real-time monitoring and control of safety-critical industrial systems, including nuclear plants, chemical processes, transportation, etc [2]. Dealing with OOD inputs is essential for practical applications of ML [3].

An important subtopic of OOD is open-set recognition (OSR) which aims to detect multiclass unknown samples and avoid overconfident behavior [4]. OSR requires that the multiclass classifier not only classify known classes accurately but also detect unknown classes in the test samples. **Fig. 1** provides a visual comparison of traditional classification and OSR. The methods used in OSR scenarios are mainly classification-based methods since ID data comprises multiple classes during the training processing. In addition, distance-based methods, reconstruction-error-based methods, and density-based methods are developed [5]. In classification-based methods, redistributing the logits that are the unscaled output of the penultimate layer in a neural network can reduce the probability of overconfidence. Extreme value theory (EVT) has been widely applied in OSR tasks [6]. It can analyze the data distribution of abnormally high or low values and rely on the probabilistic model such as the Weibull distribution. However, selecting a proper probabilistic model is challenging. Accurate uncertainty predictions can help interpret the confidence levels and capture semantic shifts in OSR samples that are drawn from multiple classes [7].

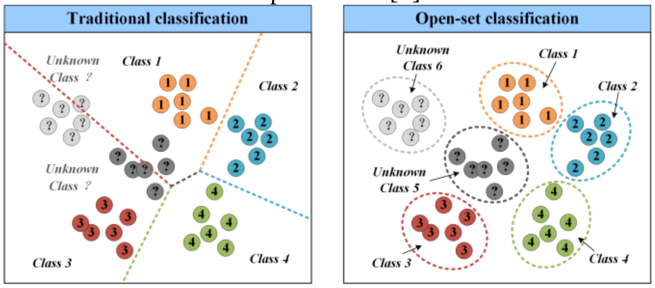


Fig 1. Difference between traditional classification and open-set recognition.

Traditional uncertainty estimation approaches mainly rely on sampling methods, such as Monte Carlo (MC) sampling [8]. MC is an effective method to approximate an exact posterior inference, which has been a popular method for uncertainty estimation. However, it is a slow and computationally expensive method when integrated into a deep architecture [9]. To combat this problem, the MC dropout method is introduced by using the dropouts as a regularization term to compute the uncertainty. Dropouts is an effective technique that has been widely used to solve the overfitting problem in deep neural networks (DNNs) [10]. The dropout methods, including MC dropout, Bernoulli dropout, and Gaussian dropout, can be accurate in quantifying the prediction uncertainty by sampling weights. However, dropout-based methods require many repeated feed-forward calculations with high computational costs [11]. Bayesian deep learning can provide a theoretical framework for uncertainty estimation by modeling the distributions for the parameters. It is robust to overfitting problems. However, it requires expensive MC sampling [9]. Other uncertainty estimation methods using deep ensembles have been used to detect misclassification and out-of-distribution inputs, warning of the potential untrustworthy diagnosis [12]. The probabilistic ensemble approach and Bayesian nonparametric ensemble approach can robustly estimate the uncertainty from different sources [13]. However, dropouts and deep ensemble approaches can be computationally expensive due to the high demand for memory, and hyperparameters are hard to determine. Also, it is difficult to directly infer the posterior distribution of the weights and to choose a weight prior [14].

Recently, prior networks that parameterize a Dirichlet prior over output distributions have been proven to have better uncertainty estimation performance than MC dropout [14]. Furthermore, the Dirichlet prior distribution is not overly dependent on the training samples and can adapt to the data changes in the multiclass OSR scenarios [15]. It has been used to estimate data uncertainties in DNNs by directly estimating parameters of the predictive posterior as their output [16]. However, the available Dirichlet prior distribution relies on auxiliary losses in DNNs to achieve good classification performance and uncertainty estimation [17]. Enhanced uncertainty estimation can be achieved by deriving the properties of the new loss function. The most important loss functions, such as the maximum likelihood loss, the cross-entropy loss, and the sum of squares loss, can be applied for uncertainty estimation in DNNs [16]. Dirichlet networks with appropriate loss functions require no sampling and minimal changes to the standard neural network structures, which can effectively reduce the risk of underestimation of uncertainty [8]. This approach is promising for predictive uncertainty estimation in practical settings, which can provide robust and trustworthy prediction results. It has been investigated in image recognition tasks and achieved good performance.

This paper develops a new framework for trustworthy fault diagnosis in open set fault diagnosis (OSFD) tasks. Inspired by the implicit density models, the proposed method links the Dirichlet distribution parameter to the evidence assigned over classes, obtaining the Dirichlet distribution instead of the point estimation of the probabilities. To build the classifier, the VGG network (developed by the Visual Geometry Group [18]), has been chosen as the basic structure. Then, an improved evidential VGG-architecture network (EVGG) has been developed which can provide categorical probabilities and pignistic probabilities, respectively. The proposed approach assigns more evidence to the correct labels and decreases the misleading evidence from misclassified samples by applying the risk-calibrated evidential loss function. Two benchmark datasets are used to test the performance. Experimental results show that the proposed method can accurately diagnose known classes and detect unknown classes. The main contributions of the paper are:

1) A new trustworthy fault diagnosis framework in OSFD tasks is proposed. It only requires minimal changes of DL approaches for general neural network structures and can capture uncertainty in prediction. By integrating evidence theory, the proposed method can achieve good diagnosis performance for known classes and detect unknown classes through effective estimation of prediction uncertainty.

2) The EVGG model treats the predictions of the classifier as evidence and replaces the point estimation of the probabilities over classes from *softmax* with the Dirichlet distribution. In addition, the weights of the standard backpropagation neural network are optimized through the risk-calibrated evidential loss function by assigning more evidence to correct classification.

3) By developing an improved EVGG model with evidence theory, the present work achieves end-to-end trustworthy intelligent fault diagnosis. The proposed method achieves effective uncertainty estimation and high diagnostic performance with known and unknown classes. The uncertainty estimation provides support to trustworthy predictions with the capability of detecting OOD samples.

The remainder of the present paper is organized as follows. Section II introduces the theoretical background. Section III presents the details of the proposed method. The experimental study is presented in Section IV. Section V gives the conclusion and possible future work.

II. Background

A. Problem description

In the practical task of fault diagnosis with possible new faults, it is critical to not only classify the known faults contained in the training process but also detect the new faults in testing. The goal of the proposed method is to classify known classes accurately and recognize unknown classes that could happen after the deployment of the trained models. Suppose $I=\{(x_1,y_1),...,(x_n,y_n)\}$ is the training dataset. x_n is the n^{th} sample and $y_n\in[1,2,...,M]$ is the label of the M known classes in the training dataset. Correspondingly, $I'=\{x'_1,...,x'_n,x_{n+1},...,x_m\}$ is the testing dataset, which includes the samples of known classes and samples of new classes. The task is to achieve an accurate diagnosis of known class samples $\{x'_1,...,x'_n\}$ and identify samples $\{x_{n+1},...,x_m\}$ that are from new classes.

