ISA Transactions 119 (2022) 152–171

Contents lists available at ScienceDirect

ISA Transactions

journal homepage: www.elsevier.com/locate/isatrans

Research article

Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions

Tianci Zhang [a], Jinglong Chen [a,*], Fudong Li [a], Kaiyu Zhang [a], Haixin Lv [a], Shuilong He [b,*], Enyong Xu [c,d]

[a] State Key Laboratory for Manufacturing and Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, PR China
[b] School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology, Guilin 541004, China
[c] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, 430074, China
[d] Dongfeng Liuzhou Motor Co., Ltd., Liuzhou 545005, China

ARTICLE INFO

Article history:
Received 22 October 2020
Received in revised form 24 February 2021
Accepted 24 February 2021
Available online 8 March 2021

Keywords:
Intelligent fault diagnosis
Small & imbalanced data
Data augmentation
Feature learning
Classifier design
Meta-learning
Zero-shot learning

ABSTRACT

The research on intelligent fault diagnosis has yielded remarkable achievements based on artificial intelligence-related technologies. In engineering scenarios, machines usually work in a normal condition, which means limited fault data can be collected. Intelligent fault diagnosis with small & imbalanced data (S&I-IFD), which refers to build intelligent diagnosis models using limited machine faulty samples to achieve accurate fault identification, has been attracting the attention of researchers. Nowadays, the research on S&I-IFD has achieved fruitful results, but a review of the latest achievements is still lacking, and the future research directions are not clear enough. To address this, we review the research results on S&I-IFD and provides some future perspectives in this paper. The existing research results are divided into three categories: the data augmentation-based, the feature learning-based, and the classifier design-based. Data augmentation-based strategy improves the performance of diagnosis models by augmenting training data. Feature learning-based strategy identifies faults accurately by extracting features from small & imbalanced data. Classifier design-based strategy achieves high diagnosis accuracy by constructing classifiers suitable for small & imbalanced data. Finally, this paper points out the research challenges faced by S&I-IFD and provides some directions that may bring breakthroughs, including meta-learning and zero-shot learning.

1. Introduction

Fault diagnosis plays an essential link in machine health management as it builds a bridge between machine monitoring data and its health conditions. Intelligent fault diagnosis utilizes artificial intelligence technologies in the process of fault diagnosis to make it intelligent and automatic [1]. Recently, deep neural network such as deep auto-encoder (DAE) [2,3], deep convolutional neural network (DCNN) [4,5], and other deep networks [6,7], have been widely used to build end-to-end intelligent diagnosis models, which reduces the dependence on manual labor and expert knowledge, and greatly promotes the development of intelligent fault diagnosis [8].

Intelligent fault diagnosis with small & imbalanced data (S&I-IFD) refers to build intelligent diagnosis models using a few machine faulty samples to achieve accurate fault identification. Generally speaking, intelligent diagnosis models with deep networks are built on sufficient machine monitoring data analysis [8]. The more sufficient the training data is, the more abundant the fault types in the training set are, the higher the diagnosis accuracies of intelligent diagnosis models are. However, in engineering scenarios, it is difficult to build an ideal dataset for the training of intelligent diagnosis models for the following three reasons.

(1) In engineering scenarios, machines usually work in a normal condition and faults are rare. Therefore, despite the condition monitoring system composed of multiple sensors can collect data from machines constantly, the majority of the collected data is healthy data, and the volume of the fault data is small. Thus, it is hard to obtain sufficient fault data from engineering scenarios directly to support the training of intelligent diagnosis models.
(2) It is expensive to carry out fault simulation experiments to collect machine fault data in the laboratory. For example, to obtain fault data of gears in the laboratory, researchers need to purchase gear specimens and manufacture faults

* Corresponding authors.
E-mail addresses: jlstrive2008@mail.xjtu.edu.cn (J. Chen), xiaofeilonghe@guet.edu.cn (S. He).

https://doi.org/10.1016/j.isatra.2021.02.042
0019-0578/© 2021 ISA. Published by Elsevier Ltd. All rights reserved.

by wire-electrode cutting or other ways artificially. Moreover, it is necessary to build a fault simulation test bench to collect data. Such an experiment is not only expensive but also consumes a lot of human labor. Besides, some common faults like gear tooth surface bonding are difficult to be simulated by artificial fault manufacturing. Thus, it is difficult to collect fault data by conducting fault simulation experiments in the laboratory.

(3) The fault data obtained by computer simulation is not practical. Some fault simulation software can simulate faults of equipment and output fault data. For example, Gasturb is a performance calculation software of aero-engines [9]. Researchers use Gasturb to simulate faults of aero-engines to obtain fault data. However, although Gasturb can perform precise mathematical operations, it cannot simulate the complex working environment of aero-engines. Different working environments and working conditions have a significant impact on fault data. Therefore, the fault data obtained by simulation is usually not practical enough.

In short, intelligent fault diagnosis in engineering scenarios is a typical small & imbalanced data problem. In this case, if the intelligent diagnosis model is trained with limited fault data directly, it is prone to poor generalization performance and low fault identification accuracy. Therefore, the lack of fault samples makes it difficult to build an effective intelligent diagnosis model and achieve accurate fault identification in engineering scenarios.

How to solve the S&I-IFD problem has been the research interest of scholars for a long time. For example, some researchers use Synthetic Minority Over-sampling Technique [10] to expand faulty sample number or develop fault classifiers with Support Vector Machines [11], so diagnosis models can have relatively high identification accuracies under the condition of insufficient fault data samples. Recently, the research on S&I-IFD has yielded fruitful achievements with new machine learning algorithms. For instance, researchers use generative adversarial networks (GAN) to emulate data distributions of machine faulty samples so that more faulty samples are generated to expand the limited fault dataset [12]. Besides, transfer learning-related diagnosis models reuse the previously learned diagnosis knowledge to the new diagnosis task, so that accurate fault identification can also be achieved using a few faulty samples [13].

At present, there have been many research achievements on S&I-IFD, however, the research directions for future development are not clear enough, and a review for existing results is still lacking. Although some reviews about intelligent fault diagnosis have been published, these reviews mainly aim at the utilization of some theory like deep learning to specific objects like induction motors [14,15], not on the problem of lacking fault data samples. There is no doubt that small & imbalanced data learning is a common problem in many areas of the real world, such as medical, financial, and so on [16]. For example, the detection of invalid transactions and financial fraud in the trading system of banks is also a typical small & imbalanced data problem. Therefore, many reviews on imbalanced data classification have also been published [16–19]. However, these existing reviews pay little attention to the new machine learning theory and algorithms like GAN and transfer learning, which have been widely applied to S&I-IFD in recent years. Moreover, these existing reviews are mainly a summary of research methods and do not take mechanical equipment as a special research object. From the perspective of data analysis, the analysis of machine monitoring data often involves frequency domain analysis, and so on, which is different from other data analysis such as image data analysis. Besides, as far as the authors know, similar review papers for S&I-IFD are neither under consideration nor already published in another venue. Therefore, it is necessary to present a review for S&I-IFD to summarize the existing achievements and give some future directions for further exploration.

This paper provides a review of S&I-IFD. The contributions of this paper include two aspects. First, this paper focuses on the small & imbalanced data problem in intelligent machine fault diagnosis, which is a significant research point, but the related review is still lacking in intelligent fault diagnosis. Taking mechanical equipment as the research object, this paper reviews the related work on S&I-IFD in the past 10 years, and focuses on the latest research results represented by GAN and transfer learning. Different from other reviews on small & imbalanced data learning [16–19], this paper divides the achievements of S&I-IFD into three categories: data augmentation-based strategy, feature learning-based strategy, and classifier design-based strategy, according to the general process of machine fault diagnosis (MFD), as shown in Fig. 1. In particular, MFD contains three main stages: data preprocessing, feature extraction, and conditions classification [1]. For S&I-IFD, solutions can also be found from the three steps, as shown in Fig. 1. From the perspective of data preprocessing, scholars augment the limited fault data through data generation or data oversampling, and the augmented data can be directly used to train intelligent diagnosis models. In terms of feature extraction, fault features can be learned from limited fault data directly by designing regularized neural networks or feature adaptation without data augmentation. On the aspects of conditions classification, the health conditions of machines can be classified directly by designing fault classifiers suitable for small & imbalanced data without data augmentation or the designing of feature extraction models. Compared with the other reviews on small & imbalanced data learning [16–19], the presented review has stronger field characteristics due to the special classification mode of the research achievements. As a result, this paper may be more enlightening for researchers in this field.

Second, based on the existing research results and the latest machine learning theories, this paper provides some research challenges and directions for further development. Specifically, in the aspect of data augmentation, the current researches mainly focused on expanding the number of fault samples, while how to measure and enhance the samples' quality needs to be paid more attention to. How to prevent negative transfer in the diagnosis models is a key to the application in engineering scenarios. Besides, as a new machine learning theory, meta-learning [20] has initially shown its advantages in dealing with small sample problems. Thus, the applications of meta-learning theory on S&I-IFD may increase greatly. Finally, zero-shot learning [21] may bring a breakthrough for S&I-IFD in the extreme case where there are no fault samples available at all.

