An Attention-Based Meta Learning Network via Prior Meta-training Strategy for Intelligent Fault Detection of Shipboard Antenna under Small Samples Prerequisite

Yuanhong Chang

Xi'an Jiaotong University

Jinglong Chen (✉ jlstrive2008@163.com)

Xi'an Jiaotong University

Shuilong He

Guilin University of Electronic Technology

Kaiyu Zhang

Xi'an Jiaotong University

Tongyang Pan

Xi'an Jiaotong University

Original Article

Keywords: Shipboard antenna, Intelligent fault diagnosis, Small samples, Meta learning, Attention mechanism

Posted Date: August 31st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-845702/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

An Attention-Based Meta Learning Network via Prior Meta-training Strategy for Intelligent Fault Detection of Shipboard Antenna under Small Samples Prerequisite

Yuanhong Chang¹, Jinglong Chen^{1, *}, Shuilong He², Kaiyu Zhang¹ and Tongyang Pan¹

Abstract

Fault detection of shipboard antenna is of great significance to ensure the safe operation and smooth completion of astronautic measurement ship. With the development of data-oriented technology, intelligent fault detection is desiderated to improve self-management of entire shipboard antenna system. However, insufficient fault data results in intelligent algorithms stagnation. In this paper, a meta-learning network is specially designed for fault identification of shipboard antenna under small samples prerequisite, which is named affiliation network (AN). The AN consists of a random sampler, a feature extractor, an auxiliary classifier and a discriminator. The former three are utilized to extract and concatenate the features from training and testing samples, while the latter trains an adaptive pseudo-distance to evaluate the affiliation degree between concatenation features for identifying unknown data. Besides, a prior sufficient meta-training strategy is specially designed to realize metric-based knowledge transfer for acquiring the more generic AN, thus avoiding reiterative training of the AN in different application scenarios. Effectiveness of proposed method are validated by three experimental cases. Results indicate that, comparing with conventional intelligent models, the prior trained AN only utilized few samples to effectively identify failure categories of shipboard antenna even with complex operating conditions.

Keywords: Shipboard antenna; Intelligent fault diagnosis; Small samples; Meta learning; Attention mechanism.

* Correspondence: jlstrive2008@163.com

¹ State Key Laboratory for Manufacturing and Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, PR China

² School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology, Guilin 541004, China

Full list of author information is available at the end of the article

1. Introduction

Nowadays, aerospace engineering has gradually become the focus of international attention [1]. With the frequent implementation of aerospace launch missions, astronautic measurement ship, as an indispensable link is mainly responsible for the completion of space and ground information transmission and satellite positioning tasks. Among them, the shipborne antenna plays a key role in receiving and transmitting information. As the key drive units of shipboard antenna, bearings and gearboxes would inevitably exist various defects after continuous long-term operation [2]. At the same time, the harsh environment of the ocean, such as sea beaten and seawater corrosion, increases the risk of damage to these key mechanical components, which can easily cause the breakdown of the entire communication system. Consequently, the research on fault diagnosis of shipboard antenna is greatly significant for its health maintenance, so as to ensure the successful completion of the aerospace measurement mission of astronautic measurement ship.

Traditional diagnostic techniques generally rely on classical signal analysis to extract artificial characteristics, such as time-domain & frequency-domain analysis [3-4]. Due to the low efficiency of traditional diagnosis algorithms, it is difficult to play its role in current big data era [5-6]. Therefore, the technique of intelligent data analysis is considered as a high-efficiency way to maintain performance and detect failures of shipboard antenna system [7-8]. For example, Chang et al. [9] designed a parallel convolutional structure based on inner product principles, which was used to diagnose bearing faults of generator reducer under different operating conditions. Chen et al. [10] proposed a multi-kernel convolutional neural network (CNN) combined the boosting algorithm, which was utilized to locate and identify various fault impulses of rolling bearings. Nevertheless, the effectiveness of these methods depends on adequately training on massive labeled data, otherwise they will not work properly. This phenomenon brings challenges to the development of intelligent diagnosis methods, which can be summarized as the following two aspects.

- 1) In the practical application, most of the data collected from various industrial sites belongs to the health

state of mechanical equipment, the data containing fault information is difficult to capture and record.

- 2) The manual fault data collected in laboratory is difficult to fully simulate the occurrence of real component failures. Because there exists significant characteristic distribution discrepancies between the two data-domains from the laboratory collection and industrial site.

Consequently, small samples problem has greatly hindered the development of artificial intelligent-based diagnosis approaches [11], which gradually becomes a research hotspot. Currently, many researchers consider utilizing transferable knowledge [12-14] to address these problems. For example, Lu et al. [15] used source-domain data and target-domain labeled data to train neural network for fault diagnosis of bearings under varying operating conditions. Wen et al. [16] adopted sparse auto-encoders to extract features from bearings data, then minimized the maximum mean discrepancy to obtain the transferable knowledge. However, the above studies about knowledge transfer still suffer from several weaknesses, such as poor generalization capability, requiring repeated fine-tuning and so on. The proposal of meta-learning further makes up for the shortcomings of the above methods and provides ideas to solve the shortage of samples.

Meta-learning [17-18] as the future promotion direction of machine learning has gradually become the research hotspot in numerous fields. The intention behind meta-learning is to make the network more “intelligent”, able to learn autonomously and adapt to different application scenarios under small sample conditions. Currently, the applications of meta-learning are various, which are roughly divided into three types, that is, metric-based, model-based and optimization-based approach. Among them, metric-based meta-learning network is a fairly easy and effective approach. Metric means to express the correlation between two samples in a certain way. It can be considered that in a certain projection space, the closer the sample is, the more similar it is, that is, it can be classified into the same category. Meta-learning has been studied in many aspects such as image recognition [19] and speech processing [20], but it is seldom studied in the industrial field. For example, Zhang et al. [21] proposed a meta-learning-based matching network (MatchNet) for fault diagnosis of motor bearings. Combining

the specially designed selective sample reuse strategy, the MatchNet could effectively reduce the data distribution discrepancies under various working conditions by generating pseudo-labels of unlabeled samples, and then the model reused these data to iteratively update parameters for realizing faulty bearing classification. Feng et al. [22] developed a domain-adversarial similarity-based meta-learning network (DASMN) for realizing cross-domain fault diagnosis. The DASMN was conducted by minimizing and maximizing the domain-discriminative errors to acquire optimal domain adaptation. Combining with semi-supervised training strategy, the DASMN could successfully detect various bearing failures under small samples condition. Chen et al. [23] proposed a kind of squeeze-and-excitation meta-learning network (SEMN) for bearing vibration data analysis. The SEMN could extract representative prototype features and then utilize unlabeled samples to refine prototype features, which successfully solved the problem of bearing fault diagnosis under insufficient fault samples condition.