B. Evidence theory

Dempster-Shafer evidence theory (*DST*) is a generalization of Bayesian theory to include subjective probabilities [19]. On assigning belief mass to subsets of the discriminative framework, the belief truth can be any of the possible states. Then, subjective logic (*SL*) formalizes *DST*'s notion of belief distribution over a discernment framework as a Dirichlet distribution. The theoretical framework to quantify belief mass and uncertainty based on the principles of evidence theory can be found [16]. For each sample, *SL* provides a belief mass a_m and uncertainty w, satisfying the following equation:

$$w+\sum\nolimits_{m=1}^{M}a_m=1 \tag{1}$$

where $a_m=\frac{v_m}{\sum\nolimits_{m=1}^{M}(v_m+1)},w=\frac{M}{\sum\nolimits_{m=1}^{M}(v_m+1)}$. v_m means the evidence of the m^{th} class. $\mathbf{v}=[v_1,..,v_M]$ is the evidence vector. It is evident that there is an inverse relationship between belief mass and uncertainty. Typically, a sample will have a low level of uncertainty when it obtains sufficient evidence supporting or belief mass. In contrast, a diagnosis result accompanied by a high uncertainty value is a lack of evidence supporting the classification as a known class. Then, the sample is identified as an unknown class.

C. The Dirichlet distribution integrated into evidence theory

In the classification task, the i^{th} data sample has the observation x_i and the class label y_i. This label corresponds to a latent class distribution $\mathbf{p} = [p_1, ..., p_M]$, representing the probability over M categories. Neural network classifiers can estimate the probability by using the *softmax* function. However, *softmax* provides a point estimate for the class probability of a sample without associated uncertainty, which introduces the risk of an overconfident diagnosis. Dirichlet distribution is used as a distribution of all possible *softmax* outputs for the classification of any given samples. The Dirichlet distribution is a probability density function for categorical distributions and can be characterized by parameter $\boldsymbol{\beta} = [\beta_1, ..., \beta_M]$ [20], given by

$$D(\mathbf{p} \mid \boldsymbol{\beta}) = \begin{cases} \dfrac{1}{B(\boldsymbol{\beta})} \prod_{i=1}^{M} p_i^{\beta_i - 1} & for\ \mathbf{p} \in V_M, \\ 0 & otherwise \end{cases} \tag{2}$$

where V_M is the K-dimensional unit simplex, given by

$$V_M = \left\{ \mathbf{p} \mid \sum_{i=1}^{M} p_i = 1, 0 \le p_i \le 1 \right\} \tag{3}$$

The Dirichlet distribution parameter can be viewed as real-valued pseudocounts or evidence, where the higher pseudocounts indicate more evidence over classes [8, 21]. If there is no evidence for the assignment of the sample to classes, Dirichlet distribution can be seen as a uniform prior $D(\mathbf{p} \mid 1, ..., 1)$ (i.e. $\boldsymbol{\beta}_0 = [1, ..., 1]$). As a result, its belief mass value is zero and the uncertainty value is one. When the evidence v_m exists over classes, the relevant parameter would be updated ($\beta_m = 1 + v_m$), generating the new Dirichlet distribution $D(\mathbf{p} \mid \boldsymbol{\beta}') = D(\mathbf{p} \mid \mathbf{v}{+}\boldsymbol{\beta}_0)$.

The mean and the variance of a Dirichlet distribution for the class probability p_m are computed as follows:

$$\hat{p}_m = \mathrm{E}[p_m] = \frac{\beta_m}{V},\ \mathrm{V}[p_m] = \frac{\beta_m (V - \beta_m)}{V^2 (V+1)} \tag{4}$$

where $V = \sum_{m=1}^{M} \beta_m$ refers to the Dirichlet strength (or the total evidence of the sample). $\hat{p}_m$ means the expected probability of the m^{th} class.

III. The Proposed Method

A. The evidential deep classifier

In this paper, the evidence theory has been integrated to develop the evidential VGG networks, which can provide reliable diagnosis results by assigning more evidence to the correct labels and decreasing the misleading evidence from misclassified samples. The logic of the evidential deep classifier is presented in **Fig. 2.** The procedure is summarized in **Algorithms 1**.

The deep classifier first estimates evidence $\mathbf{v}$ over each class. Then, the belief mass $\mathbf{a}$ is obtained which can be used to calculate the uncertainty $\mathbf{w}$ of a sample. With the prior parameter $\boldsymbol{\beta}_0$, the uniform Dirichlet distribution is generated. With the new evidence, the Dirichlet distribution will be updated which generates class probabilities $\mathbf{p}$. The predicted label y with uncertainty value provides a trustworthy prediction. The VGG-architecture network, an improved convolutional neural network (CNN), has been modified and used as the basic classifier [18].

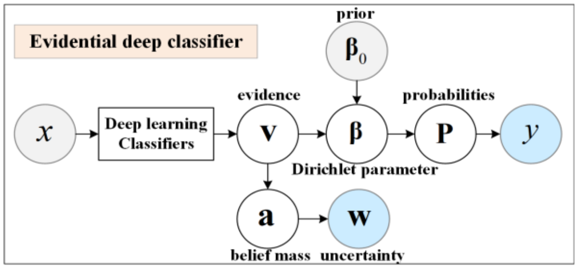


Fig. 2. The logic of the evidential deep classifier.

Algorithm 1: Evidential deep classifier for open-set recognition

```
Input: dataset $\mathbf{X}$, prior Dirichlet distribution parameter $\boldsymbol{\beta}_0$
1: Obtain evidence $\mathbf{v}$ on each class by a deep classifier.
2. Compute the belief mass $a_m = v_m / \sum_{i=1}^{M}(v_m + 1)$ and uncertainty
   $w = 1 - \sum_{m=1}^{M} a_m$.
3: Update Dirichlet distribution $D(\mathbf{p} \mid \boldsymbol{\beta})$ by the updated Dirichlet
   distribution parameter $\boldsymbol{\beta}{=}\mathbf{v}{+}\boldsymbol{\beta}_0$.
4. Compute the class probability $\hat{p}_m = \beta_m / \sum_{m=1}^{M} \beta_m$.
Return: predicted label $\mathbf{y}$, uncertainty estimation value $\mathbf{w}$
```

B. The evidential loss function

For i^{th} sample, x_i is the observation and y_i is the class label that is one hot encoding. This paper defines the evidential loss function using the maximum likelihood loss (MLL), the cross-entropy loss (CEL), and the sum-of-squares loss (SSL). In this paper, the selection of the optimal evidential loss function is incorporated into the hyperparameter optimization process. The maximum likelihood results in

$$\begin{aligned} \mathrm{L}_i(\Theta) &= -\log\left(\int \prod_{m=1}^{M} p_{im}^{y_{im}} \frac{1}{B(\boldsymbol{\beta}_i)} \prod_{m=1}^{M} p_{im}^{\beta_{im}-1} d\mathbf{p}_i\right) \\ &= -\sum_{m=1}^{M} \log\left(\int p_{im}^{y_{im}} \frac{1}{B(\boldsymbol{\beta}_i)} \prod_{m=1}^{M} p_{im}^{\beta_{im}-1} d\mathbf{p}_i\right) \\ &= -\sum_{m=1}^{M} y_{im} \log(\hat{p}_m) \\ &= -\sum_{m=1}^{M} y_{im} \log\left(\frac{\beta_{im}}{V_i}\right) \\ &= \sum_{m=1}^{M} y_{im} (\log(V_i) - \log(\beta_{im})) \end{aligned} \tag{5}$$

where Θ represents the network parameters. $B(\boldsymbol{\beta}_i)$ is the K-dimensional multinomial beta function. The loss function is minimized by searching parameters $\boldsymbol{\beta}_i$. For the cross-entropy loss, similarly, the Bayes risk concerning the class predictor is:

$$\begin{aligned} \mathrm{L}_i(\Theta) &= \int \left[\sum_{m=1}^{M} -y_{im} \log(p_{im})\right] \prod_{m=1}^{M} p_{im}^{\beta_{im}-1} d\mathbf{p}_i \\ &= \sum_{m=1}^{M} y_{im} (\psi(V_i) - \psi(\beta_{im})) \end{aligned} \tag{6}$$

where $\psi(\cdot)$ is the digamma function. The same approach can be applied to compute the sum-of-squares $\|\mathbf{y}_i - \mathbf{p}_i\|_2^2$ using

$$
\begin{aligned}
\mathrm{L}_i(\Theta) &= \int \|\mathbf{y}_i - \mathbf{p}_i\|_2^2 \frac{1}{B(\boldsymbol{\beta}_i)} \prod_{i=1}^{M} p_{im}^{\beta_{im}-1} d\mathbf{p}_i \\
&= \sum_{m=1}^{M} E\,[y_{im}^2 - 2y_{im}p_{im} + p_{im}^2] \\
&= \sum_{m=1}^{M} (y_{im}^2 - 2y_{im}E\,[p_{im}] + E\,[p_{im}^2]) \\
&= \sum_{m=1}^{M} (y_{im}^2 - E\,[p_{im}])^2 + \mathrm{var}[p_{im}]) \\
&= \sum_{m=1}^{M} (y_{im} - \beta_{im}/V_i)^2 + \frac{\beta_{im}(V_i - \beta_{im})}{V_i^2(V_i+1)} \\
&= \sum_{m=1}^{M} \underbrace{(y_{im} - \hat{p}_{im})^2}_{\mathrm{L}_m^{err}} + \underbrace{\frac{p_{im}(1-p_{im})}{V_i+1}}_{\mathrm{L}_{im}^{\mathrm{var}}}
\end{aligned} \tag{7}
$$

L_{im}^{err} is the fitting error in the data prediction, and L_{im}^{var} is the variance of the Dirichlet distribution generated by the EVGG model. The SSL used in the paper is to minimize the sum of error and variance.

To reduce the evidence assigning for misclassification classes, KL divergence is introduced to the loss function as a penalty term, reducing the evidence for misclassification classes. It measures the difference between the target Dirichlet distribution and the prior Dirichlet distribution, calculated as follows:

$$KL[D(\mathbf{p}_i \,|\, \tilde{\boldsymbol{\beta}}_i) \| D(\mathbf{p}_i \,|\, \boldsymbol{\beta}_0)], \;\; \tilde{\boldsymbol{\beta}}_i = \mathbf{y}_i + (1-\mathbf{y}_i) \square \; \boldsymbol{\beta}_i \tag{8}$$

where □ represents the element-wise product, $\Gamma(\cdot)$ is the gamma function, and $\tilde{\boldsymbol{\beta}}_i$ represents the updated Dirichlet parameter. Then evidential loss can then be calculated as

$$\mathrm{L}_{total}(\Theta) = \sum_{i=1}^{N} \mathrm{L}_i(\Theta) + \lambda_t \sum_{i=1}^{N} KL[D(\mathbf{p}_i \,|\, \tilde{\boldsymbol{\beta}}_i) \| D(\mathbf{p}_i \,|\, \boldsymbol{\beta}_0)] \tag{9}$$

where $\lambda_t = \min(1, t/10) \in [0,1]$ represents the annealing coefficient, t is an index for the current training epoch.

C. The risk-calibrated evidential loss function

To measure the uncertainty and risk in assigning the evidence over classes, pignistic probability (***q***) has been introduced to calculate the risk of misclassification [22]. The pignistic probabilities over classes can be set as a Dirichlet distribution[20] using:

$$g_\theta(\mathbf{q} \,|\, \mathbf{x}) = D(\mathbf{q} \,|\, \boldsymbol{\beta}) = D(\mathbf{q} \,|\, \mathbf{v}_\theta(\mathbf{x}) + \boldsymbol{\varphi}_\theta(\mathbf{x})) \tag{10}$$

where $\mathbf{v}_\theta(\mathbf{x})$ means evidence over classes. $\boldsymbol{\varphi}_\theta(\mathbf{x})$ is the prior count for the sample, computed by:

$$\boldsymbol{\varphi}_\theta(\mathbf{x}) = M * f_\theta(\mathbf{W} * f'_\theta(\mathbf{x}) + \mathbf{b}), \sum\nolimits_i \varphi_{\theta j}(\mathbf{x}) = M \tag{11}$$

in which $f'_\theta(\cdot)$ is the output of the logits layer, ***W*** and ***b*** are the weight and bias variables, respectively. $f_\theta(\cdot)$ is the activation function e.g., *softmax* function. The average risk of misclassified samples is given by

$$risk(\mathbf{x}) = \sum\nolimits_{i=1}^{M} R_{yi} q_i \tag{12}$$

where R_{yi} represents the risk of misclassifying sample *x* from class *y* to class *i*. $\mathbf{R} \in [0,\infty)^{M \times M}$ is the *M*-dimensional non-negative square matrix based on subjective criteria. When the sample is correctly classified, the risk R_{yy} decreases to zero. Relying on the definition of the Dirichlet distribution parameter $\beta_i = v_{\theta i}(\mathbf{x}) + \varphi_{\theta i}(\mathbf{x})$ in pignistic probabilities, the expected risk over classes can be computed by

$$E\,[risk(x)] = \frac{\sum_{i=1}^{M} R_{yi}(v_{\theta i}(\mathbf{x}) + \varphi_{\theta i}(\mathbf{x}))}{M + \sum_{i=1}^{M} \varphi_{\theta i}(\mathbf{x})} \tag{13}$$

This paper integrates misclassification risk into the loss function to form a risk-calibrated evidential loss function, minimizing the risk of misclassification. The loss function is

$$
\begin{aligned}
\mathrm{L}_{total}(\Theta) = &\sum_{i=1}^{N} \mathrm{L}_i(\Theta) + \lambda_t \sum_{i=1}^{N} KL[D(\mathbf{p}_i \,|\, \tilde{\boldsymbol{\beta}}_i) \| D(\mathbf{p}_i \,|\, \boldsymbol{\beta}_0)] \\
&+ E\,[risk(x)]
\end{aligned} \tag{14}
$$

The *L2* regularization has been introduced to regularize the weights of the fully connected layer, thereby reducing model complexity and preventing overfitting.

D.EVGG with uncertainty estimation in OSFD

The framework of the proposed trustworthy intelligent fault diagnosis approach with uncertainty estimation, called EVGG, is shown in **Fig. 3.** The proposed method outputs the classification result together with a prediction uncertainty. The EVGG model as the multiclass classifier achieves accurate classification with test samples from known classes and effective detection of unknown classes with high uncertainty values. Quantifying uncertainty can avoid overconfident risk and make reliable fault diagnosis decisions in practical systems. The procedure is summarized as follows:

Step 1: Measure and collect vibration data of the monitored machine with different fault conditions. Preprocess the data and construct the two-dimensional feature maps of the samples to form the dataset.

Step 2: Construct EVGG by integrating evidence theory and an improved VGG-architecture network. Use the risk-calibrated evidential loss function as the loss function. To reduce the complexity of models, L2 regularization is applied to the weights of full-connected layers. Category probabilities and pignistic probabilities can be obtained.

Step 3: Train the EVGG model with the training samples of the known classes. Test the proposed method with samples from known classes and unknown classes. The EVGG model can achieve good diagnostic accuracy for known classes and detect unknown samples with higher uncertainty values.