For the rest of this review, Section 2 describes both the research methodology and the initial data analysis. Section 3, 4, and 5 review the research achievements from the perspective of the data augmentation, the feature learning, and the classifier design respectively. Section 6 gives some possible extensions for S&I-IFD in the future. Section 7 presents a conclusion for this review.

2. Research methodology and initial analysis

2.1. Research methodology

This paper mainly searched and collected the publications on S&I-IFD published from 2010 to November 2020. Four library databases covering the natural science research field were selected for the literature search, which are Science Direct, IEEE Xplore, Springer, and ACM. Besides, the Scopus and the Web of Science were also used to search the papers in some individual publishers [22].

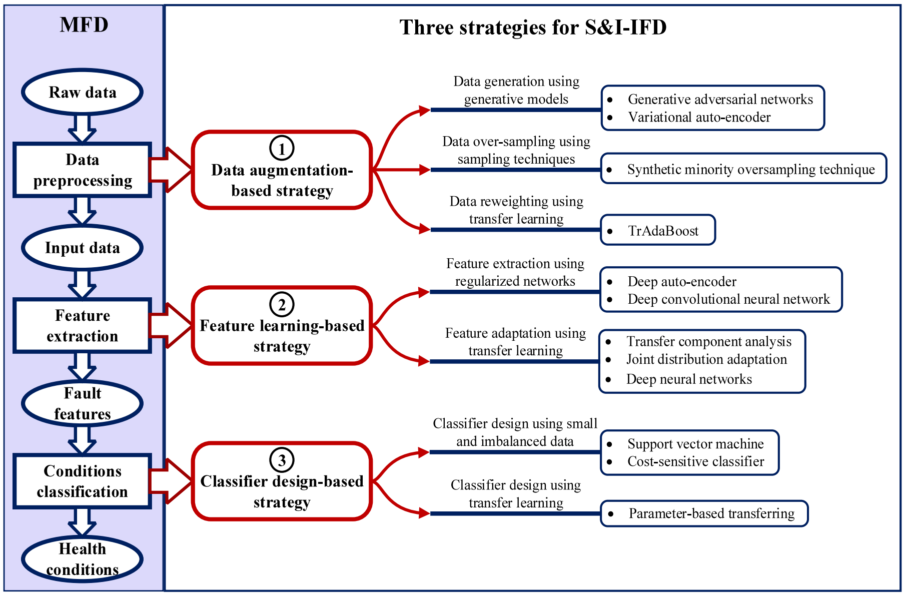

MFD
Raw data
Data preprocessing
Input data
Feature extraction
Fault features
Conditions classification
Health conditions
Three strategies for S&I-IFD
1 Data augmentation-based strategy
2 Feature learning-based strategy
3 Classifier design-based strategy
Data generation using generative models
Data over-sampling using sampling techniques
Data reweighting using transfer learning
Feature extraction using regularized networks
Feature adaptation using transfer learning
Classifier design using small and imbalanced data
Classifier design using transfer learning
• Generative adversarial networks
• Variational auto-encoder
• Synthetic minority oversampling technique
• TrAdaBoost
• Deep auto-encoder
• Deep convolutional neural network
• Transfer component analysis
• Joint distribution adaptation
• Deep neural networks
• Support vector machine
• Cost-sensitive classifier
• Parameter-based transferring

Fig. 1. The process of machine fault diagnosis and the three strategies for S&I-IFD.

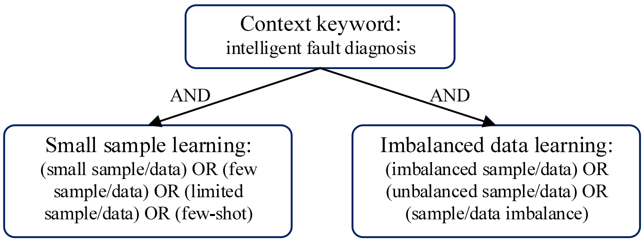

Context keyword:
intelligent fault diagnosis
AND
AND
Small sample learning:
(small sample/data) OR (few sample/data) OR (limited sample/data) OR (few-shot)
Imbalanced data learning:
(imbalanced sample/data) OR (unbalanced sample/data) OR (sample/data imbalance)

Fig. 2. The two-level keywords tree.

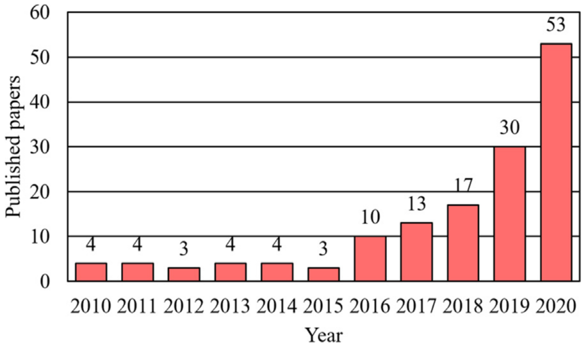


Fig. 3. The publishing trends of S&I-IFD.

Inspired by [16], a two-level keywords tree was constructed to collect published papers on S&I-IFD as comprehensive as possible, as given in Fig. 2. Since intelligent fault diagnosis in the small & imbalanced data case was reviewed in this paper, the search keyword of the first level was restricted to intelligent fault diagnosis. For S&I-IFD, some scholars regard it as the problem of imbalanced data classification [23,24], because the volume of health data is larger than fault data. On the other hand, some scholars regard it as the problem of small sample classification [12,25,26], that is, the volume of health data is set to be the same as the fault data to avoid the problem of data distribution imbalance. Therefore, the search keywords of the second level were divided into two parts, i.e., small sample learning and imbalanced data learning respectively, as shown in Fig. 2. A total of 249 English journal papers were collected in the initial search. After further review, 145 papers were related to the theme of this paper, which will be the main data source of this review. Besides, in the citations of these papers, we found 9 related conference papers and included them in the references for this review.

In the process of literature search, due to the inaccurate or incomplete keywords, there may be a lack of some related literature. For example, some scholars regard the imbalanced data as "skewed data" [16]. However, in the literature search process, we did not list "skewed data" as the search keyword, which is the main limitation and threat to the validity of the literature search.

2.2. Initial analysis

Fig. 3 shows the number of S&I-IFD-related publications in 2010–2020. It can be seen that there are few English journal papers about S&I-IFD from 2010 to 2015, while the number of published papers increased rapidly since 2016, which is mainly due to the emergence and application of new machine learning models like GAN [12]. The trends in Fig. 3 show that S&I-IFD is a valuable research problem and may continue to be a research hotspot in the next few years.

After a careful review, the collected papers are classified into data augmentation-based strategy, feature learning-based strategy, and classifier design-based strategy, as shown in Fig. 1. Inspired by the general process of machine fault diagnosis, the classification mode of the collected papers in this paper has stronger field characteristics than that in the existing related reviews [16–19]. Specifically, in the aspect of data augmentation, the data generation and data over-sampling models can effectively expand the fault dataset [25,27,28], and the data reweighting methods based on transfer learning can also augment the limited fault data with the help of other related datasets [13,29]. The research achievements indicate that augmented data improve the diagnosis accuracies in S&I-IFD effectively. In the aspect of feature learning, fault features can be extracted directly from small & imbalanced data by designing the regularized neural networks [23,30,31], and feature adaptation based on transfer learning is also useful for learning features from limited fault data to achieve accurate fault identification [32–34]. In the aspect of classifier design, it is expected to achieve accurate fault identification by modifying SVM or designing a cost-sensitive fault

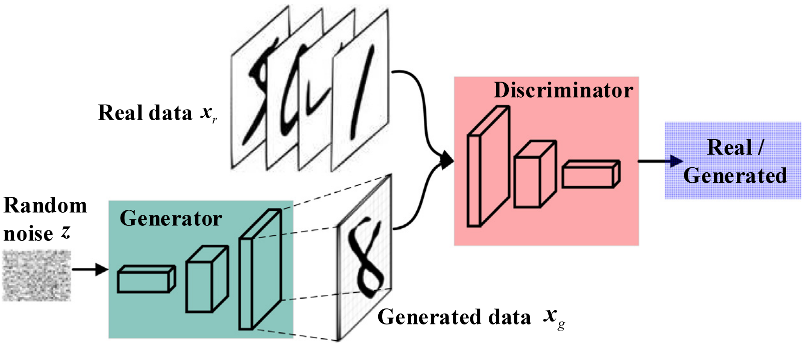


Fig. 4. Structure of GAN.

classifier [35–40]. Besides, the classifier design scheme based on parameter-transfer learning also shows effectiveness in the case of limited fault samples [41–43].

3. Data augmentation-based strategy for S&I-IFD

3.1. Motivation

Data-driven intelligent fault diagnosis has been widely studied. Research results have demonstrated that data-driven intelligent fault diagnosis models can usually achieve good diagnosis performance [44]. However, in engineering scenarios, machine faulty samples are hard to be collected, which is an important factor restricting the data-driven intelligent diagnosis models to be utilized. As an efficient approach to enhance the neural networks' generalization performance, data augmentation [17] presents a good solution for S&I-IFD. With a few faulty data for data generation [45,46], data over-sampling [47,48], or data reweighting [13,29], the limited fault dataset can be augmented to train the intelligent diagnosis models effectively. As a result, intelligent diagnosis models are expected to have strong diagnosis ability in the case of lacking fault data samples.