At present, most of researches on meta-learning are limited in field of computer vision, its application in field of industrial machinery is still in infancy. In addition, the above mentioned metric-based networks usually use well-defined Euclidean Distance or Cosine Distance, which results in its poor performance in complex recognition tasks. This is because industrial data is often presented as a highly nonlinear complex mapping relation. In response to this actuality, we proposed a novel metric-based meta-learning network named affiliation network (AN) to solve insufficient sample difficulties for fault diagnosis of shipboard antenna, which consisted of four substructures, including a random sampler, a feature extractor, an auxiliary classifier with attention mechanism, and a discriminator. The random sampler samples and partitions the sufficient training data into multiple subsets, which provides data preparation for the prior training implementation of the network. Afterwards, the feature extractor is utilized to extract universal features from partitioned subsets in turn. Finally, the discriminator evaluates the affiliation degree between different labelled data from the feature information filtered by the auxiliary classifier to implement the fault identification task.

The contributions of this paper are summarized as follows.

1) Firstly, we proposed a meta-learning-based approach suitable for fault analysis with insufficient data. No predefined features and fine-tuning training are required. The proposed network adaptively trains a pseudo-distance to evaluate the similarity relationship between few known samples and unknown samples for realizing fault diagnosis. Moreover, specially designed auxiliary classifier combines attention mechanism to help the AN filter useless information, which can greatly improve the identification efficiency of the AN.

2) Secondly, a general prior training strategy was designed in this paper. The portioning of prior training data is used to simulate the small sample conditions, which helps the AN gain transferable knowledge to improve its generalization. During the prior training process, multiple iterative training of simulated testing process can effectively reduce the distribution discrepancies between different samples to acquire better pseudo-distance assessment. Finally, the obtained optimal weights of the prior trained AN can be directly delivered for testing process of different applied scenarios.

3) Finally, strong learning and generalization ability of the AN were verified by three cross-domain experiments. The network prior trained by public dataset was used to test three data from different mechanical equipment. Comparing with the state-of-art algorithms, the AN is able to accomplish the state identification task even if training and testing data are not from the same devices, furthermore, it also achieves good classification results even when the devices work at different rotational speeds, which surpasses other small samples intelligent diagnosis methods.

The rest of this paper is organized as follows. The fundamental of similarity metric-based meta-learning is presented in Section 2. Section 3 describes details of each sub-structure of the proposed approach. The brief introduction of the prior training datasets establishment are given in Section 4. Section 5 verifies the effectiveness of proposed method by three different bearing data. In Section 6, attention mechanism and other aspects of the AN are investigated and discussed. Finally, conclusion is drawn in Section 7.

2. Similarity metric-based meta-learning

Meta-learning has attracted extensive attention in recent years because of its outstanding generalization ability. The purpose of meta-learning is to make the neural network learn to learn autonomously [20], which is usually applied in the research of small sample problems. It seeks rapid and precise adaptation from two distinct data domains. Meta-learning models learn the common internal structure of discrepancies but related tasks to accomplish the data-domain adaptation and generalization. The implementations of the Meta-learning are often not limited to a specific network or algorithm, and many scholars have proved that it can be realized in different forms, such as convolution-based network, recurrent-based network or reinforcement-based network [22-23]. The following section will give a brief introduction to the idea of similarity metric-based meter-learning.

Fig. 1. Illustration of the similarity metric-based meta-learning implementation process.

Generally, a meta-task is composed of a meta-model M , support set S and a validation set V , that is, $\Gamma_{meta} = \{M, S, V\}$. $S = \{(x_i, y_i)\}_{i=1}^{n_s}$ and $V = \{(x_j, y_j)\}_{j=1}^{n_v}$ share the same label space $\mathcal{Y}$. Similarity metric-based meta-model adaptively defines a metric function as the classification criterion through measuring the similarity relationship between the embedded features of support sample S_i and validation sample V_j . The implementation process of metric-based meta-learning is shown in the Fig.1. Firstly, the average representation P_C of each category from S is acquired as follows:

$$P_C = \frac{\sum_{\substack{(x_i, y_i): s \\ y_i = C}} f_M(x_i)}{n_s} \quad (1)$$

where f_M is the embedded abstract features extracted by the meta-model. Then the similarity relationship

between support features and validation features is computed by certain distance assessment criterion. Finally, the predicted probability of validation samples can be obtained as follows:

$$p(y = k | x_i \in S, x_j \in V) = \frac{\exp[-F_M(P_k, f_M(x_i) \Theta f_M(x_j))]}{\sum_{i=1, j=1}^C \exp[-F_M(P_j, f_M(x_i) \Theta f_M(x_j))]} \quad (2)$$

where F_M represents the similarity metric matrix, and Θ is the connect function. k is the number of categories.

Besides the meta-model structure, the learning strategy of meta-learning can't be ignored for improving the generalization ability for the meta-model. For adapting various meta-tasks, sufficient inner training iteration is very necessary on the S and V datasets. The specific learning strategy will be explained in following sections.

3. Proposed method

The proposed method will be presented in this section. The overall architecture of the AN is depicted in Fig. 2. Firstly, three network sub-structures are demonstrated in details respectively, including a feature extractor, an auxiliary classifier and a discriminator. After that, corresponding learning strategy is given at the last part.

Fig. 2. Overview of the Affiliation Network.

3.1 Feature extractor based on convolutional framework

To better extract robust and sensitive characteristics, each sample should be pre-processed by same normalization before inputting network. The implementation process is as follows.

$$\begin{cases} \bar{p} = \frac{1}{n} \sum_{i=1}^n p_i \\ s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (p_i - \bar{p})^2} \\ P_i = \frac{p_i - \bar{p}}{s} \end{cases} \quad (3)$$

where n is the number of data points in each sample. p_i represents the value of i^{th} data point. $\bar{p}$ is the mean value of each sample. s is the standard deviation of each sample. P_i is the updated data value.

The feature extractor mainly involves the convolution layers and the pooling layers. The specific structure of the feature extractor are shown in Table 1. It's worth noting that the size of the convolution kernel is small, which can make network pay more attention to details, reduce the computation amount and time consumption of training. The convolution layer can be calculated as follows.

$$Y_j^l = f_{ReLU}^{\phi}(X_j^l) = f_{ReLU}^{\phi} \left(\sum_{i \in M_j} X_i^{l-1} \otimes w_{ij}^l + b_j^l \right) \quad (4)$$

$$f_{ReLU}(X_j^l) = \max\{0, X_j^l\} \quad (5)$$

where Y_j^l denotes an intermediate convolution output. $\otimes$ is the convolution operation. w_{ij}^l denotes a weight matrix connecting the $l-1^{th}$ layer to the l^{th} layer. b_j^l represents the additive bias given to each output. f_{ReLU} is the *ReLU* activation function. In addition, the maximum pooling layer is calculated as follows.

$$Y_j^l = \text{Max}(\beta_j^l \cdot \text{down}(Y_j^{l-1}) + b_j^l) \quad (6)$$

where β and b are multiplicative bias and additive bias, while *down* is a subsampling function to search the maximum value. After that, the characteristics of different training subsets are concatenated with those of validation subsets, the synthetic characteristics are as the next inputs of the AN. The concatenate result can be represented as follows.

$$Z_{ij} = (f_{\phi}(X_i) \Theta f_{\phi}(X_j)) \quad (7)$$

where Z_{ij} is referred to a synthetic characteristic vector. $f(X_i)$ denotes to extracted features of training subset, while $f(X_j)$ denotes extracted features of validation subset. $\otimes$ represents the concatenation operation.