Step 4: The trained EVGG is deployed to achieve fault diagnosis. With unknown faults detected, further operational warnings and maintenance decisions for the machine will be triggered.

Step 5: With the cumulated unknown samples detected, the model parameters can then be adaptively optimized by maximizing the successful detection of collected samples of unknown classes. The EVGG model can be continuously optimized for online fault diagnosis.

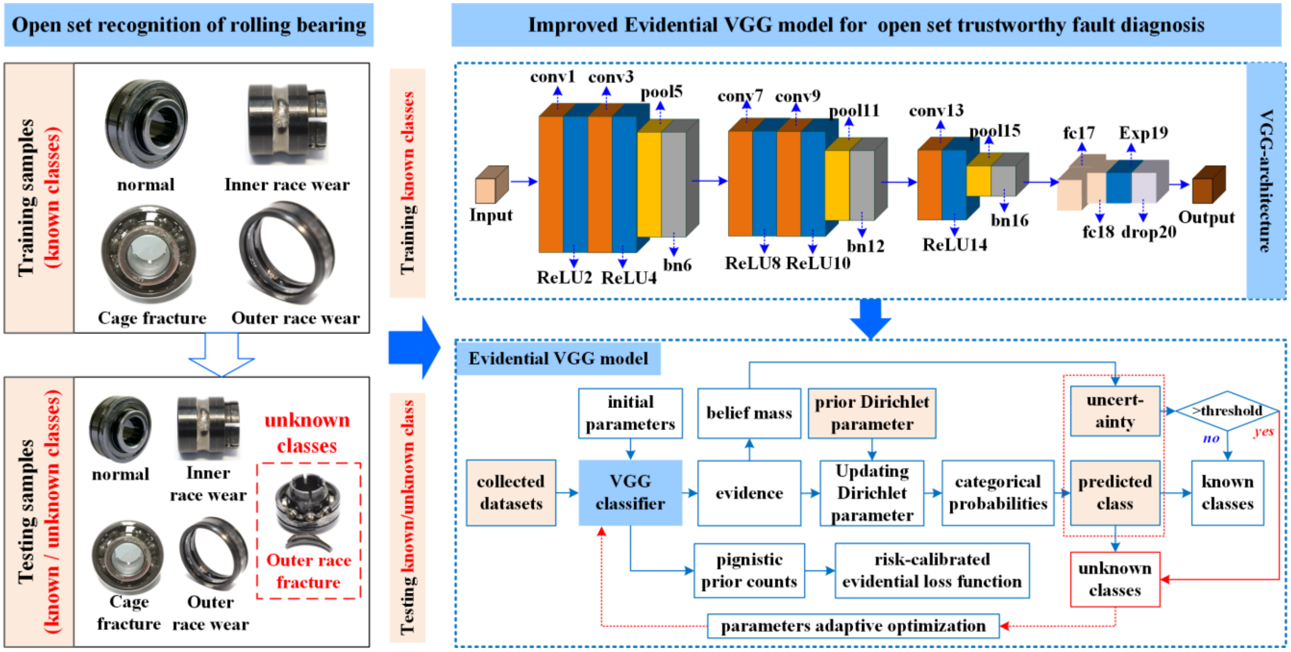


Fig. 3. The framework of the proposed EVGG method for open set fault diagnosis.

IV. Experimental Validation

CASE 1 Fault diagnosis of CWRU Data

A. CWRU Fault Dataset Description

In this experiment, the roller bearing dataset collected from a motor drive system by Case Western Reserve University (CWRU) is used. The test stand is shown in **Fig. 4**. Vibration signals of the drive-side (6205-2RS JEM SKF) bearing were acquired at a sampling frequency of 12 kHz under a 1-hp load (1772 rpm). The monitored conditions of the bearings include one normal condition and nine faulty conditions. Three types of faults are inner race fault (IRF), ball fault (BF), and outer race fault (ORF). Each failure type has three severity levels including 0.007, 0.014, and 0.021 inches that correspond to slight, moderate, and severe faults respectively.

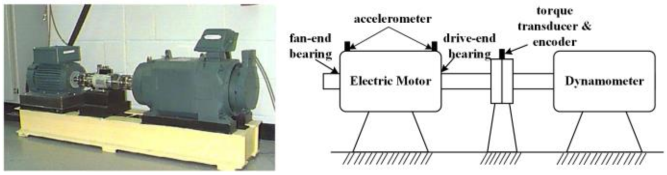


Fig. 4. Tested rolling-element bearing of CWRU.

In the data pre-processing, an image transformation method is adopted to convert the vibration signal to two-dimensional feature maps as the input of the EVGG model [23]. A vibration signal with a length of 784 data points is converted to a 28*28 two-dimensional feature map, details given in **Fig. 5**.

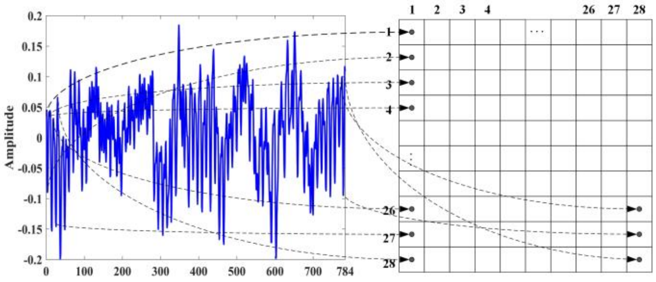


Fig. 5. Schematic diagram of data conversion.

To expand the number of training and testing samples, the overlap sliding segmentation method [24] is used. Each segmentation signal includes 784 sample points and the length of sample overlap for two neighbor segments is 684 sample points. For each class, 1000 samples are obtained. Then dataset is split randomly into subsets of training data (70%) and testing data (30%), as given in **TABLE I**.

TABLE I
Description of CWRU Datasets

Fault level	Fault type	Number of the training/testing samples	Class Label
normal	/	700/300	C1
slight fault (0.007")	IRF	700/300	C2
	BF	700/300	C3
	ORF	700/300	C4
medium fault (0.014")	IRF	700/300	C5
	BF	700/300	C6
	ORF	700/300	C7
severe fault (0.021")	IRF	700/300	C8
	BF	700/300	C9
	ORF	700/300	C10

To test the performance of the proposed approach in OSFD tasks, three settings are designed as shown in **TABLE II**. For example, in task T_1, the training dataset contains normal class (C1), BF classes (C3, C6, and C9), and ORF classes (C4, C7, and C10). The testing dataset includes all the known classes and the unknown classes, IRF classes (C2, C5, and C8).

TABLE II
The Setting of OSFD Tasks in Case1

Scenarios setting	Tasks	Training dataset	Testing dataset
unknown IRF	T_1	C1,C3,C4,C6,C7,C9,C10	unknown：C2,C5,C8 plus all known classes
unknown BF	T_2	C1,C2,C4,C5,C7,C8,C10	unknown：C3,C6,C9 plus all known classes
unknown ORF	T_3	C1,C2,C3,C5,C6,C8,C9	unknown：C4,C7,C10 plus all known classes

B. Evaluation Metrics for OSFD tasks

To evaluate the classification results, several measures are chosen for known classes: accuracy, precision, recall, and F1 score which are widely used in the literature [25].

$$Accuracy = (TP + TN) / (TP + FP + FN + TN) \tag{15}$$

$$Precision = TP / (TP + FP) \tag{16}$$

$$Recall = TP / (TP + FN) \tag{17}$$

$$F1 = 2 * Precision * Recall / (Precision + Recall) \tag{18}$$

where TP, FP, FN, and TN represent the number of true positive, false positive, false negative, and true negative outcomes, respectively.