3.2. Data generation using generative models

Recently, data generation models represented by generative adversarial networks (GAN) [49] and Variational Auto-Encoder (VAE) [50] have been deeply studied and shown bright results in many fields [51]. Fortunately, these generative models can also be used to generate mechanical signals, providing a powerful tool for data augmentation in S&I-IFD [52].

3.2.1. GAN-Based methods

3.2.1.1. Introduction to GAN. GAN has two multi-layer neural network modules named Generator and Discriminator, as depicted in Fig. 4. Generator samples random noise $\boldsymbol{z}$ from distribution p_z and then generates data $\boldsymbol{x}_g$, while Discriminator outputs a probability scalar quantity to distinguish real data $\boldsymbol{x}_r$ and generated data $\boldsymbol{x}_g$. Given $G(\cdot)$ is the operation in Generator, $D(\cdot)$ is the operation in Discriminator, L_G is the objective function of Generator.

$$L_G = \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right)\right)\right)\right].(1) \tag{1}$$

For Discriminator, L_D is the objective function.

$$L_D = -\mathrm{E}_{\boldsymbol{x}\sim p_r}\left[\log D\left(\boldsymbol{x}\right)\right] - \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right)\right)\right)\right] \tag{2}$$

where p_r represents the real data distribution.

As a result, GAN has the overall objective function as follows:

$$\min_G \max_D W_{G,D} = \mathrm{E}_{\boldsymbol{x}\sim p_r}\left[\log D\left(\boldsymbol{x}\right)\right] + \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right)\right)\right)\right]. \tag{3}$$

The specific training process of GAN can be described as follows:

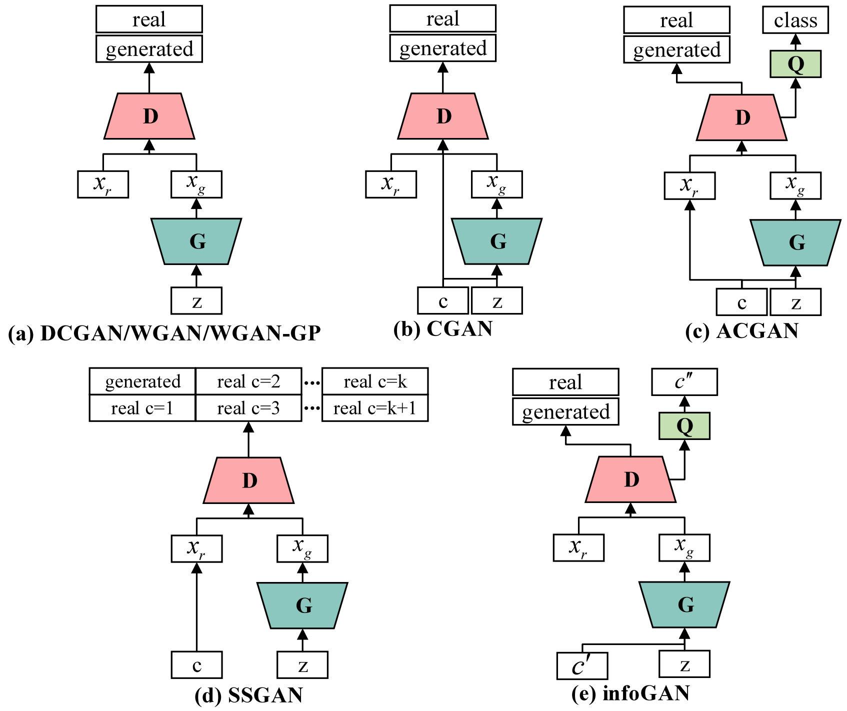
Based on the original GAN, scholars have made many improvements on it and created many variants since its birth. For example, Deep Convolutional GAN (DCGAN) [53] uses deep convolutional neural networks to build Generator and Discriminator, which makes it possible to generate high-quality images. Wasserstein GAN (WGAN) [54] applies Wasserstein distance to modify the original loss function, which makes the training process more stable than the original GAN. Wasserstein GAN with Gradient Penalty (WGAN-GP) [55] applies gradient penalty to Discriminator to further stabilize the training process. Conditional GAN (CGAN) [56] introduces the class information of the real data into the training of GAN, which enables the model to generate labeled data samples. Auxiliary Classifier GAN (ACGAN) [57] adds a classifier to Discriminator to generate labeled data samples. Semi-supervised GAN (SSGAN) [58] realizes semi-supervised data classification by constructing pseudo labels for the unlabeled data samples. Information maximizing GAN (infoGAN) [59] can learn the disentangled feature representation by inputting latent code into Generator so that the learned features are interpretable.

For simplicity, we use G and D to represent Generator and Discriminator. Q represents classifier. c is the class information of the input data. k is the class number. λ and α are real numbers less than 1. c' and c'' denote the input latent code and the reconstructed latent code. $L_I(\cdot)$ represents the calculation of mutual information. As shown in Fig. 5, we summarize several common variants of GAN, and their objective functions are given in Table 1.

3.2.1.2. Applications of GAN to data generation. The applications of GAN to generate data for S&I-IFD are summarized in Table 2. The research achievements show that the fault data augmented by GAN can improve the faults identification performance of gears [65], bearings [66], rotors [52], and other components [67] effectively in the case of limited fault data. According to the data dimension, these research results can be divided into two categories: one-dimensional samples (1-D) generation and two-dimensional samples (2-D) generation. Among them, the generation of 1-D data can be classified into three types. The first is to generate raw signals directly [12,27,52,60–64,77]. GAN and its variants are applied to generate the monitoring signals of machines, and the generated signals can be used to train the intelligent diagnosis models directly. For example, Zhang et al. [12] used a deep gradient penalized GAN to generate bearings' vibration data, which expands training datasets effectively. The presented work in [12] was one of the earliest research using GAN for mechanical signals augmentation, which also designed an index based on correlation coefficients to measure the generated samples' quality. The second is to generate the frequency spectrum of the monitoring signals [65–73]. Compared with raw monitoring data, the frequency spectrum also contains abundant fault information and is widely used in machine fault identification. For example, Wang et al. [65] adopted GAN to generate the

Algorithm 1: GAN

```
Input: real data $x_r$, random noise $z$, training epochs $K$, hyper-parameter $k$
Output: network weights in Generator ($\theta_G$) and Discriminator ($\theta_D$)
for the training epochs $K$ do
    for $k$ do
        Take $m$ data samples $\{z^{(1)},\dots,z^{(m)}\}$ from $p_z$.
        Take $m$ data samples $\{x_r^{(1)},\dots,x_r^{(m)}\}$ from $p_r$.
        Ascend the stochastic gradient to update the network weights in Discriminator:
            $\nabla_{\theta_D}\frac{1}{m}\sum_{i=1}^{m}\left[\log D\left(x_r^{(i)}\right)+\log\left(1-D\left(G\left(z^{(i)}\right)\right)\right)\right].$
    end for
    Take $m$ data samples $\{z^{(1)},\dots,z^{(m)}\}$ from $p_z$.
    Ascend the stochastic gradient to update the network weights in Generator:
        $\nabla_{\theta_G}\frac{1}{m}\sum_{i=1}^{m}\log\left(1-D\left(G\left(z^{(i)}\right)\right)\right).$
end for
Note: Mini-batch stochastic gradient descent algorithm is used to update weights.
```



(a) DCGAN/WGAN/WGAN-GP

(b) CGAN

(c) ACGAN

(d) SSGAN

(e) infoGAN

Fig. 5. Variants of GAN.

gearbox's signal frequency spectrums. The generated frequency spectrums with the real ones were used to train a Stacked Auto-encoder (SAE) together, which achieves high diagnostic accuracy and good anti-noise ability. The third is to generate the extracted data features [46,74]. The generated fault features can also be used to train the fault classifier directly. For instance, Zhou et al. [74] used auto-encoder (AE) to extract fault features from monitoring data, and the extracted features were generated by a global optimization GAN. The generated and the real fault features were used for accurate fault identification by deep neural

Table 1
The objective functions of the variants of GAN.