3.2 Auxiliary classifier with attention mechanism

Attention mechanism is applied to help the AN reduce the influence brought by background information in process of recognition. So it can improves the robustness of model while keeping the identification accuracy.

The auxiliary classifier with attention mechanism is composed of the convolution layers, the pooling layers, and the maximum pooling layers. The specific structure of the auxiliary classifier are given in Table 1. The auxiliary classifier can realize two functions, i. e., the channel attention and the spatial attention. The first one can select the major channel, while the other one can select the distinguished area in image space. The mathematic formula can be defined as follows.

$$Z' = M_c(Z) \otimes f_{\sigma_c}(Z) \quad (8)$$

$$Z'' = M_s(Z') \otimes f_{\sigma_s}(Z') \quad (9)$$

where Z is referred to the output from the feature extractor. Channel attention M_c and spatial attention can be specific extended as follows.

$$\begin{aligned} M_c(Z) &= \sigma(MLP(AvgPool(Z)) + MLP(MaxPool(Z))) \\ &= \sigma(W_1(W_0(Z_{avg}^c)) + W_1(W_0(Z_{max}^c))) \end{aligned} \quad (10)$$

$$\begin{aligned} M_s(Z') &= \sigma'(f_{SoftMax}([AvgPool(Z') \otimes MaxPool(Z')])) \\ &= \sigma'(f_{SoftMax}([Z'^s_{avg} \otimes Z'^s_{max}])) \end{aligned} \quad (11)$$

$$f_{Soft} \Rightarrow \begin{cases} O_{Soft} = \sum_{i,j} w_{Soft} Z_{ij} + b_{Soft} \\ P_{Soft} = \frac{e^{o_j^T Z_{ij}}}{\sum_{i=1}^j e^{o_j^T Z_{ij}}} \end{cases} \quad (12)$$

where σ represents the channel attention parameter, which represents the probability of occurrence of strongly periodic features. σ' is the spatial attention parameter, which represents the probability of occurrence of larger pixel value. W denotes the weight matrix. *AvgPool* indicates the average pooling operation. *MaxPool*

indicates the maximum pooling operation. f_{Soft} represents the *SoftMax* activation function, while w_{Soft} and b_{Soft} are its weight parameters and additional biases. p_{Soft} represents the attention parameters of each output feature from the extractor. The weight of the auxiliary classifier can be automatically adjusted and optimized through itself, and the optimized attention parameters are utilized to guide the discriminator to speed up its learning rate. The specific details of attention mechanism will be discussed in Section 6.

3.3 Discriminator based on convolutional framework

The discriminator uses the convolution framework to discriminate the similarity degree between training subsets and validation subsets, and then finds out the sample with the largest similarity degree to training subset in validation subset. So no threshold is required in this process.

The discriminator mainly consists of the convolution layers, the pooling layers and the full-connection layers. Table 1 demonstrates the specific structure of the discriminator. Mathematical model of convolutional layer and pooling layer have been introduced above, as shown in Eq. (4)-(6). In order to ensure that the affiliation score of final output is between $[0-1]$, *Sigmoid* activation function is adopted in last full-connection layer. The full-connection layer can be described as follows.

$$r = f_{Sigm}(w_f \times Z + b_f) \quad (13)$$

$$f_{Sigm}(Z) = \frac{1}{1 + e^{-Z}} \quad (14)$$

where w_f and b_f are the weight matrix and bias. f_{Sigm} is the activation function. Finally, the discriminator can be defined as follows.

$$\begin{cases} r_{i,j} = \text{Max}[g_\phi(C(f_{\sigma_c}(Z_i), f_{\sigma_s}(Z_j)))], & (i, j = 1, 2, \dots, K) \\ \quad = \text{Max}[g_\phi(C(W_{\sigma_c} \otimes W_{\sigma_s} \otimes f_\phi(Z_i), W_{\sigma_c} \otimes W_{\sigma_s} \otimes f_\phi(Z_j)))] \\ g_\phi = f_{Sigm}(f_{ReLU}(W_{c\Theta fc} \times f_\sigma(Z_{ij}) + b_{c\Theta fc})) \\ f_\sigma = f_\phi \times \sigma_c \times \sigma_s \end{cases} \quad (15)$$

where $r_{i,j}$ represents the affiliation score, the maximum value is taken as the final prediction result of sample.

$f_\sigma(Z)$ comes from the auxiliary classifier, and $f_\phi(Z)$ comes from the feature extractor. $C(\cdot, \cdot)$ is the

concatenate operation. W_{σ_c} and W_{σ_s} are the channel and spatial parameters. g_ϕ represents the discriminator, while $W_{c\phi c}$ and $b_{c\phi c}$ are the united weights and biases of convolution layer and full-connection layer of the discriminator. K denotes the total number of categories.

Table 1 Hyper-parameter settings of the AN

Sub-structure	Operation	Kernel size/strides	No. of kernels	Activation	Padding	Output shape
Feature extractor	Input	\	\	\	\	\
	Convolution	3*3/1	64	ReLU	Valid	84×84×64
	Maximum Pooling	2/1	64	\	Valid	42×42×64
	Convolution	3*3/1	64	ReLU	Valid	42×42×64
	Maximum Pooling	2/1	64	\	Valid	21×21×64
	Convolution	3*3/1	64	ReLU	Same	21×21×64
	Convolution	3*3/1	64	ReLU	Same	21×21×64
	Feature layer concatenation of training subset and validation subset					
Auxiliary classifier	Convolution	3*3/1	64	ReLU	Same	42×21×64
	Average Pooling	2/2	64	\	Valid	21×10×64
	Convolution	3*3/1	64	ReLU	Same	21×10×64
	Maximum Pooling	2/2	64	\	Valid	10×5×64
	Convolution	3*3/1	64	ReLU	Same	10×5×64
	Average Pooling	2/2	64	\	Valid	5×2×64
	Convolution	3*3/1	64	ReLU	Same	5×2×64
	Maximum Pooling	2/2	64	\	Valid	2×1×64
Discriminator	Feature layer product of the auxiliary classifier and the discriminator					
	Convolution	3*3/1	128	ReLU	Valid	42×21×128
	Maximum Pooling	4/1	128	\	Valid	10×5×128
	Convolution	3*3/1	64	ReLU	Valid	10×5×64
	Maximum Pooling	4/1	64	\	Valid	2×1×64
	Full-Connection	\	100	ReLU	\	100×1
	Full-Connection	\	1	Sigmoid	\	Categories×1

3.4 Network training and the flow of algorithm

The loss function L of the proposed method is calculated by minimum Mean Square Error (MSE) function.

$$\varphi, \phi, \sigma \leftarrow L = \begin{cases} \arg \min_{\varphi, \phi} \sum_{i=1}^m \sum_{j=1}^n (r_{i,j} - 1)^2 & \text{if } (\hat{Z}_i = \hat{Z}_j) \\ \arg \min_{\varphi, \phi} \sum_{i=1}^m \sum_{j=1}^n r_{i,j}^2 & \text{if } (\hat{Z}_i \neq \hat{Z}_j) \end{cases} \quad (16)$$

where φ, σ, ϕ are to be optimized weight parameters of each sub-network respectively. m is the number of samples in training subset, and n is the number of samples in validation subset. $\hat{Z}_i$ indicates the real label, and $\hat{Z}_j$ indicates the predicted label. The MSE function is adopted to reduce the error between real label and predicted label in training process, the distances between the features of validation subsets and those of training subsets are guaranteed to be the closest under the premise of same label.