For the unknown classes, some quantities [4, 6] are defined:
TK : The number of correctly classified known classes;
FU : The number of misclassified known classes;
TK : The number of successfully detected unknown samples;
FU : The number of fairly detect unknown samples.

TU and *FU* rely on the uncertainty threshold u_θ that needs to be determined first. Given sample *i*, its uncertainty u_i can be calculated by Eq.(4). If $u_i \le u_\theta$, a sample will be identified as a known class and otherwise an unknown class. In this paper, we define u_θ to be the mean value plus one standard deviation of the uncertainty estimations from all samples of known classes in the training process which is more effective than a fixed threshold [26]. Then, several evaluation metrics are employed. The accuracy of known classes *AKS* can be computed by:

$$AKS = TK / (TK + FK) \tag{19}$$

The accuracy of unknown classes *AUS* is given by:

$$AUS = TU / (TU + FU) \tag{20}$$

In addition, the accuracy of all testing samples *ALL*, including known classes and unknown samples, is given by:

$$ALL = (TK + TU) / (TK + FK + TU + FU) \tag{21}$$

Another comprehensive metric *H-score* is defined as:

$$H - score = 2 * AKS * AUS / (AKS + AUS) \tag{22}$$

C. Trustworthy diagnosis results of the CWRU dataset

The hyperparameters of the EVGG model are determined through grid search using a training dataset. Theoretically, all the hyperparameters can be optimized which could be computationally expensive. Here, we first determine the main model structure based on VGG which is shown in **TABLE III**. Then, batch size and loss function type are selected as the hyperparameters to be searched to obtain the best prediction results including accuracy and uncertainty for known classes. The diagnostic performance using different batch sizes is displayed in **TABLE IV**. A batch size of 32 achieves the best accuracy and uncertainty prediction. The performance of different loss functions is shown in the lower section of **TABLE IV** where SSL achieves the best results.

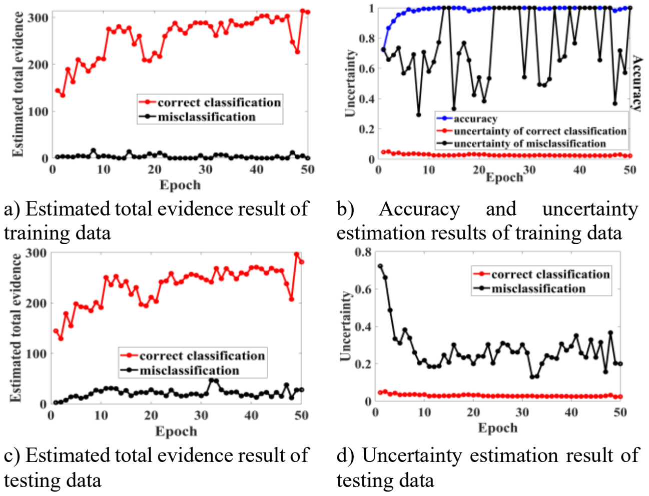
In the experiments, the Adam optimizer is used as the default setting for training. All experiments are carried out on a computer with an Intel Core i7 CPU, 16 GB RAM, and GeForce RTX 3050Ti GPU. The training process converges rapidly during 50 epochs with high training accuracy for the known classes, as shown in **Fig 6**. The total evidence of correct classification gradually increases, while the misclassification gets low evidence supporting as shown in subplots (a) and (c) of **Fig.6**. For the uncertainty estimation results in subplots (b) and (d) of **Fig.6**, the uncertainty value of misclassifications is significantly higher than the correct classification samples.

TABLE III
HYPERPARAMETERS OF THE MAIN MODEL STRUCTURE

Description	Value
input	28*28*1
convolution layers 1	Kernel 3*3*1*20, stride [1 1]
activation layers 2	ReLU
convolution layers 3	Kernel 3*3*1*20, stride [1 1]
activation layers 4	ReLU
pooling layers 5	Maximum pooling [2 2]
batch_normalization 6	Batch_normalization
convolution layers 7	Kernel 3*3*20*40, stride [1 1]
activation layers 8	ReLU
convolution layers 9	Kernel 3*3*20*40, stride [1 1]
activation layers 10	ReLU
pooling layers 11	Maximum pooling [2 2]
batch_normalization 12	Batch_normalization
convolution layers 13	Kernel 3*3*40*60, stride [1 1]
activation layers 14	ReLU
pooling layers 15	Maximum pooling [2 2]
batch_normalization 16	Batch_normalization
full-connected layers 17	500 fully connected layer
full-connected layers 18	100 fully connected layer
activation layers 19	Exponential
dropout 20	50% dropout
output	10 classes

TABLE IV
DIAGNOSTIC PERFORMANCE WITH DIFFERENT PARAMETERS FOR TASK T_1 (%)

Parameters		Training performance		Testing performance	
		Accuracy	U	Accuracy	U
batch size (MLL)	16	99.87	5.25	99.40	8.37
	32	**100.00**	**2.62**	**99.51**	**5.15**
	64	100.00	2.48	99.55	5.37
	128	100.00	2.98	99.13	7.14
loss function (batch size 32)	MLL	100.00	2.62	99.51	5.15
	CEL	99.95	4.06	99.40	6.02
	SSL	**100.00**	**3.09**	**99.74**	**5.32**



a) Estimated total evidence result of training data

b) Accuracy and uncertainty estimation results of training data

c) Estimated total evidence result of testing data

d) Uncertainty estimation result of testing data

Fig. 6. Training and testing process of task T_1.

The diagnostic performance of the model for known classes is compared and analyzed for three scenarios. Fault diagnosis performance is quantified using metrics of accuracy, recall, precision, F1 score, uncertainty(U), and F1 score [25], shown in **TABLE V**. The testing accuracy of the known classes in the three different tasks reaches 99.65%, 99.96%, and 99.47%, respectively. The average uncertainty values of known classes are 6.70%, 4.47%, and 7.14%, respectively. Overall, the average diagnostic accuracy is 99.69% with a low average uncertainty of 6.10%. The recall indicator achieves 98.76%, 99.86%, and 98.41%, respectively, which shows a good diagnosis performance. The results show that the proposed method can provide a reliable diagnostic result for known classes with high accuracy and low uncertainty.

TABLE V
FAULT DIAGNOSIS RESULTS ON KNOWN CLASSES (%)

Task		Accuracy	Recall	Precision	F1 score	U
T_2	C1	100.00	100.00	100.00	100.00	3.30
	C2	100.00	100.00	100.00	100.00	3.73
	C4	100.00	100.00	100.00	100.00	3.27
	C5	99.86	99.00	100.00	99.50	7.77
	C7	100.00	100.00	100.00	100.00	4.86
	C8	99.90	100.00	99.34	99.67	2.93
	C10	99.95	100.00	99.67	99.83	5.42
T_1 average		99.65	98.76	98.78	98.75	6.70
T_2 average		99.96	99.86	99.86	99.86	4.47
T_3 average		99.47	98.14	98.18	98.13	7.14
All average		99.69	98.92	98.94	98.91	6.10

The detection results of unknown class samples and comprehensive evaluation results of all testing samples that contain known classes and unknown samples are presented in **TABLE VI**. First, the proposed EVGG model has a strong detection capability of unknown classes with higher estimated uncertainty than the known classes. The average uncertainty of different tasks is 52.58%, 49.25%, and 52.61%, respectively. The average detection accuracy is 90.22%, 79.78%, and 86.78% for unknown IRF, unknown BF, and unknown ORF scenarios, respectively. The positive correlation observed here between uncertainty and the detection accuracy of unknown samples verifies the proper setting of the uncertainty threshold u_θ .