Name	Objective function
DCGAN	$L_D^{DCGAN} = -\mathrm{E}_{\boldsymbol{x}\sim p_r}\left[\log D\left(\boldsymbol{x}\right)\right] - \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right)\right)\right)\right]$ $L_G^{DCGAN} = \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right)\right)\right)\right]$
WGAN	$L_D^{WGAN} = -\mathrm{E}_{\boldsymbol{x}\sim p_r}\left[D\left(\boldsymbol{x}\right)\right] + \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[D\left(G\left(\boldsymbol{z}\right)\right)\right]$ $L_G^{WGAN} = -\mathrm{E}_{\boldsymbol{z}\sim p_z}\left[D\left(G\left(\boldsymbol{z}\right)\right)\right]$
WGAN_GP	$L_D^{WGAN\_GP} = L_D^{WGAN} + \lambda\mathrm{E}_{(\boldsymbol{x},\boldsymbol{z})\sim(P_r,P_z)}\left[\left(\left\|\nabla D\left(\alpha\boldsymbol{x} - \left(1 - \alpha G\left(\boldsymbol{z}\right)\right)\right)\right\| - 1\right)^2\right]$ $L_G^{WGAN\_GP} = L_G^{WGAN}$
CGAN	$L_D^{CGAN} = -\mathrm{E}_{\boldsymbol{x}\sim p_r}\left[\log D\left(\boldsymbol{x}, c\right)\right] - \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right), c\right)\right)\right]$ $L_G^{CGAN} = \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right), c\right)\right)\right]$
ACGAN	$L_D^{ACGAN} = L_D^{DCGAN} - \mathrm{E}_{\boldsymbol{x}\sim P_r}\left[P\left(class = c \mid \boldsymbol{x}\right)\right] - \mathrm{E}_{\boldsymbol{z}\sim P_z}\left[P\left(class = c \mid G\left(\boldsymbol{z}\right)\right)\right]$ $L_G^{ACGAN} = L_G^{DCGAN} - \mathrm{E}_{\boldsymbol{z}\sim P_z}\left[P\left(class = c \mid G\left(\boldsymbol{z}\right)\right)\right]$
SSGAN	$L_D^{SSGAN} = L_D^{WGAN} - \mathrm{E}_{\boldsymbol{x}\sim P_r}\left[P\left(class = c \mid \boldsymbol{x}, c < k + 1\right)\right]$ $L_G^{SSGAN} = L_G^{WGAN} + \left\|\mathrm{E}_{\boldsymbol{x}\sim P_r} f\left(\boldsymbol{x}\right) - \mathrm{E}_{\boldsymbol{z}\sim P_z} f\left(G\left(\boldsymbol{z}\right)\right)\right\|^2$
infoGAN	$L_D^{infoGAN} = L_D^{DCGAN} - \lambda L_I\left(\boldsymbol{c}', \boldsymbol{c}''\right)$ $L_G^{infoGAN} = L_G^{DCGAN} - \lambda L_I\left(\boldsymbol{c}', \boldsymbol{c}''\right)$

Table 2
Applications of GAN to generate data in S&I-IFD.

Data dimension	Data types	Models	References
1-D	Raw signal	GAN/WGAN/WGAN-GP	Zhang et al. [12], Liu et al. [27], Yin et al. [60], Gao et al. [61], Zhang et al. [62], Zhang et al. [63]
		ACGAN	Shao et al. [52]
		infoGAN	Wu et al. [64]
	Frequency spectrum	GAN/WGAN/WGAN-GP	Wang et al. [65], Zou et al. [66], Wang et al. [67], Ding et al. [68], Mao et al. [69]
		CGAN	Wang et al. [70], Zheng et al. [71], Zheng et al. [72]
		ACGAN	Li et al. [73]
	Extracted feature	GAN/WGAN/WGAN-GP	Pan et al. [46], Zhou et al. [74]
2-D	Time–frequency spectrum	GAN/WGAN/WGAN-GP	Cabrera et al. [75]
		CGAN	Liu et al. [26], Yu et al. [45]
		SSGAN	Liang et al. [76]

networks. Since the dimension of features is generally lower than that of raw data, the generation of data features is easier and faster than that of raw data. However, the fault information contained in the generated features may not be as rich as the one in the raw data, which is one of the drawbacks of fault feature generation.

On the other hand, GAN is used for 2-D image generation originally, therefore, it is handy for processing 2-D data. In the field of machine fault diagnosis, researchers usually use wavelet transform (WT) and other methods [45,75,76] to obtain the time–frequency domain features of raw signals, which are 2-D data. GAN can generate time–frequency features of raw monitoring signals to serve the training of intelligent diagnosis models. Cabrera et al. [75] presented a deep diagnosis scheme based on GAN for imbalanced fault diagnosis, in which the 2-D time–frequency features are extracted using wavelet packet transform and augmented by GAN. Liang et al. [76] used continuous wavelet transform to extract time–frequency features of gearboxes' vibration data and a GAN was adapted to expanding the number of 2-D time–frequency features to train the diagnosis model.

As a popular data generation method, GAN has the ability to generate faulty samples similar to the real faulty samples collected from engineering scenarios, thus expanding the training dataset of the intelligent diagnosis model. However, there are still two problems when GAN is applied for fault data generation. First, GAN is difficult to train. In order to generate sufficient fault data, GAN consumes a large number of computing resources and needs a long training time. Second, although GAN can expand the volume of fault data, the data generation ability is limited when the training data is insufficient. Specifically, the original GAN needs massive data for training. The more training data is, the closer the data distribution learned by GAN is to the real data distribution. However, when only a few training data is available, it is easy to fall into mode collapse [55]. In this case, the generated samples approximate the copies of the real samples, which means that the fault information contained in the generated data is very limited. As a result, the fault identification accuracy of the diagnosis model cannot meet the requirement of engineering using such low-quality generated samples as training data. Therefore, despite many achievements have been yielded using GAN, there is huge research space on how to reduce the consumption of computing time and improve the data generation ability when the training data is insufficient.

3.2.2. VAE-Based methods

3.2.2.1. Introduction to VAE. Variational Auto-Encoder (VAE) [50] is another commonly used deep generative model, as shown in Fig. 6. In terms of data generation, VAE can sample from hidden variables and then generate more data.

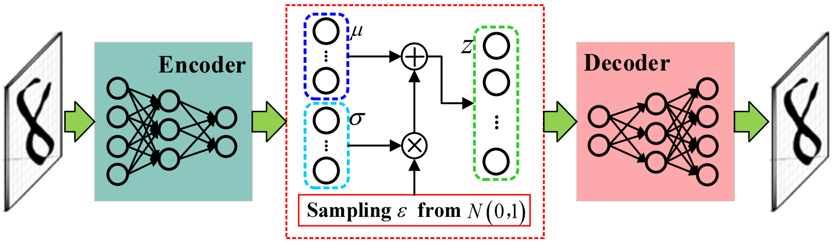


Fig. 6. Structure of VAE.

The input of the encoder is data $\boldsymbol{x}$, the output is the hidden variable $\boldsymbol{z}$, which is composed of μ and σ, and the weights and biases of the encoder are θ. In the training, the posterior distribution $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ will be learned by the encoder. The hidden variable $\boldsymbol{z}$ will be input into the decoder to reconstruct data, and the weights and biases of the decoder are ϑ. The distribution $p_\vartheta(\boldsymbol{x}|\boldsymbol{z})$ will be learned by the decoder.

The objective function can be expressed as

$$L_i(\theta,\vartheta) = -E_{\boldsymbol{z}\sim q_\theta(\boldsymbol{z}|\boldsymbol{x}_i)}\left[\log\left(p_\vartheta(\boldsymbol{x}_i|\boldsymbol{z})\right)\right] + KL\left(q_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\,\|\,p(\boldsymbol{z})\right) \quad (4)$$

where $p(\boldsymbol{z})$ is the hidden variable's prior distribution. $KL(\cdot)$ denotes the Kullback–Leibler divergence. In VAE, $p(\boldsymbol{z})$ is the normal distribution $N(\boldsymbol{z};0,1)$. $q_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$ is the normal distribution $N\left(\boldsymbol{z};\mu_i,\sigma_i^2\right)$. Thus, the $KL(\cdot)$ between $q_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$ and $p(\boldsymbol{z})$ can be described as

$$KL\left(q_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\,\|\,p(\boldsymbol{z})\right) = -\frac{1}{2}\sum_{j=1}^{J}\left(1+\log\left(\left(\sigma_i^j\right)^2\right)-\left(\mu_i^j\right)^2-\left(\sigma_i^j\right)^2\right) \quad (5)$$

where J is the dimension of the hidden variable $\boldsymbol{z}$.

In Eq. (5), μ_i and σ_i can be computed by the encoder directly. The hidden variable $\boldsymbol{z}$ is calculated by

$$z_i = \mu_i + \sigma_i\varepsilon \quad (6)$$

where $\varepsilon \sim N(0,1)$ is a noise variable, as given in Fig. 6.

In VAE, the output data has a high similarity to the input because the data reconstruction loss is optimized in the training process. Meanwhile, due to the addition of the noise variable ε, the generated data will not be completely consistent with the input data, thus achieving data augmentation.

3.2.2.2. Applications of VAE to data generation. In intelligent fault diagnosis, VAE has been utilized to generate fault data of gearboxes [70] and bearings [25,78]. For example, in [70], a diagnosis scheme based on VAE and GAN was proposed for imbalanced fault diagnosis, in which VAE was applied to generate the frequency spectrums of gearbox in different working conditions. Different from the traditional GAN, this scheme used VAE as the data generator and further improved the data generation ability of VAE through adversarial training. Dixit et al. [25] adopted a Conditional Variational Auto-encoder (CVAE) to generate faulty data of bearings, in which a centroid loss term was added to the original loss function of VAE. Zhao et al. [78] proposed an intelligent diagnosis model suitable for small and unbalanced monitoring data, in which a VAE was used to generate the vibration signals of machines. The signals generated by VAE had high similarity to the real signals in terms of time–frequency domain, which made the proposed diagnosis method possible to obtain higher accuracy than related works.