In order to make affiliation score regress to an integer, the following operations can be performed. When the probability of the predicted label corresponding to the real label is high, affiliation score is denoted as 1, otherwise, it is denoted as 0. Consequently, the back propagation algorithm can be used to calculate the errors of each layer, thereby accelerate the convergence of the network. Besides, the entire AN is optimized by the *Adam* optimizer.

Fig. 3. The flow chart of the Affiliation Network.

Considering the difficulties of cross-domain identification of network under small samples prerequisite, this paper designs an implementation procedure of affiliation network, which mainly includes two parts, namely prior training part and testing part. The flow chart is as follows in Fig. 3.

Prior training Process:

- Step 1: Transform sufficient normalized raw data into spectral format by the Short-Time Fourier Transform (STFT) as a sufficient prior training dataset.
- Step 2: Establish the architecture of the affiliation network by Table 1.
- Step 3: The random sampler divides the training data acquired from Step 1 into training subsets and validation subsets, which is used to simulate the test process under the small sample conditions.
- Step 4: The feature extractor extracts and concatenates the features of various subsets from Step 3.
- Step 5: The features extracted from Step 4 are used to train and update the network parameters of the auxiliary classifier and discriminator.
- Step 6: The prior training process of Step 5 is iterated many times to meet the preset requirements, and the obtained weights with the highest verification accuracy is taken as the optimal weight for other testing scenarios.

Testing process:

- Step 1: For the small pseudo-training samples differing from the prior training data, transform the insufficient pseudo-training data into spectral format as a small pseudo-training dataset.
- Step 2: The extractor extracts the features of pseudo-training dataset, which are used to concatenate with the features of unknown testing data.
- Step 3: The discriminator utilizes the optimal weight from Step 6 in prior training process to identify the unknown concatenation features from Step 2, the AN actually does not need to be trained in this process.

Step 4: Obtain the final fault identification results under small samples condition.

4. Prior training data setup

4.1 Introduction of the CWRU data

The Case Western Reserve University (CWRU) bearing data is widely utilized to verify the feasibility and effectiveness of various methods in fault diagnosis fields. To guarantee the stringency of the proposed method, the CWRU data are also used as sufficient prior training data for the implementation of proposed method. The test rig with motor, torque transducer and dynamometer is shown in Fig. 4. Drive-end bearing are SKF deep-groove ball bearings: 6205-2RS JEM SKF. Table 2 shows the relevant bearing specifications. The bearing with single-point defect was manufactured and reinstalled into test rig. Vibration data were collected under 48 kHz sampling frequency. The motor rotational speed was 1800rpm. Please visit the CWRU Bearing Data Center website [24] to acquire more details.

Fig. 4. The CWRU test rig.

Table 2 Parameter of 6205-2RS JEM SKF

Inner diameter (mm)	Outer diameter (mm)	Thickness (mm)	Roller diameter (mm)	Pitch diameter (mm)
25	52	15	7.94	39.04

4.2 Construct the prior training sub-dataset

The essence of meta-learning is to enable the network to learn independently. This requires the AN to learn the rules of general versatility in training process, thus it can play a good role in testing process. Therefore, it is necessary to divide prior training datasets to simulate the testing process when training the network, which can make the network learn general knowledge in small samples conditions. The prior training datasets are randomly divided into training subsets and validation subsets. The specific implementation process in this paper is to randomly select small samples from training dataset as training subset, while the remaining samples are appropriately randomly selected as a validation subset. Repeat this process as much as possible to ensure that the entire dataset is traversed. The training subset represents pseudo-training dataset which belongs to the testing process and the quantity remains consistent, while the validation subset represents testing dataset in same way. Besides, in practical application, the pseudo-training dataset does not participate in network training, which is only used for feature extraction.

To train the AN, the CWRU data were utilized for establishing the prior training subsets. For the sake of simplicity, Four health states of bearings, i.e., normal condition, inner raceway fault, roller fault, and outer raceway fault, are simplified as NC, IF, RF and OF. Further, each fault category contained 400 samples while each sample contained 2048 data points. So we obtained a total of 1600 prior training samples. Referred to [25-26], spectrogram features are more applicable to neural network compared with raw vibration features. Consequently, the acquired training data is transformed into the spectrograms by the Short-Time Fourier Transform (STFT), which is easier to highlight the failure information. The transformation process of STFT can be calculated as follows.

$$STFT_i(x_i, m) = \sum_{n=0}^{N-1} x_i(n) \omega(m-n) e^{-j\omega n} \quad (17)$$

where x_i is the i^{th} time-domain signal. $\omega(\cdot)$ is the window function, Hamming window is adopted in this paper. $STFT_i(x_i, m)$ is the spectrogram of x_i . Each spectrogram was cropped to the same size (84×84 in this paper). Further, we regarded 5 samples as a small sample condition, because it only accounts for 0.3% of entire

training dataset. So every 5 samples formed a training subset, while every 10 samples formed a validation subset. The purpose is to speed up the training of network, and the number of validation subsets can be manually increased. Training subsets and validation subsets could be sampled repeatedly to enable network to fully learn versatile knowledge. The process is done by the random sampler.

$$S_{Training\ subset_i} = \{(X_1, \hat{Z}_1), (X_2, \hat{Z}_2), \dots, (X_k, \hat{Z}_k)\} \quad (18)$$

$$S_{Validation\ subset_i} = \{(X_1, \hat{Z}_1), (X_2, \hat{Z}_2), \dots, (X_K, \hat{Z}_K)\} \quad (19)$$

where X is the sample vector, $\hat{Z}$ is the corresponding data label, which contains four categories of failures and is defined as $[0, 1, 2, 3]$. i is the number of subsets, $i = 1, 2, \dots, N$. k represents the number of extremely small samples of training subset (k is set as 5 in case study), while K is the number of samples of validation subset (K is set as 10 in case study).