TABLE VI
UNKNOWN CLASS DETECTION AND EVALUATION RESULTS (%)

Task	Unknown samples		Total testing samples	
	AUS	U	All	H_score
T_1	90.22	52.58	96.20	94.30
T_2	79.78	49.25	93.83	88.70
T_3	86.78	52.61	94.73	92.11

Overall, the accuracy of all testing samples in the three tasks is 96.20%, 93.83%, and 94.73% respectively. The *H_score* is 94.30%, 88.70%, and 92.11% respectively. Comparing the metrics in the three tasks, it is clear that the reliability and robustness of the diagnostic results are all at a high level. To our limited knowledge, this is the first time that uncertainty estimation has been introduced into an OSFD task and achieved good diagnosis performance.

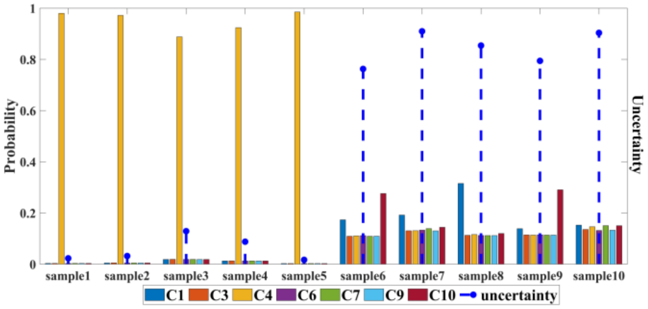


Fig. 7 Diagnostic probability assignment and uncertainty estimation of task T_1.

For task T_1, 5 samples from known classes (samples 1 to 5) and 5 samples from unknown classes (samples 6 to 10) are randomly selected and displayed. The classification probabilities and uncertainty estimates are shown in **Fig. 7**. Results of the known classes show that high classification probabilities are obtained for a specific class with low uncertainty. For the unknown classes, similar probabilities are assigned for each class with higher uncertainty than known classes. The results show that the proposed method can accurately classify known classes and identify unknown samples with uncertainty estimation.

After initial training, the EVGG model can detect samples of unknown classes with high uncertainty values. With cumulated samples of unknown classes being detected, the evaluation metrics listed in **TABLE IV** are used to adaptively optimize the hyperparameters of the EVGG models. To illustrate how the process works, we also select optimal batch size and loss function type as the parameters to be searched. The experimental results indicate a batch size of 32 and the loss function of SSL are still the optimal choices. This adaptive process can be continued with increasing identified unknown samples to continuously improve the model performance.

D. Comparison with state-of-the-art methods

To verify the diagnostic performance of the EVGG approach, other state-of-the-art methods are compared first under the condition that all classes are known. **TABLE VII** presents the comparative results. The proposed method achieves 99.89% diagnostic accuracy, which is better than other deep-learning methods. While the proposed trustworthy fault diagnosis framework has competitive diagnostic performance, VGG architecture needs more computational cost than other CNN-based methods, which should be circumvented by using more compact structures in the real-time application. In addition, the proposed method estimates a predictive uncertainty of 8.92%, which is close to the uncertainty of the known classes listed in **TABLE V** and significantly lower than the uncertainty value of the unknown classes shown in **TABLE VI**. Overall, the proposed method achieves good predictive accuracy and overcomes the overconfident diagnostic behavior by providing uncertainty indicators in OSFD tasks.

TABLE VII
DIAGNOSTIC PERFORMANCE COMPARISON ON KNOWN CLASSES

Method	Accuracy (%)	Uncertainty (%)
Proposed method	99.89	8.92
Deep Transfer Learning [27]	97.95	/
CWT- CNN-gcForest [28]	99.19	/
Improved-DCGAN [29]	99.80	/
MCNN-LSTM [25]	98.46	/
RNGPT-RBF [30]	99.39	/
SDANN [31]	93.84	/
KPCA-SAE-GPC [32]	94.29	/

Second, the effectiveness of the proposed method in OSFD tasks is validated compared with other advanced OSFD methods. The diagnostic tasks as shown in **TABLE VIII**. The results are presented in **TABLE IX**, in which the proposed method can provide additional uncertainty estimation values of the testing classes and unknown samples.

TABLE VIII
THE SETTING OF OSFD TASKS IN LITERATURE [6]

Task	Load	Training label	Testing label
K_0	0hp	C1,C2,C3,C5,C6,C8,C9	C1,C2,C3,C4,C6,C7,C9,C10
K_1	1hp	C1,C2,C3,C5,C6,C9	C1,C2,C3,C4,C6,C7,C9
K_2	2hp	C1,C2,C3,C5,C6,C9	C1,C3,C4,C8
K_3	3hp	C1,C3,C4,C5,C9,C10	C1,C2,C3,C5,C6,C9

According to the average *H_score* values, the proposed approach is superior to other OSFD methods. The compared results validate the effectiveness and robustness of the proposed method in different tasks. Moreover, it can realize trustworthy fault diagnosis in open set conditions through the correct classification of known classes and detection of unknown classes with high uncertainty values.

TABLE IX
DIAGNOSTIC PERFORMANCE COMPARISON ON UNKNOWN CLASSES (%)

Task	1DCNN +KNN	1DCNN +SVDD	1DCNN +EVT	Proposed method		
	H_score	H_score	H_score	H_score	U_{known}	$U_{unknown}$
K_0	68.1	84.7	**94.0**	91.2	6.9	48.6
K_1	77.4	90.1	**97.8**	92.4	7.9	50.0
K_2	51.6	77.0	80.7	**93.8**	3.0	44.3
K_3	67.9	76.5	85.8	**91.9**	5.7	26.0
Ave.	66.2	82.1	89.6	**92.3**	5.9	42.2

CASE 2 Fault diagnosis of XJTU-SY Data

A. XJTU-SY Fault Data Description

In this case study, the publicly available roller bearing dataset collected from a motor drive system (shown in **Fig. 8)** by Xi'an Jiaotong University and the Changxing Sumyoung Technology Co. (XJTU-SY) is used [33]. The type of tested bearings was LDK UER204. Vibration data were acquired at a sampling frequency of 25.6 kHz. There were three different operating conditions: condition 1, 2100 rpm; condition 2, 2250 rpm; condition 3, 2400 rpm. The horizontal vibration signals were selected which contain more bearing health information. Fault types include inner race fault (IRF), cage fault (CF), outer race fault (ORF), inner race and outer race fault (IORF), and mix fault (MF, including inner race, ball, cage, and outer race fault) as listed in **TABLE X**. The data contains one normal condition and six faulty conditions are selected from the dataset where each condition contains 1000 samples. Then, 70% of the data is used for training and 30% for testing.

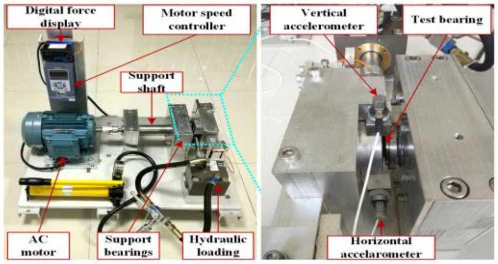


Fig. 8. Tested rolling-element bearing of XJTU-SY.