Similar to GAN, VAE can also be used for fault data generation, and the research achievements above have proved the effectiveness of VAE in S&I-IFD. Compared with GAN, the training process of VAE is more stable, and there is no problem of mode collapse [79]. However, due to the difference in the loss function, the data generated by VAE is usually not as real as the data generated by GAN [80]. As a result, the application of GAN to data augmentation is more popular than that of VAE [80]. Some scholars have tried to combine VAE and GAN to generate mechanical data [70]. In the future, how to make the data samples generated by VAE more real is a problem that needs to be solved.

3.3. Data over-sampling using sampling techniques

Although deep generative models like GAN and VAE can generate fault data to support the training of intelligent diagnosis models, these deep generative models are often difficult to train and require a large number of computing resources [51]. Taking into account this problem, data over-sampling using sampling techniques is another important way to augment limited data [19]. Some sampling techniques like Synthetic Minority Over-sampling Technique (SMOTE) [10], have yielded many achievements in S&I-IFD.

3.3.1. SMOTE-Based methods

3.3.1.1. Introduction to SMOTE. In general, researchers over-sample the minority classes or under-sample the majority classes to balance the dataset [19]. However, under-sampling will lose some valuable information that might be useful for data classification. On the other hand, over-sampling replicates training data randomly, which may lead to overfitting of classifiers [18]. Based on the random over-sampling, an improved method named SMOTE is proposed [81]. By analyzing the samples in the minority classes, SMOTE can synthesize more new samples. As given in Fig. 7, the process of SMOTE is described as follows:

(1) The Euclidean distance between the sample $\boldsymbol{x}$ and all the samples in the same class is calculated to obtain the k-nearest neighbors.
(2) For each sample $\boldsymbol{x}$, n samples $\{\boldsymbol{x}_i\}_{i=1}^n$ are randomly chosen within the range of the k-nearest neighbors.
(3) For each sample $\boldsymbol{x}_i$, the new synthesized sample $\boldsymbol{x}_{new}$ can be obtained as follows:

$$\boldsymbol{x}_{new} = \boldsymbol{x} + rand(0,1) * (\boldsymbol{x}_i - \boldsymbol{x}). \quad (7)$$

3.3.1.2. Applications of SMOTE to data over-sampling. Some scholars [28,47,48,82] have introduced SMOTE and its modified variants to over-sample the machine faulty samples. For example, Martin-Diaz et al. [10] used SMOTE to synthesize the fault samples of induction motors (IMs), in which the stator current signals in the minority classes were synthesized to balance the dataset. The results indicated that the balanced data constructed by SMOTE could help to improve the diagnosis performance effectively. An effective imbalanced data learning scheme named Easy-SMT was presented in [82]. Easy-SMT used SMOTE to augment the minority fault classes of wind turbines and Easy-Ensemble algorithm to transfer the imbalanced fault classification problem

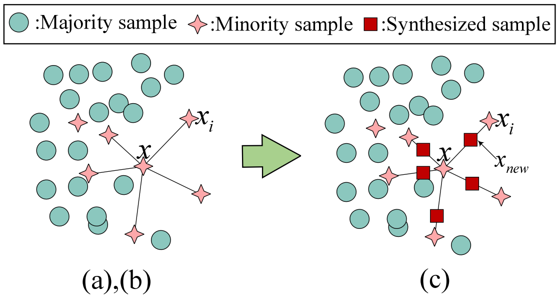


Fig. 7. New samples synthesized using SMOTE.

to a balanced one, making it possible to achieve good diagnosis performance. In [47], Wu et al. proposed an expectation–maximization minority over-sampling method based on SMOTE, in which a local-weighted strategy was applied to the expectation–maximization algorithm to learn and identify the hard-to-learn informative fault samples.

Compared with deep generative models, SMOTE requires fewer computing resources, so it is able to synthesize a large quantity of fault data samples to meet the demand of intelligent diagnosis models. However, SMOTE has the problem of data distributional marginalization when it is applied to synthesize data in the minority class. Specifically, if a fault sample is at the edge of the fault data distribution, the samples synthesized using this fault sample will also be at the edge of the distribution, which will blur the classification boundary [83]. Therefore, despite SMOTE improves the balance of the training dataset, it may increase the difficulty of fault classification when it falls into distributional marginalization.

3.4. Data reweighting using transfer learning

In addition to data generation and data over-sampling, data augmentation can also be achieved by reweighting data samples using transfer learning-based approaches with the help of other related datasets [13,29,84].

3.4.1. Introduction to transfer learning

In the case of lacking fault data, it is difficult to train a new intelligent diagnosis model [85]. However, this problem could be solved if the existing diagnosis knowledge learned by the trained diagnosis model could be reused. For example, we can use the bearing fault data collected in the laboratory to train a diagnosis model. The bearing fault diagnosis knowledge learned by this diagnosis model may be helpful for bearing fault identification in engineering scenarios. Transfer learning, which means that the knowledge learned from one task is reused in another task, is a promising tool for achieving this goal [86].

Generally speaking, transfer learning has three categories: instance-based transferring, feature-based transferring, and parameter-based transferring, depending on the components being transferred [86]. Among them, instance-based transferring aims to select some data samples from the source domain to improve the target task's performance in the case of limited target samples. Data reweighting is one of the most commonly used strategies of instance-based transferring. The weights of the selected target domain data samples will be increased while the weights of the selected source domain ones will be decreased. TrAdaBoost [87] is the most representative data reweighting algorithm in transfer learning.

3.4.2. TrAdaBoost-based methods

The source domain samples and the target domain ones will be reweighted by TrAdaBoost, so the contributions of the source and the target domain samples to the diagnosis model training can be balanced. In TrAdaBoost, if a target domain sample is misclassified by the diagnosis model, the weight of this sample will be increased because this sample is hard to be classified correctly. On the other hand, if a source domain sample is misclassified by the diagnosis model, the weight of this sample will be decreased because this sample is considered to be of little help to the training of the diagnosis model. Consequently, the classification boundary is moved to the direction of accurately identified the target data, as given in Fig. 8. As a result, the obtained diagnosis model based on TrAdaBoost algorithm will have a good classification accuracy on the target diagnosis task.

In intelligent fault diagnosis, TrAdaBoost algorithm has been used to handle the small sample condition. For example, Xiao et al. [13] presented a transfer learning scheme for machine fault diagnosis under the small sample condition, in which a TrAdaBoost algorithm was applied to assign weights to each training sample. The weighted samples helped to train a convolutional neural network-based learner. The proposed scheme obtained the highest diagnosis accuracy compared with related works in the case of inadequate target data. Shen et al. [29] applied the TrAdaBoost algorithm to update the weights of the selected auxiliary samples, the experimental results showed that the presented work was effective in bearing fault identification using small target data samples.

As a data reweighting algorithm, TrAdaBoost only operates on the data and does not participate in feature extraction and conditions identification. Therefore, it is easy to be combined with various advanced data classification models like deep belief networks and convolutional neural networks. However, the performance of data reweighting is connected with the similarity of the source and the target domain data distributions, if there is a large deviation between them, the TrAdaBoost-based data reweighting strategy may lead to negative transfer in the target diagnosis task [8], which means the reweighted fault samples may lead to a poor diagnosis performance.

3.5. Epilog

This section reviews the research results using data augmentation-based strategy in S&I-IFD. The data augmentation-based strategy in S&I-IFD has three categories: data generation using generative model, data over-sampling using sampling technique, and data reweighting using transfer learning. The first two methods can expand the volume of fault data effectively. However, they have the following two problems to be solved. First, deep generative models like GAN and VAE are often difficult to train and require many computing resources, which means they are not friendly to practical application. Moreover, when only a few samples are available for training, the generated faulty samples' quality is too low to meet the requirement of intelligent fault diagnosis models because these deep generative models usually need massive data to learn an authentic data distribution. Second, the sampling techniques represented by SMOTE have the problem of data marginal distribution, which may even increase the difficulty of accurate fault classification. Based on transfer learning, data reweighting can also augment limited fault data samples by increasing the selected data samples' weights with the help of other related datasets. However, data reweighting relies on the similarity of the source and the target domain data distributions, which is prone to reduce the performance of the diagnosis model. Therefore, it is necessary to find new data augmentation methods with high efficiency to improve the diagnosis performance on S&I-IFD further.