5. Case study

5.1 Methods for comparison

In this section, three intelligent diagnosis methods are designed for comparative analysis under small samples prerequisite. Firstly, convolutional neural network (CNN) [10] is widely used as a representative method for state recognition of mechanical equipment due to its excellent feature extraction ability. Secondly, in order to highlight the superiority of the proposed method, a metric-based meta-learning network named Prototypical Network (ProtoNet) [27] is used to compare with the proposed method. The main difference from the proposed method is that the ProtoNet realizes the measurement by calculating the Cosine Distance, while the proposed method realizes the measurement by using the neural network. Finally, an improved meta-learning network based on cosine metric is utilized as a contrast, which is named Matching Network (MatchNet) [21]. More details of the three intelligent algorithms can be found in [10], [21], [27]. During case studies, in order to make experiments comparable, we set the number of layers and kernels of comparison methods to be consistent with the proposed

method in Table 1. Meanwhile, the experimental settings of comparison methods are same as the proposed method.

The experimental platform of relevant algorithms is configured as the 64-bit operating system of Windows8, and the CPU belongs to Intel's Core i3-4150 @3.50GHz. The programming language is Python 3.8.2 and the program running environment is TensorFlow 2.8.0.

5.2 Case 1: SQ data from rotor system

Fig. 5. Graph of the SQ rotor system fault simulation experimental platform and photographs of the manufactured bearings. Left: Inner raceway defect. Right: Outer raceway defect.

To verify the effectiveness of the AN, the investigation was conducted on the rotor system fault simulation experimental platform developed by Spectral Quest (SQ) Company. Fig. 4 demonstrates the specific setup of the experimental platform, which consists of a rotor system, a generator, a magnetic brake and some accelerometers. Generator speed can be set at three rotational speeds (approximately 600, 1200 and 1800rpm) by the speed controller, while the rotational speeds of each experiment were recorded by the tachometer. Two accelerometers were installed at the generator output end for data acquisition. The vibration signals were recorded by the

CoCo80 acquisition instrument with 25.6 kHz sampling frequency. Besides, the detective bearing was also installed into the generator output end. The bearing model is NSK 6203 bearing, and its parameters are shown in Table 3. Two bearings were manufactured with single-point defects, as shown in Fig. 5. One belonged to inner raceway fault (IF) and another one was outer raceway fault (OF). A healthy bearing was used as normal condition (NC).

Table 3 Parameter of NSK 6203

Inner diameter (mm)	Outer diameter (mm)	Thickness (mm)	Roller diameter (mm)	Pitch diameter (mm)
17	40	12	6.747	28.50

Since the operation of AN network is carried out through the similarity metric-based meta-learning training method, whether there exists a certain similarity relationship between the data becomes an important factor to set the following experiment. For this purpose, in order to verify the feasibility of the proposed AN, we first carried out experiments under the condition that the training data and testing data had similar rotational speeds. Further, we reduced the similarity relationship between the training data and testing data, that is, we added the data with various rotational speeds into the testing dataset, and then evaluated the performance and its similarity metric-based principle of the proposed method.

Considering that the prior training datasets from CWRU were collected under 1800rpm, the SQ data at same rotational speed were firstly validated for experiments under small sample conditions. Similarly, each SQ sample contained 2048 data points. And each fault category included 155 spectrogram samples. Finally, we acquired a total of 465 samples with three labels (NC, IF, and OF).

Five samples of each fault category were used as pseudo-training dataset and the rest was for testing. Therefore, the pseudo-training dataset contained 15 samples, while the testing dataset contained 450 samples. It's worth noting that the pseudo-training samples only account for 3.2% of total data volume, such experimental setup is consistent with the requirement of small sample Prerequisite. The learning rate was set as 0.001. After

initializing, the AN was trained though 200 loops per iteration by the established CWRU prior training datasets, the network was continuously trained for 10 iterations, and then the optimal weights with the highest validation accuracy were taken for testing SQ datasets. The final identification results of the AN and each comparative intelligent algorithms are given in Table 4. It can be found that the proposed method gains 100% identification accuracy, which is superior to the comparative state-of-art methods.

Fig. 6. Waveforms and spectrograms of samples created from the SQ test rig. From left to right are different classes (NC, IF, and OF). From top to bottom are different rotational speeds (approximately 600, 1200, and 1800rpm).

Table 4 Classification results of SQ Datasets

Method	Mean accuracy (%)		Modeling time	Testing time
	Single rotational speed	Hybrid rotational speed	(s/loop)	(s/step)
CNN [10]	91.1	82.0	4.3	0.06
ProtoNet [27]	81.3	64.2	13.4	0.13
MatchNet [21]	93.0	91.0	25.9	0.18
AN	100.0	95.1	31.8	0.21

Further, the hybrid rotational speed experiments were also carried out. Some figures of data acquired at multiple speeds are shown in Fig. 6. The same implementation way of dataset establishment as above was adopted, we constructed 155 samples from each fault category at three rotational speeds and obtained their spectrograms though Eq. (17). So we ended up with a total of 1395 samples. 5 samples of each fault category at

each rotational speed were taken as pseudo-training dataset, and the rest were as testing dataset. Consequently, the pseudo-training dataset contained 45 samples, while the testing dataset contained 1350 samples. The network still conducted 200 loops per iteration for training by established CWRU prior training datasets, and training iteration with a total of 10 times were performed to obtain the optimal weight for testing the hybrid SQ data. The final results of various intelligent models are also given in Table 4.

According to analyzing the recognition results of the CNN, ProtoNet and MatchNet in single & hybrid rotational speed experiments, the ProtoNet obtains the lowest accuracy of 64.2%. It indicates that the general meta-learning network is difficult to evaluate the similarity relationship between unknown samples by simply calculating the Cosine Distance. Therefore, the ProtoNet is easy to misjudge fault category due to the interference of background noise and rotational speeds. However, the AN is able to learn a kind of adaptive metric distance criterion without explicit mathematical expression for similarity relationship evaluation between unknown data, so as to greatly improve the recognition accuracy. Besides, comparing the more advanced MatchNet with the AN, the AN still obtains the highest accuracy of 95.1%, which is slightly ahead. But it's worth mentioning that the AN does not need to be retrained like the MatchNet to adapt to the new classification task, which is the superiority of the proposed approach. To sum up, the feasibility and effectiveness of the proposed AN are verified by cross-domain experiments under small samples prerequisite.

5.3 Case 2: Gearbox reducer data from Shipboard antenna system

A classification task under extreme conditions was carried out to further test the performance of proposed method. As shown in Fig. 7, the experiment setup consists of 110SU5502 type generator and PLS90-32 type planetary gearbox reducer to precisely simulate the complicated mechanical transmission system of shipboard antenna. The gearbox output shaft bearing type is 32007. Two faulty Bearings with the RF and OF labels were used for experiments, which were installed into to output end of the gearbox. Two accelerometers were attached to the gearbox output end in the horizontal and vertical directions for the acquisition of vibration signals. The

signals were sampling at 5 kHz by the CoCo80 data-acquisition instrument and the generator rotational speed could be set as 600 rpm, 1200 rpm and 1500 rpm.