TABLE X
DESCRIPTION OF XJTU-SY DATASETS

Fault type	Bearing	Operating	Files number	Class Label
normal	Bearing 2_3	Condtion 2	10-13	C1
ORF	Bearing 2_5	Condtion 2	174-177	C2
IRF	Bearing 2_1	Condtion 2	464-467	C3
CF	Bearing 2_3	Condtion 2	494-497	C4
IORF	Bearing 1_5	Condtion 1	35-38	C5
MF	Bearing 3_2	Condtion 3	2268-2271	C6

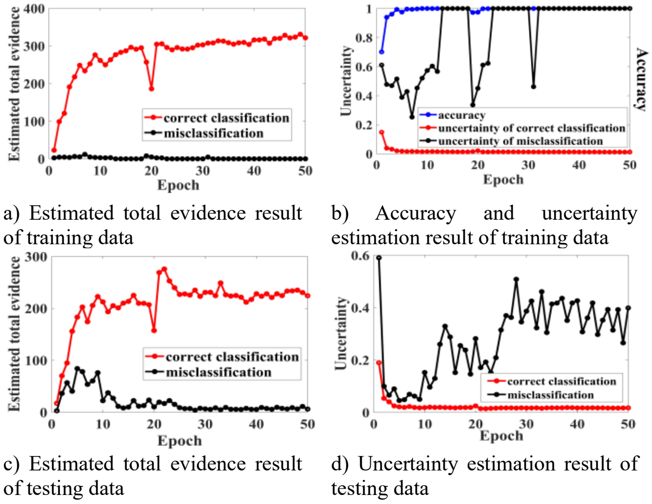
TABLE XI
THE SETTING OF OSTFD TASKS OF CASE2

Scenarios setting	Task	Training dataset	Testing dataset
unknown IORF	X_1	C1,C2,C3,C4	unknown: C5 plus all known classes
unknown MF	X_2	C1,C2,C3,C4	unknown: C6 plus all known classes
unknown IORF+MF	X_3	C1,C2,C3,C4	unknown:C5,C6 plus all known classes

The description of scenario settings for OSFD tasks is illustrated in **TABLE XI**. The training dataset includes known classes (C1, C2, C3, and C4) and the testing dataset includes known classes and unknown classes.

B. Trustworthy diagnosis results of the XJTU-SY dataset

Fig. 9 presents the effectiveness of trustworthy diagnosis in the training and testing process for task X_3 using the proposed method. The differences in evidence support and uncertainty value between correct classification and misclassification are evident.



a) Estimated total evidence result of training data

b) Accuracy and uncertainty estimation result of training data

c) Estimated total evidence result of testing data

d) Uncertainty estimation result of testing data

Fig. 9. Training and testing process of task X_3.

The diagnostic performance on known classes is presented in **TABLE XII.** The average accuracy on all tasks using the proposed EVGG method is 99.56% and the average uncertainty is 5.73%. Recall on diagnosis tasks of X_1, X_2, and X_3, is 99.75%, 99.33%, and 98.25%, respectively. Experiment results validate that the proposed method can correctly classify known classes and provide reliable evidence support.

TABLE XII
FAULT DIAGNOSIS RESULTS ON KNOWN CLASSES (%)

Task		Accuracy	Recall	Precision	F1 score	U
X_1	C1	99.75	99.33	99.67	99.50	6.29
	C2	100	100	100	100	3.90
	C3	99.92	99.67	100	99.83	7.82
	C4	99.83	100	99.34	99.67	5.67
X_1 average		99.88	99.75	99.75	99.75	5.92
X_2 average		99.67	99.33	99.34	99.33	5.02
X_3 average		99.13	98.25	98.26	98.25	6.25
All average		**99.56**	**99.11**	**99.12**	**99.11**	**5.73**

TABLE XIII
UNKNOWN CLASS DETECTION AND EVALUATION RESULTS (%)

Task	Unknown samples		Total testing samples	
	AUS	U	All	H_score
X_1	86.00	41.38	97.00	92.37
X_2	85.33	25.82	96.53	91.80
X_3	78.17	35.74	91.56	87.07
average	**83.17**	**34.31**	**95.03**	**90.41**

TABLE XIII presents the detection capability of unknown classes, with the comprehensive evaluation results of all testing samples. AUS on the three tasks is 86%, 85.33%, and 78.17%, respectively, which demonstrates the detection capability of an unknown class of the proposed. The uncertainty of unknown classes (34.31%) is significantly higher than that of known classes (5.73% in **TABLE XII**). For all test samples, the proposed method achieves good diagnosis in known classes and can detect unknown classes with 95.03% *All* and 90.41% *H_score*.

For task X_1, 5 samples from known classes (samples 1 to 5) and 5 samples from unknown classes (samples 6 to 10) are randomly selected and displayed. The classification probabilities and uncertainty estimates are shown in **Fig. 10**. For known classes, a high probability is obtained for a specific class with low uncertainty. For the unknown classes, the classification probability assigned to the four known classes is similar. Therefore, higher uncertainty estimation values are obtained. The proposed method can effectively distinguish known and unknown samples with uncertainty estimation.

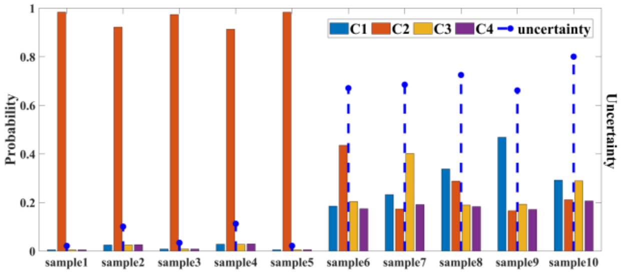


Fig.10. Diagnostic probability assignment and uncertainty estimation of task X_1.

To test model robustness, ten trials are run with all testing samples and the obtained uncertainty estimations are shown in **Fig. 11**. The average uncertainty estimation of known classes for normal, ORF, IRF, and CF is 6.06%, 2.65%, 4.75%, and 6.05%, respectively. The average uncertainty of the unknown class is nearly 40%, significantly higher than that of the known classes. The experiment results demonstrate a reliable uncertainty estimation.

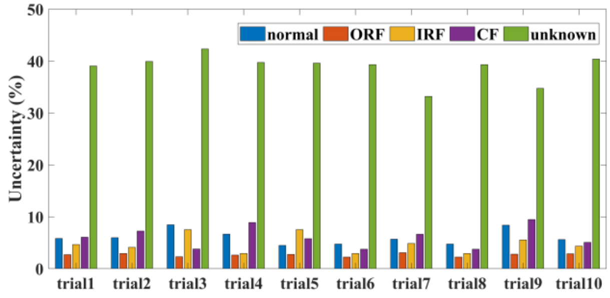


Fig. 11. The repeated results for uncertainty estimation in task X_1.

C. Comparison with state-of-the-art methods

The diagnosis performance on known classes is compared with state-of-the-art methods as shown in **TABLE XIV.** The proposed method achieves a better accuracy of fault diagnosis. Also, the predictive uncertainty has been quantified as an important guarantee for trustworthy diagnosis by using the EVGG model.