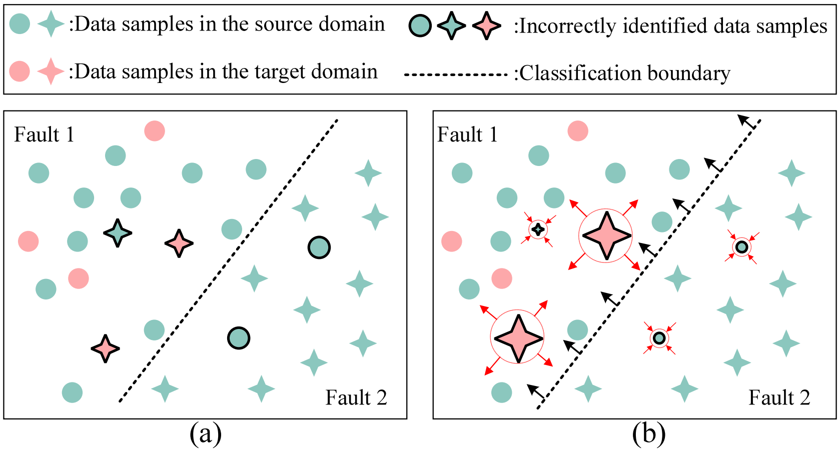

:Data samples in the source domain
:Incorrectly identified data samples
:Data samples in the target domain
:Classification boundary
Fault 1
Fault 2
(a)
Fault 1
Fault 2
(b)

Fig. 8. Illustration of TrAdaBoost: (a) the diagnosis model training with the source and the target domain samples directly, and (b) the diagnosis model training based on TrAdaBoost algorithm.

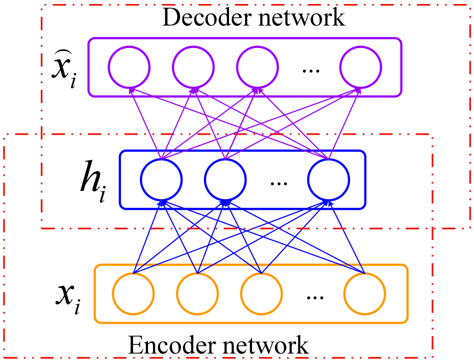


Fig. 9. Structure of AE.

4. Feature learning-based strategy for S&I-IFD

4.1. Motivation

In intelligent fault diagnosis, fault feature learning from machine monitoring data is the core link. The quality of the learned fault feature will affect the performance of machine fault diagnosis to a great extent. In addition to data augmentation, the problem S&I-IFD can also be solved if diagnosis models can learn effective fault features from small & imbalanced data. Scholars have done many works on how to learn fault features from small & imbalanced data. According to the existing results, the research ideas are mainly divided into the following two kinds. First, by designing regularized neural networks like sparse ones [23,30,31], diagnosis models can extract fault features from small & imbalanced data directly. Second, with the help of other related datasets, feature adaptation based on transfer learning can also learn fault features from small & imbalanced data to achieve accurate fault identification [32–34].

4.2. Feature extraction using regularized neural networks

The use of neural networks for fault feature extraction from monitoring data has been studied deeply. Recent research achievements show that regularized neural networks can process small & imbalanced data effectively [88–91]. Moreover, in these achievements, deep auto-encoders (DAE) and deep convolutional neural networks (DCNN) are favored as a basic model.

4.2.1. DAE And DCNN-based methods

4.2.1.1. Introduction to DAE. As shown in Fig. 9, Auto-encoder (AE) is a typical unsupervised model [8], which can reconstruct input data through the operation of encoder and decoder. The input is $\boldsymbol{x}_i$, $\boldsymbol{w}_e$ and $\boldsymbol{b}_e$ are the weight and bias of the encoding layer. The data features of the hidden layer $\boldsymbol{h}_i$ are expressed as

$$\boldsymbol{h}_i = f_e\left(\boldsymbol{w}_e \cdot \boldsymbol{x}_i + \boldsymbol{b}_e\right) \tag{8}$$

where f_e is the activation function in the encoder network. The weight and bias of the decoding layer are $\boldsymbol{w}_d$ and $\boldsymbol{b}_d$, the reconstructed data $\widehat{\boldsymbol{x}}_i$ can be defined as

$$\widehat{\boldsymbol{x}}_i = f_d\left(\boldsymbol{w}_d \cdot \boldsymbol{h}_i + \boldsymbol{b}_d\right) \tag{9}$$

where f_d is the activation function in the decoder network. By minimizing the loss $L\left(\boldsymbol{x}_i, \widehat{\boldsymbol{x}}_i\right)$, the input data can be reconstructed by AE.

$$L\left(\boldsymbol{x}_i, \widehat{\boldsymbol{x}}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i\right\|^2 \tag{10}$$

where n is the data points number.

In the decoder network, the low-dimensional data $\boldsymbol{h}_i$ is used to reconstruct the high-dimensional data $\widehat{\boldsymbol{x}}_i$. Thus, $\boldsymbol{h}_i$ can be regarded as the features of input data $\boldsymbol{x}_i$. By stacking multiple encoding layers and multiple decoding layers, DAE is constructed. Deep features of the input data can be collected using DAE through pre-training layer by layer, and the collected deep features are available for data classification using classifiers like Softmax [12].

4.2.1.2. Introduction to DCNN. Compared to AE, convolutional neural network (CNN) has fewer training parameters and stronger feature extraction ability [92]. CNN contains convolutional and pooling layers. The convolutional layer learns the feature vector of the input data by convolution operation. As given in Fig. 10(a), in the mth convolutional layer, the convolution kernel $\boldsymbol{k}^m \in \Re^{W\times D\times H}$ is used to learn the feature vector $\boldsymbol{x}^m$, where W is the kernel number, D is the kernel depth. H represents the kernel height. The wth feature vector $\boldsymbol{x}_w^m$ is obtained by

$$\boldsymbol{x}_w^m = \sigma\left(\sum_d \boldsymbol{k}_{w,d}^m \times \boldsymbol{x}_d^{m-1} + \boldsymbol{b}_w^m\right) \tag{11}$$

where σ denotes the activation function. $d = 1, 2, \ldots, D$, $w = 1, 2, \ldots, W$, $\boldsymbol{x}_d^{m-1}$ is the dth feature vector in the $m-1$th layer, and $\boldsymbol{b}_w^m$ is the bias of the wth layer.

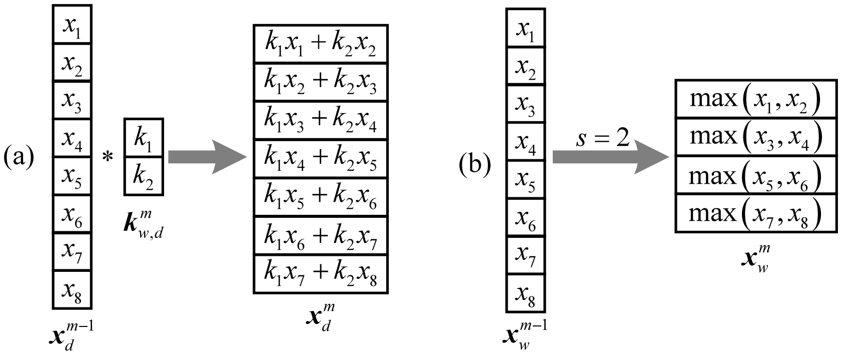


Fig. 10. Illustration of CNN. (a) The convolution operation, and (b) the pooling operation.

On the other hand, the pooling layer plays a role of down-sampling, it can reduce the size of the feature vector and the number of parameters, which is meaningful for accelerating convergence. Max pooling is the most common used pooling method, as shown in Fig. 10(b), in the mth pooling layer, the max-pooling is calculated by

$$\boldsymbol{x}_w^m = \text{down}\left(\boldsymbol{x}_w^{m-1}, s\right) \tag{12}$$

where down$(\cdot)$ is the function of down-sampling, and s is the pooling size.

Similar to DAE, DCNN can also be built by stacking convolutional and pooling layers. Benefiting from deep network structure, DCNN has stronger feature extraction capability than the shallow CNN so the high-dimensional complex data can be processed handily with DCNN.

4.2.1.3. Applications of regularized DAE and DCNN to feature extraction. DAE and DCNN with deep network structures often need a large volume of data for training, so they are not suitable for processing small & imbalanced data directly. Fortunately, regularization can help the training of DAE and DCNN with fewer training data while ensuring generalization ability. In intelligent fault diagnosis, regularized neural networks can extract fault features from a few fault samples and realize accurate fault classification. There are three commonly used regularized neural networks, i.e. sparse ones [23,30,31], normalized ones [24,93,94], and ensemble ones [95–97]. Among them, sparse neural networks will reduce the parameters of the network to decrease the risk of overfitting through weight decay, thus ensuring the generalization ability with limited training data. For example, Saufi et al. [31] presented a stacked sparse auto-encoder (SSAE) for gearbox fault diagnosis with limited fault data. Taking the Kullback–Leibler divergence as the sparse penalty term, the parameters to be trained in SSAE was reduced so diagnosis model can achieve better generalization performance and higher diagnosis accuracy than other deep neural networks using fewer training samples. Second, normalized neural networks will reduce the adverse effect of data imbalance on the training process by normalizing the weights, which ensures strong data classification ability in the case of imbalanced data distribution. For example, normalized convolutional networks (DNCNN) were used in [94] for imbalanced bearing faults identification. By applying a weights normalization strategy to construct the normalized convolutional and fully connected layer, the proposed DNCNN reduced the negative impact of data imbalance on fault classification. As a result, the proposed DNCNN was more effective in dealing with imbalanced fault classification than traditional CNNs. Finally, ensemble neural networks fuse data to prevent networks from overfitting in the case of small sample. In particular, there are two kinds of fused data, i.e. the extracted features [23,95] and the classification results [96,97]. For instance, Ren et al. [95] used a capsule network-based auto-encoder (CaAE) for intelligent fault identification of bearings, in which different local features were fused to construct the feature capsules. The feature capsules were input into a classifier for faults identification, and the experimental results showed that fused feature capsules were easier to obtain better diagnosis accuracies with small training samples than independent local features. An ensemble convolutional neural network (EnCNN) was proposed in [96] for imbalanced faults identification of machinery. In EnCNN, the imbalanced raw data were spilt into different training subsets to train a CNN-based classifier, the classification results from multiple basic classifiers were integrated by voting strategy. The integrated results were more conducive to realize accurate fault identification than a single result in the case of imbalanced training data.