Fig. 7. Experiment setup of the shipboard antenna gearbox reducer test rig and photographs of the manufactured bearings. Left: Roller defect. Right: Outer raceway defect.

Fig. 8. Waveforms and spectrograms of samples created from the gearbox reducer test rig. From left to right are different classes (NC, RF, and OF). From top to bottom are different rotational speeds (approximately 600, 1200, and 1500rpm).

As mentioned in Case 1 above, this case would also include single & hybrid rotational speed experiments.

Consistent with the Case 1, each sample contained 2048 data points, and we formed 205 samples with each fault

category at same rotational speed. So a total of 1845 samples were obtained. Thereinto, the pseudo-training dataset contained 45 samples, while the testing dataset contained 1800 samples. Fig. 8 shows some waveforms and spectrograms of constructed datasets. It can be found that the samples with different labels are more difficult to separate by observation. Even the samples with same label can vary a lot under different rotational speeds, which implies a more difficult classification task than the Case 1.

Similarly, single rotational speed experiment was carried out. The experimental setup was same as the Case 1, 5 samples of each fault category at 1500 rpm were used as small pseudo-training dataset and the rest were used for testing. The learning rate of optimizer was set as 0.001. Afterwards, the AN was still trained 200 loops per iteration for a total of 10 iterations by the established CWRU prior training datasets, which was used for obtaining the optimal weights. Table 5 presents the recognition results of the AN and other comparative intelligent methods. It can be found that the AN obtained the highest 98.2% accuracy, which is slightly lower than Case 1. This is due to the discrepancies between rotational speeds of prior training data and pseudo-training data, (the prior training data acquired at 1800rpm and the pseudo-training data acquired at 1500rpm). The specific reason will be discussed in Section 6. However, the testing results of proposed method still far exceed the comparison methods, which reflects its outstanding performance.

Table 5 Classification results of Gearbox Reducer Datasets

Method	Mean accuracy (%)		Modeling time (s/loop)	Testing time (s/batch)
	Single rotational speed	Hybrid rotational speed		
CNN [10]	69.5	59.2	5.6	0.07
ProtoNet [27]	73.3	62.6	17.5	0.16
MatchNet [21]	86.0	63.0	33.7	0.20
AN	98.2	92.0	40.3	0.25

Furthermore, we also used the AN trained by the CWRU prior training dataset to test hybrid data with 600rpm, 1200rpm and 1500rpm on the basis of previous experiments. The implementation process was the same as the

single-speed experiment. The final recognition results of various algorithms are also demonstrated in Table 5. The results indicate that the accuracies of four models drop significantly, which is due to the deterioration of data quality caused by the environmental interference, such as complex transmission paths and mass of background noise. Besides, the serious lack of training data also results in various models not learning enough diversity of knowledge. Nevertheless, the AN is still far ahead of other comparative approaches. In summary, the AN is liable to adapt to new tasks and achieves better results than other methods, which indicates that the AN can effectively learn from a small amount of data and make itself more generic.

5.4 Case 3: Engineering verification

Fig. 9. Graph of the real shipboard antenna gearbox reducer.

Fig. 10. Waveforms and spectrograms of samples created from the measured data: (a) waveform with NC label; (b) waveform with RF label; (c) spectrogram with NC label; (d) spectrogram with RF label.

The feasibility of the proposed AN has been adequately verified. Further, its practical effect in engineering

application will be discussed in following section. The measured data were derived from a gearbox reducer installed in the shipboard antenna of astronautic measurement ship. Fig. 9 shows the graph of the real shipboard antenna gearbox reducer. Its specific structure and the sensors arrangement were consistent with Case 2. The type of gearbox output shaft bearing is also 32007. The data sampling frequency was set at 2048Hz. Meanwhile, the speed of the motor fluctuated between 1350rpm and 1500rpm, which greatly increased the difficulty of fault identification. Through long-term data acquisition, a sudden failure was detected on the bearing roller at the output end of the gearbox. Some waveforms and spectrograms with the NC and RF label are given in Fig. 10.

The dataset contained 205 samples was established from the data with NC label, while the dataset with RF label was conducted in same way. Finally, a total of 410 samples were obtained, thereinto, 10 samples were used as pseudo-training data, while the rest was used as testing data. The implementation of this experiment was consistent with the Case 2. Fig. 11 demonstrates the confusion matrix of identification result of the AN. It can be found that the AN can achieve 95.8% recognition accuracy. And the AN can identify the vast majority of health samples, it indicates that the AN has the ability to distinguish abnormal samples which is also necessary for mechanical equipment maintenance.

Fig. 11. The confusion matrix of recognition result of the AN on the measured data.

To further verify the superiority of proposed AN in practical application, three mentioned intelligent methods were used to identify the measured data as comparative experiments. Table 6 gives the recognition results. It can

be found that the recognition accuracy of proposed method is much higher than other three methods. Although the testing process of the AN is more time-consuming than other models, it's still acceptable to take into account the recognition accuracy. Excellent generalization performance of the AN once again proves its engineering practical value, which provides a new idea to solve insufficient fault sample challenge.

Table 6 Classification results of measured data

Method	Mean accuracy (%)	Testing time (s/step)
CNN [10]	65.5	0.02
ProtoNet [27]	71.4	0.07
MatchNet [21]	75.0	0.11
AN	95.8	0.13

6. Discussion

6.1 Attention mechanism influence

Attention mechanism is widely used in machine vision to reduce interference of background information [28-29], which happens to coincide with the demand of fault diagnosis. In this paper, attention mechanism is adopted to make network pay more attention to the prominent information in spectrograms, and the specific mathematical model has been given in Section 2. These auxiliary parameters are used to improve network's learning of key features. Fig. 12 shows the region where the key features are learned by the auxiliary classifier, the yellow stripe in red box contains characteristic frequency and rotational frequency information of faulty bearings in Case 1, while the remaining blue part is the background noise. The goal of the AN is to find the potential link between training data and testing data to achieve cross-domain classification tasks. Consequently, the spatial attention enables network focus on the width and relative position of yellow stripe in spectrograms. Since the raw vibration data is converted into spectrogram, the dimension of data is expanded from one to three, which enriches data diversity and highlights the fault information in a sense. Therefore, the channel attention helps network focus more on the bright patches in spectrograms, which represents the amplitude mutation of

mechanical data.

Fig. 12. The capture parts of attention mechanism.

In order to quantitatively describe the improvement brought by attention mechanism, this paper conducted a comparative experiment on two cases before and after applying attention mechanism, and the comparison results are presented in Table 7. The recognition accuracy of Case 1 are improved by 0.2%, 1.3%, while the recognition accuracy of Case 2 are improved by 8.7%, 12.5%. It can be found that the attention mechanism greatly improves the recognition accuracy of Case 2, because the data of Case 2 is more complex, and the network successfully eliminates interference relying on the attention mechanism. In addition, the training time of each loops is reduced by approximately 10% after adding attention mechanism, which enhances the engineering practicability of proposed method.