TABLE XIV
DIAGNOSTIC PERFORMANCE COMPARISON ON KNOWN CLASSES

Method	Accuracy (%)	Uncertainty (%)
Proposed method	99.93	5.95
SDANN [34]	93.84	/
MDAAN[35]	95.45	/
CWT- CNN-gcForest [28]	99.80	/
improved DCGAN [29]	98.99	/
SDANN [31]	98.97	/

V. CONCLUSIONS

In this paper, an improved evidential VGG (EVGG) method based on evidence theory was developed for trustworthy fault diagnosis, overcoming the overconfident prediction of existing approaches in open set fault diagnosis tasks. The risk-calibrated evidential loss function that can assign more evidence to the correct classification and decrease the misleading evidence from misclassified samples was developed. The EVGG model can predict not only the classification probability but also can estimate the prediction uncertainty, which avoids overconfident and undesirable misclassification results. Experimental studies on two rolling bearing fault diagnosis datasets verified the fault diagnosis performance of the proposed approach. It achieved accurate predictions of the known classes and detected unknown classes with high uncertainty values. The comparison between the proposed method and the state-of-the-art methods showed that the improved EVGG model could achieve higher performance and more reliable diagnosis results with effective uncertainty estimation. It showed high potential in detecting the out-of-distribution samples and provides trustworthy prediction results in open set fault diagnosis.

The limitations of this study should be addressed in future studies. Given the accuracy and robustness, the VGG-architecture was chosen as the basic classifier. However, VGG is comparatively complex with more computational resources needed. So, more compact structures should be developed to achieve faster training and online testing. In addition, the setting of misclassifying risk matrix still relies on expert knowledge, which can be improved. Furthermore, the application of the proposed approach in real industrial equipment or production system needs to be verified. Furthermore, the application of the proposed approach in real industrial equipment or production system needs to be verified where strong noise inference should be considered.

REFERENCES

[1] Y. Ovadia *et al.*, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.

[2] S. Yin, J. J. Rodriguez-Andina, and Y. Jiang, "Real-time monitoring and control of industrial cyberphysical systems: With integrated plant-wide monitoring and control framework," *IEEE Industrial Electronics Magazine*, vol. 13, no. 4, pp. 38-47, 2019.

[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. J. a. p. a. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv: 1606.06565*, 2016.

[4] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE transactions on pattern analysis machine intelligence*, vol. 43, no. 10, pp. 3614-3631, 2020.

[5] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv: 2110.11334*, 2021.

[6] X. Yu *et al.*, "Deep-learning-based open set fault diagnosis by extreme value theory," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 1, pp. 185-196, 2021.

[7] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14927-14937, 2020.

[8] M. Sensoy, L. Kaplan, F. Cerutti, and M. Saleki, "Uncertainty-aware deep classifiers using generative models," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020, vol. 34, no. 04, pp. 5620-5627.

[9] M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion,* vol. 76, pp. 243-297, 2021.

[10] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International conference on machine learning*, 2016: PMLR, pp. 1050-1059.

[11] Y. Mae, W. Kumagai, and T. Kanamori, "Uncertainty propagation for dropout-based Bayesian neural networks," *Neural Networks,* vol. 144, pp. 394-406, 2021.

[12] T. Han and Y.-F. Li, "Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles," *Reliability Engineering & System Safety,* vol. 226, p. 108648, 2022.

[13] J. Caceres, D. Gonzalez, T. Zhou, and E. L. Droguett, "A probabilistic Bayesian recurrent neural network for remaining useful life prognostics considering epistemic and aleatory uncertainties," *Structural Control and Health Monitoring,* vol. 28, no. 10, p. e2811, 2021.

[14] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in neural information processing systems,* vol. 31, 2018.

[15] B. Wang, Z. Li, Z. Dai, N. Lawrence, and X. Yan, "Data-driven mode identification and unsupervised fault detection for nonlinear multimode processes," *IEEE Transactions on Industrial Informatics,* vol. 16, no. 6, pp. 3651-3661, 2019.

[16] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in Neural Information Processing Systems,* vol. 31, 2018.

[17] A. Malinin and M. Gales, "Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness," *Advances in Neural Information Processing Systems,* vol. 32, 2019.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[19] A. Chowdhury, G. Karmakar, J. Kamruzzaman, and S. Islam, "Trustworthiness of self-driving vehicles for intelligent transportation systems in industry applications," *IEEE Transactions on Industrial Informatics,* vol. 17, no. 2, pp. 961-970, 2020.

[20] M. Sensoy, M. Saleki, S. Julier, R. Aydogan, and J. Reid, "Misclassification risk and uncertainty quantification in deep classifiers," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2484-2492.

[21] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[22] D. Dubois, H. Prade, and P. Smets, "A definition of subjective possibility," *International journal of approximate reasoning,* vol. 48, no. 2, pp. 352-364, 2008.

[23] M. Xia, T. Li, L. Xu, L. Liu, and C. W. De Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME transactions on mechatronics,* vol. 23, no. 1, pp. 101-110, 2017.

[24] H. Han, H. Wang, Z. Liu, and J. Wang, "Intelligent vibration signal denoising method based on non-local fully convolutional neural network for rolling bearings," *ISA transactions,* vol. 122, pp. 13-23, 2022.

[25] X. Chen, B. Zhang, and D. Gao, "Bearing fault diagnosis base on multi-scale CNN and LSTM model," *Journal of Intelligent Manufacturing,* vol. 32, no. 4, pp. 971-987, 2021.

[26] D. Chakraborty and H. Elzarka, "Early detection of faults in HVAC systems using an XGBoost model with a dynamic threshold," *Energy Buildings,* vol. 185, pp. 326-344, 2019.

[27] X. Li, W. Zhang, Q. Ding, and X. Li, "Diagnosing rotating machines with weakly supervised data using deep transfer learning," *IEEE transactions on industrial informatics,* vol. 16, no. 3, pp. 1688-1697, 2019.

[28] Y. Xu, Z. Li, S. Wang, W. Li, T. Sarkodie-Gyan, and S. Feng, "A hybrid deep-learning model for fault diagnosis of rolling bearings," *Measurement,* vol. 169, p. 108502, 2021.

[29] B. Zhao and Q. Yuan, "Improved generative adversarial network for vibration-based fault diagnosis with imbalanced data," *Measurement,* vol. 169, p. 108522, 2021.

[30] J. Chen *et al.*, "Gaussian process kernel transfer enabled method for electric machines intelligent faults detection with limited samples," *IEEE Transactions on Energy Conversion,* vol. 36, no. 4, pp. 3481-3490, 2021.

[31] W. Mao, Y. Liu, L. Ding, A. Safian, and X. Liang, "A new structured domain adversarial neural network for transfer fault diagnosis of rolling bearings under different working conditions," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-13, 2020.

[32] M. Liang and K. Zhou, "Probabilistic bearing fault diagnosis using Gaussian process with tailored feature extraction," *The International Journal of Advanced Manufacturing Technology,* vol. 119, no. 3, pp. 2059-2076, 2022.

[33] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Transactions on Reliability,* vol. 69, no. 1, pp. 401-412, 2018.

[34] W. Mao, Y. Liu, L. Ding, A. Safian, and X. Liang, "A new structured domain adversarial neural network for transfer fault diagnosis of rolling bearings under different working conditions," *IEEE Transactions on Instrumentation Measurement,* vol. 70, pp. 1-13, 2020.

[35] Z. Huang *et al.*, "A multi-source dense adaptation adversarial network for fault diagnosis of machinery," *IEEE Transactions on Industrial Electronics,* 2021.