In summary, DAE and DCNN have powerful data processing capability and can extract fault features from massive monitoring data automatically. However, such deep models update parameters by minimizing empirical risk, which means they are prone to overfitting when the training samples are insufficient [8]. Although recent studies have shown that regularized networks can improve their generalization ability, it must be noted that how to design high-quality regularization schemes for deep neural networks is a difficult problem requiring a large amount of research experience because there are many choices of regularization methods. Moreover, compared with the standard DAE and DCNN, the regularized network structure is generally more complex and difficult to train due to the introduction of other factors such as sparse penalty term.

4.2.2. Other algorithms-based methods

In addition to regularized DAE and DCNN, other neural networks have also achieved some results in feature learning from small and imbalanced data [89–91,98–100]. For example, Geng et al. [98] presented a diagnosis method based on a residual network with 17 convolutional layers for faults identification of bogie under imbalanced data condition. The deep residual learning framework with stacked non-linear rectification layers made it possible to learn discriminative fault features from imbalanced Fast kurtogram images of mechanical signals. Liu et al. [99] used the noise-assisted empirical mode decomposition for fault feature extraction from raw signals, and the extracted features were input into an enhanced fuzzy network for faults classification. Qian et al. [100] proposed an imbalanced learning scheme based on sparse filtering for fault feature extraction, which introduced a balancing matrix to balance the feature learning abilities of different classes. The results demonstrated that the presented feature learning model was effective for bearing fault diagnosis. In short, by modifying neural networks, fault features can be learned from small and imbalanced data, which is an important means to deal with S&I-IFD.

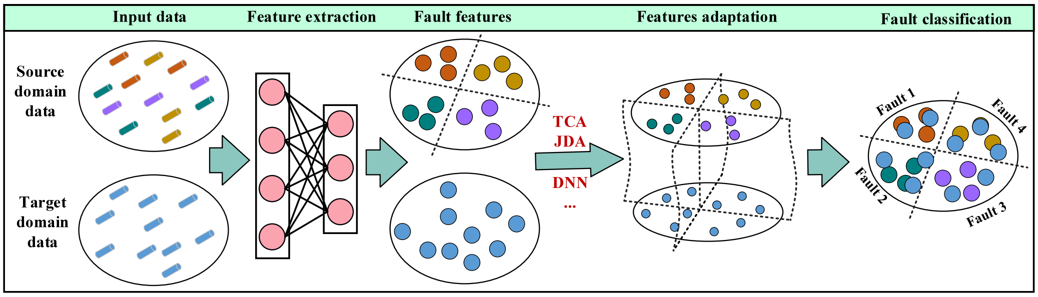


Fig. 11. Feature adaptation based on transfer learning.

4.3. Feature adaptation using transfer learning

In addition to extracting fault features directly, feature adaptation with the help of other related datasets is another important way to learn fault features from small and imbalanced data. In the transfer learning scenario, the volume of target domain data samples is usually much smaller than that in the source domain. Moreover, because of the difference between the source domain and the target domain data distributions, their features are generally different. Feature adaptation based on transfer learning attempts to minimize the discrepancy between the feature distributions in the two domains, so the feature of the target domain data can also be learned well by models, as shown in Fig. 11. Not only transfer component analysis (TCA) [101], many achievements in S&I-IFD have also been achieved by using joint distribution adaptation (JDA) [102], deep neural networks (DNN) [34], and other ways [103].

4.3.1. TCA And JDA-based methods

4.3.1.1. Introduction to TCA and JDA. TCA is a traditional feature adaptation method [101]. When the source domain data X^S has different distributions with the target domain data X^T, a feature mapping Φ is utilized to map them to high-dimensional Hilbert spaces, where the target domain data has the minimized distance with the source domain data.

The maximum mean discrepancy (MMD) is used by TCA to calculate the distance between $\Phi\left(X^S\right)$ and $\Phi\left(X^T\right)$, which is described as follows

$$dist\left(\Phi\left(X^S\right),\Phi\left(X^T\right)\right)=\left\|\frac{1}{n^S}\sum_{i=1}^{n^S}\Phi\left(\boldsymbol{x}_i^S\right)-\frac{1}{n^T}\sum_{i=1}^{n^T}\Phi\left(\boldsymbol{x}_i^T\right)\right\| \tag{13}$$

By using a kernel matrix $\boldsymbol{K}$ and $\boldsymbol{L}$, the MMD between $\Phi\left(X^S\right)$ and $\Phi\left(X^T\right)$ is rewritten to another form.

$$\boldsymbol{K}=\begin{bmatrix}\boldsymbol{K}_{S,S} & \boldsymbol{K}_{S,T}\\ \boldsymbol{K}_{T,S} & \boldsymbol{K}_{T,T}\end{bmatrix},L_{ij}=\begin{cases}\frac{1}{\left(n^S\right)^2},\boldsymbol{x}_i, & \boldsymbol{x}_j\in \mathrm{X}^S\\ \frac{1}{\left(n^T\right)^2},\boldsymbol{x}_i, & \boldsymbol{x}_j\in \mathrm{X}^T\\ -\frac{1}{n^S n^T}, & otherwise\end{cases} \tag{14}$$

$$dist\left(\Phi\left(X^S\right),\Phi\left(X^T\right)\right)=trace\left(\boldsymbol{KL}\right)-\lambda*trace\left(\boldsymbol{K}\right) \tag{15}$$

where λ is a tradeoff parameter. λ can be used to keep the balance of distributions adaptation and parameters complexity.

Finally, the optimization goal of TCA can be described as

$$\begin{aligned}&\min_{\boldsymbol{W}} trace(\boldsymbol{W^T KLKW})+\lambda*trace(\boldsymbol{W^T W})\\ &s.t.\boldsymbol{W^T KHKW}=\boldsymbol{I}_m\end{aligned} \tag{16}$$

where $\boldsymbol{H}=\boldsymbol{I}_{n^S+n^T}-1/\left(n^S+n^T\right)\boldsymbol{1}\boldsymbol{1}^T$ is a centering matrix, and $\boldsymbol{1}\in\mathbb{R}^{n^S+n^T}$ is an n^S+n^T dimensional column vector with elements of 1.

JDA is an improved variant based on TCA [102]. TCA only adapts the marginal probability distribution, while JDA adapts not only the marginal probability but also the conditional probability distribution between the source and the target domain data. As a result, the optimization goal of JDA is described as

$$\begin{aligned}&\min_{\boldsymbol{W}}\sum_{c=0}^{C} trace(\boldsymbol{W^T X L_c X^T W})+\lambda\left\|\boldsymbol{W}\right\|_F^2\\ &\text{s.t.}\boldsymbol{W^T X H X^T W}=\boldsymbol{I}\end{aligned} \tag{17}$$

where c is the class information. And $\boldsymbol{L}_c$ is

$$(\boldsymbol{L}_c)_{ij}=\begin{cases}\frac{1}{\left(n^{S,c}\right)^2}, & \boldsymbol{x}_i,\boldsymbol{x}_j\in\mathrm{X}^{S,c}\\ \frac{1}{\left(n^{T,c}\right)^2}, & \boldsymbol{x}_i,\boldsymbol{x}_j\in\mathrm{X}^{T,c}\\ -\frac{1}{n^{S,c}n^{T,c}}, & \begin{cases}\boldsymbol{x}_i\in\mathrm{X}^{S,c},\boldsymbol{x}_j\in\mathrm{X}^{T,c}\\ \boldsymbol{x}_i\in\mathrm{X}^{T,c},\boldsymbol{x}_j\in\mathrm{X}^{S,c}\end{cases}\\ 0, & otherwise\end{cases} \tag{18}$$

where the sample number in class c from the source domain is $n^{S,c}$ and that from the target domain is $n^{T,c}$.

4.3.1.2. Applications of TCA and JDA to feature adaptation. Some scholars introduced TCA and JDA to their transfer learning scheme for feature adaptation. For instance, Chen et al. [104] used a transfer learning faults identification method for rolling bearings using a few faulty samples, in which TCA was applied for feature adaptation to learn the transferable fault features from raw data. Xie et al. [105] and Duan et al. [106] extracted transferable fault feature from gearbox vibration signals using TCA, and the experimental results showed that their models were effective for gearbox faults identification in the small sample case. Besides, Han et al. [107] and Qian et al. [108] applied JDA for transferable features learning considering the problem of lacking target domain samples, the effectiveness of feature adaptation was verified using a bearing dataset and a gearbox dataset respectively.