Table 7 Comparison results before and after applying attention mechanism

Testing dataset	Mean accuracy (%): Case 1		Mean accuracy (%): Case 2	
	Before	After	Before	After
Single rotational speed	99.8	100.0	90.3	98.2
Hybrid rotational speed	93.9	95.1	81.8	92.0

6.2 Rotational speed relationship between prior training data and pseudo-training data

Rotational speed interference can't be ignored during the process of fault diagnosis, which is often one of the most important factors affecting recognition of intelligent methods. The classification ability of AN has been verified by three cases in Section 5. It can be found from Table 3 and Table 4 that the identification accuracy of

hybrid-speed experiment is significantly lower than that of single-speed experiment. Thence, based on the Case 1 and Case 2, this section carried out single-speed experiments with various speeds to discuss its influence on meta-learning network. The prior training data was still collected from CWRU dataset at 1800rpm, and the pseudo-training data was acquired from the SQ and Gearbox Datasets at various rotational speeds. Table 8 demonstrates the relevant single-speed experiment settings.

Table 8 Single-speed experiment Settings

	Prior training data	Pseudo-training data		Prior training data	Pseudo-training data
Case 1	CWRU data at 1800rpm	SQ data at 600rpm	Case 2	CWRU data at 1800rpm	Gearbox data at 600rpm
	CWRU data at 1800rpm	SQ data at 1200rpm		CWRU data at 1800rpm	Gearbox data at 1200rpm
	CWRU data at 1800rpm	SQ data at 1800rpm		CWRU data at 1800rpm	Gearbox data at 1500rpm

Fig. 13. The relationship between classification accuracy and rotational speed.

The final identification results of various single-speed experiments are shown in Fig. 13. It can be found that the closer the speed of pseudo-training data is to prior training data, the better the network classification effect is. The cause of this phenomenon is that, under same sampling frequency, the number of fault impulses contained in

a rotation period at different speeds is different, which results in obvious changes of fault information in the frequency domain. Besides, many bearing-based dynamics simulation researches [30] show that the failure frequency of different components is closely related to the rotational speed in frequency domain, as shown in Eq. (20). The passing rate f_i , f_o and the rotating frequency f_r of rollers modulate the natural frequency of raceway, and the rotational speed directly affects the motion state of rollers. These factors result in significant changes in width and position of bright stripes in spectrograms, which affects the identification results of AN. In conclusion, since the AN trains a pseudo-distance to discriminate the similarity between training data and testing data, the above phenomenon strongly supports the rationality of proposed method.

$$\begin{cases} f_i = 0.5z(1 + \frac{d}{E}\cos\alpha) & (\text{Inner raceway failure}) \\ f_o = 0.5z(1 - \frac{d}{E}\cos\alpha) & (\text{Outer raceway failure}) \\ f_r = 0.5\frac{E}{d}f(1 - (\frac{d}{E})^2\cos^2\alpha) & (\text{Roller failure}) \end{cases} \quad (20)$$

where z is the number of rollers. d is the diameter of roller, E is the pitch diameter of raceway and α is the contact angle. f represents the rotating frequency.

6.3 Advantages and drawbacks

Feasibility and effectiveness of the proposed AN was verified by three heterogeneous data. The advantages of the proposed AN can be summarized into three aspects.

Firstly, during the prior training process, the sufficient prior training data is randomly divided into multiple training subsets and validations subsets to simulate the practical testing process, which effectively enhances the domain adaptation capability of the AN. The pseudo-distance assessment acquired by multi-loops prior training can effectively reduce the distribution discrepancies between samples from varying domains and operating conditions in embedding space.

Secondly, the auxiliary classifier with attention mechanism is designed to help the AN capture the effective fault information and remove the useless background noise. The relevant experiments indicate that the attention

mechanism indeed reduce the interference of random noise and make the effective iterative update of the AN for higher identification accuracy and faster convergence speed.

Thirdly, a general prior training strategy is specially designed to gain high-generalization diagnosis model. The specific implementation process has introduced in Section 2. Two cross-domain experiments show that, without secondary training or fine-tuning, the trained network can be directly used for bearing fault identification between different equipment under small sample conditions.

There are still some shortcomings in the proposed method. First, the rotational speed discrepancies between prior training samples and pseudo-training samples have an impact on recognition accuracy of the AN, which is due to differences of the spectrograms under varying rotational speeds. Supplement of various rotational speeds prior training data may ameliorate this problem. Besides, from Table 4 and Table 5, the prior training process is comparatively time-consuming, because the AN needs to acquire better pseudo-distance evaluation to ensure its high-generalization performance. Considering that offline training and online testing of models are usually adopted in most industrial scenarios, the time-consumption of the AN is also acceptable. In addition, Pruning strategy [31] may be an effective way to lighten the weight of the AN, which can remove the useless channels of the network and speed up the training of the network.

7. Conclusions

This paper proposes a novel meta-learning network called affiliation network for fault classification under small samples condition. The main conclusions can be summarized as follows.

(1) This method utilizes sufficient laboratory data for realizing network prior training, which is used to simulate the process of testing and adaptively captures the universal knowledge for failure mode identification. The proposed AN trains a pseudo-distance to evaluate the affiliation degree between different samples with various fault categories, then the prior trained AN is able to directly applied for other scenarios without additional adjustments.

(2) The application of attention mechanism effectively helps the AN to improve the recognition accuracy by 1%-10%. Three cross-domain experiments verified its excellent performance and generalization capability. And the rotational speed interference and other factors affecting the performance of proposed method are further discussed and summarized.

(3) Our work indicates that the proposed method has competitive capacity for fault classification in the face of small samples. Meanwhile, comparing with conventional intelligent algorithms, the AN is not required secondary retraining for different application scenarios and can achieve higher identification accuracy.

Future improvements include enhancing network performance in face of extreme conditions such as mass of noise, and the possibility of training network under zero fault sample prerequisite.

Acknowledgements

The authors sincerely thank to professor Yanyang Zi and associate professor Yu Wang of Xi'an Jiaotong University for their critical discussion and reading during manuscript preparation.

Authors' contributions

YC was in charge of the whole trial and wrote the manuscript; JC proposed useful suggestions for the analysis; JP, KZ, TP made some correction of the manuscript. All authors read and approved the final manuscript.

Authors' Information

Yuanhong Chang is currently pursuing the Ph.D. degree at College of Mechanical Engineering, Xi'an Jiaotong University, China.

Jinglong Chen is currently an associate professor at College of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China. He received the Ph.D. degree in Mechanical Engineering from Xi'an Jiaotong University, China, in 2014.

Jun Pan received the Ph.D. degree in Mechanical Engineering from Xi'an Jiaotong University, China, in 2019.

Kaiyu Zhang is currently pursuing the Ph.D. degree at College of Mechanical Engineering, Xi'an Jiaotong University, China.

Tongyang Pan is currently pursuing the Ph.D. degree at College of Mechanical Engineering, Xi'an Jiaotong University, China.

Funding

Supported fiscally by the National Key Research and Development Program of China (No.2019YFF0302204), Major Research Program of National Natural Science Foundation of China (No. 91960106), National Natural Science Foundation of China (No. U1933101, No. 51875436), China Postdoctoral Science Foundation (No. 2018M631145) and Natural Science Foundation of Shaanxi (No. 2019JM-041).

Competing Interests

The authors declare no competing financial interests.

Author Details

¹ State Key Laboratory for Manufacturing and Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, PR China.

References

- [1] J. Xu, C. Li, X. Miao, C. Zhang, X. Yuan. An overview of bearing candidates for the next generation of reusable liquid rocket turbopumps. *Chinese Journal of Mechanical Engineering*, 2020, 33: 26.
- [2] Y. Jin, J. Lu, W. Ouyang, Z. Liu, K. Lao. Vibration reduction performance of damping-enhanced water-lubricated bearing using fluid-saturated perforated slabs. *Chinese Journal of Mechanical*

Engineering, 2020, 33: 92.

- [3] Z. Shang, W. Li, M. Gao, X. Liu, Y. Yu. An intelligent fault diagnosis method of multi-Scale deep feature fusion based on information entropy. *Chinese Journal of Mechanical Engineering*, 2021 34: 58.
- [4] H. Ahmed, A.K. Nandi. Connected components-based colour Image representations of vibrations for a two-stage fault diagnosis of roller bearings using convolutional neural networks. *Chinese Journal of Mechanical Engineering*, 2021, 34: 37.
- [5] Y. Lv, Q. Zhou, Y. Li, W. Li. A predictive maintenance system for multi-granularity faults based on AdaBelief-BP neural network and fuzzy decision making. *Advanced Engineering Informatics*, 2021, 49: 101318.
- [6] X. Wang, L. Luo, L. Tang, Z. Yang. Automatic representation and detection of fault bearings in in-wheel motors under variable load conditions. *Advanced Engineering Informatics*, 2021, 49: 101321.
- [7] D. You, X. Gao, and S. Katayama. WPD-PCA-based laser welding process monitoring and defects diagnosis by using FNN and SVM. *IEEE Transactions Industrial Electronics*, 2015, 62(1): 628–638.
- [8] Y. Cheng, K. Hu, J. Wu, H. Zhu, X. Shao. A convolutional neural network based degradation indicator construction and health prognosis using bidirectional long short-term memory network for rolling bearings. *Advanced Engineering Informatics*, 2021, 48: 101247.
- [9] Y. Chang, J. Chen, et al. Intelligent fault diagnosis of wind turbines via a deep learning network using parallel convolution layers with multi-scale kernels. *Renewable Energy*, 2020, 15: 205-213.
- [10] J. Chen, Y. Chang, et al. Intelligent Impulse Finder: A boosting multi-kernel learning network using raw data for mechanical fault identification in big data era. *ISA Transactions*, 2020, 107: 402-414.
- [11] S. Shao, P. Wang, R. Yan. Generative adversarial networks for data augmentation in machine fault diagnosis. *Computers in Industry*, 2019, 106: 85-93.
- [12] S.H. Bang, R. AK, A. Narayanan, Y.T. Lee, H. Cho. A survey on knowledge transfer for manufacturing

data analytics. *Computers in Industry*, 2019, 104: 116-130.

- [13] I.F. Amo, J.A. Erkoyuncu, R. Roy, R. Palmarini, D. Onoufriou. A systematic review of Augmented Reality content-related techniques for knowledge transfer in maintenance applications. *Computers in Industry*, 2018, 103: 47-71.
- [14] W. Chen, Y. Qiu, Y. Feng, Y. Li, A. Kusiak. Diagnosis of wind turbine faults with transfer learning algorithms. *Renewable Energy*, 2021, 163: 2053-2067.
- [15] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, T. Zhang. Deep model based domain adaptation for fault diagnosis. *IEEE Transactions Industrial Electronics*, 2017, 64: 2296–2305.
- [16] L. Wen, L. Gao, X. Li. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Transactions on Systems Man and Cybernetics: Systems*, 2019, 49: 136– 144.
- [17] Z. Xu, L. Cao, X. Chen. Learning to learn: Hierarchical meta-critic networks. *IEEE Access*, 2019, 7: 57069-57077.
- [18] D. Lechevalier, A. Narayanan, S. Rachuri, S. Foufou. A methodology for the semi-automatic generation of analytical models in manufacturing. *Computers in Industry*, 2018, 95: 54-67.
- [19] Vinyals, Oriol, Blundell *et al.* Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 2016: 3637-3645.
- [20] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks. In: *International Conference on Learning Representations (ICLR 2017)*, Toulon, France, 2017.
<https://arxiv.org/abs/1703.03400v3>.
- [21] K. Zhang, J. Chen, T. Zhang, et al. Intelligent fault diagnosis of mechanical equipment under varying working condition via iterative matching network augmented with selective Signal reuse strategy. *Journal of Manufacturing Systems*, 2020, 57: 400-415.

- [22] Y. Feng, J. Chen, Z. Yang, et al. Similarity-based meta-learning network with adversarial domain adaptation for cross-domain fault identification. *Knowledge-Based Systems*, 2021, 217: 106829.
- [23] Y. Feng, J. Chen, T. Zhang, et al. Semi-supervised meta-learning networks with squeeze-and-excitation attention for few-shot fault diagnosis. *ISA Transactions*, 2021. <https://doi.org/10.1016/j.isatra.2021.03.013>
- [24] Case Western Reserve University Bearing Data Center Website
<http://csegroups.case.edu/bearingdatacenter/home>.
- [25] L. Deng, J. Li, J.-T. Huang, K. Yao, *et al.* Recent advances in deep learning for speech research at Microsoft. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013. <https://ieeexplore.ieee.org/document/6639345>.
- [26] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton. Binary coding of speech spectrograms using a deep auto-encoder. In *Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, September. DBLP, 2011.
https://isca-speech.org/archive/interspeech_2010.
- [27] J. Snell, K. Swersky, R. Zemel. Prototypical networks for few-shot learning. 2017.
<https://arxiv.org/abs/1703.05175>.
- [28] R. Wei, P. Love, W. Fang, H. Luo, S. Xu. Recognizing people’s identity in construction sites with computer vision: A spatial and temporal attention pooling network. *Advanced Engineering Informatics*, 2019, 42: 100981.
- [29] Y. Chen, G. Peng, Z. Zhu, S. Li. A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. *Applied Soft Computing*, 2020, 86: 105919.
- [30] P. Novotny, J. Hrabovsky, et al. Effective thrust bearing model for simulations of transient rotor dynamics. *International Journal of Mechanical Sciences*, 2019, 157: 374-383.

- [31] Q. Tian, T. Arbel, J. Clark. Task dependent deep LDA pruning of neural networks. *Computer Vision and Image Understanding*, 2021, 203: 103154.