Traditional TCA and JDA based feature adaptation approaches are simple in the calculation and can reduce the discrepancy of the feature distributions in the two domains. However, both TCA and JDA narrow the difference between two distributions by mapping low-dimensional raw data to high-dimensional Hilbert space. When they meet complex high-dimensional mechanical data, they cannot fit them well. Thus, the diagnosis accuracy of TCA and JDA related models on the complex diagnosis task is usually poor.

4.3.2. Deep neural networks-based methods

Different from TCA and JDA, deep neural networks can learn data features from the original data samples directly by minimizing the distribution discrepancy of the target and the source domain features. As a basic distance metric of distribution discrepancy, some scholars built deep transfer diagnosis models based on Kullback–Leibler (KL) divergence to achieve feature adaptation. For example, a transfer network was constructed by Qian et al. [109] for machine faults identification, in which a distribution discrepancy measuring metric named auto-balanced KL divergence (AHKL) was developed for fault feature adaptation. After feature extraction, the first and the higher-order moment discrepancies of the features from two domains was measured by AHKL, and the discrepancies between them were reduced by

$$\min \sum_{i=1}^{N}\left[\boldsymbol{\mu}^{i}\cdot \mathbf{L}_{1}^{i} + \left(\mathbf{1}^{i} - \boldsymbol{\mu}^{i}\right)\cdot \sum_{j=1}^{n}\mathbf{L}_{j}^{i}\right] s.t.0 \le \mu^{i} \le 1 \tag{19}$$

where the data points number of each sample is N. The order moments number is n. The discrepancy vector of the nth order moment is $\mathbf{L}_n$. $\boldsymbol{\mu}^i$ is a parameter vector to weigh between $\mathbf{L}_1$ and $\sum_{j=2}^{n}\mathbf{L}_j$.

In addition to KL divergence, another distance metric for measuring distribution discrepancies is the maximum mean discrepancy (MMD). Many research achievements based on feature adaptation using deep neural networks have applied MMD to develop their diagnosis scheme to deal with the small sample problem [110]. For example, Li et al. [32] developed a deep balanced feature adaptation model with multiple convolutional layers for gearboxes fault diagnosis using limited labeled data samples. The fault features were extracted from raw data, and then MMD was applied to measure the discrepancy of the conditional and the marginal probability distributions of the extracted features. The presented network was optimized by

$$\min_{\theta} \sum_{j=1}^{N}\left(\lambda D_{\theta}^{j}\left(X_{M}^{S}, X_{M}^{T}\right) + (1-\lambda)\sum_{i=1}^{n}\left(D_{\theta}^{i}\left(X_{C_i}^{S}, X_{C_i}^{T}\right)\right)\right) \tag{20}$$

where $D_{\theta}^{j}\left(X_{M}^{S}, X_{M}^{T}\right)$ is the discrepancy of the marginal probability distribution in the jth network layer. $D_{\theta}^{i}\left(X_{C_i}^{S}, X_{C_i}^{T}\right)$ is the discrepancy of the conditional probability distribution in ith class. The network layers number and the class number are N and n. λ is a real number less than 1. To further improve the performance of feature adaptation, many variants based on the original MMD are proposed by scholars. For instance, Yang et al. [33] constructed a convolutional adaptation scheme by minimizing multi-kernel MMD. A multi-layer MMD based feature adaptation framework was presented by Li et al. [34] to identify bearing faults using a few faulty samples.

Despite MMD is effective in measuring distribution discrepancy, the computational cost of MMD increases fast as the number of samples increases. Compared with MMD, Wasserstein distance is a more reasonable distance metric when measuring distribution discrepancy, which has also been used in the feature adaptation tasks. Cheng et al. [111] used a deep feature adaptation scheme for faults classification using a few labeled target samples, in which Wasserstein distance was utilized to calculate the discrepancy between the target and the source domain features. The proposed method was trained by minimizing the Wasserstein distance between the features from the two domains, which can be described as

$$\min_{\theta} \frac{1}{n^S}\sum_{i=1}^{n^S} f_L\left(f_\theta\left(\boldsymbol{x}_i^s\right)\right) - \frac{1}{n^T}\sum_{i=1}^{n^T} f_L\left(f_\theta\left(\boldsymbol{x}_i^T\right)\right) \tag{21}$$

where f_θ denotes the convolutional feature extractor. f_L is the Lipschitz function to satisfy the gradient constraint in calculating Wasserstein distance. The samples number in the source and the target domain are n^S and n^T respectively.

In addition to minimizing distance metric, another way for feature adaptation using deep neural networks is adversarial training. Inspired by GAN, adversarial training can also reduce the distribution discrepancy of two distributions. For example, Han et al. [103] constructed an adversarial transfer learning model for wind turbine fault diagnosis using limited training samples. In the presented work, the feature descriptor composed of multiple convolutional layers extracted fault features from the samples in the two domains. The discrepancy of the two feature distributions was minimized by a discriminative classifier through adversarial training. And the health conditions were output by a fault classifier in the end. The proposed method was trained by

$$\begin{aligned}&\min_{\theta} \frac{1}{n^S}\sum_{i=1}^{n^S} J\left(y_i^s, \widetilde{y}_i^s\right)\\ &- \left[\frac{1}{n^S}\sum_{i=1}^{n^S}\log D_\theta\left(\boldsymbol{x}_i^s\right) + \frac{1}{n^T}\sum_{j=1}^{n^T}\log\left(1 - D_\theta\left(\boldsymbol{x}_j^T\right)\right)\right]\end{aligned} \tag{22}$$

where the classification loss is the first term and the adversarial loss between the two feature distributions is the second term. After adversarial training, the diagnosis model could also serve well in the target diagnosis tasks.

Due to the strong data processing ability, deep neural networks-based feature adaptation approaches can usually output better diagnosis results than the traditional TCA and JDA. Nevertheless, the feature adaptation ability depends on the distance metric sometimes. Besides, deep neural networks-based feature adaptation schemes assume that the feature spaces of the two domains overlap to some extent, however, existing studies cannot tell whether there is overlap between them. The diagnosis models may perform poorly on the target diagnosis task if the discrepancy of the feature distributions cannot be described explicitly.

4.4. Epilog

The achievements on S&I-IFD using feature learning-based strategy are reviewed in this section, which are divided into two classes. The first is to use regularized neural networks like sparse ones to extract fault features from limited fault data directly. The second is feature adaptation with the help of other related datasets based on transfer learning. Through feature adaptation, transferable fault features are expected to be learned by diagnosis models to achieve accurate fault classification. However, the feature learning-based strategy also has shortcomings. First, since the fault information provided by a small number of fault data is always limited, the diagnosis performance improved by the feature learning-based models is also limited. Second, feature adaptation based on transfer learning requires the similarity of feature distributions between different datasets. However, in engineering scenarios, it is difficult to construct an auxiliary transferable dataset. Moreover, feature adaptation usually involves the selection of the distance metric, which makes it hard to achieve the optimal diagnosis results.

5. Classifier design-based strategy for S&I-IFD

5.1. Motivation

In the process of intelligent fault diagnosis, fault identification using a fault classifier is the last step. The classification performance of the fault classifier is an important index to determine the fault identification accuracy. In the case of lacking fault data, the trained classifier is usually over-fitted and the classification accuracy is low. If the fault classifier can be designed to have strong generalization ability for small and imbalanced data, it is hopeful to achieve accurate fault identification in the case of lacking machine fault data. Scholars have also done a lot of work on S&I-IFD from the perspective of fault classifier design. According to whether the auxiliary datasets are used or not, the design of the fault classifier follows two ideas. The first is to use the small and imbalanced data to modify the original fault classifier directly, such as constructing a cost-sensitive faults classifier [38–40]. The second is to pre-train the classifier with the help of other related datasets based on transfer learning to achieve good classification performance [41,42,112].

5.2. Fault classifier design using small and imbalanced data

In this part, fault classifiers are designed based on small and imbalanced data directly. As a specialized model for processing small samples, support vector machine (SVM) [8] and its variants can improve the faults classification accuracy with limited faulty data samples [113]. Besides, cost-sensitive learning [19] is dedicated to learning information from imbalanced data distributions by applying the cost-sensitive loss function. The cost-sensitive learning-based fault classifier can also provide an effective solution for S&I-IFD.

5.2.1. SVM-based methods

5.2.1.1. Introduction to SVM. SVM is a classical data classifier. As given in Fig. 12, SVM aims at finding a hyperplane in the features space, which is expected to correctly classify data samples as far as possible.

For a training dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{M}$, $\boldsymbol{x}_i$ is the ith sample and the sample label is $y_i \in [1, -1]$. The hyperplane $H(\boldsymbol{x})$ can be described as

$$H(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b = \sum_{i=1}^{M} \boldsymbol{w} \cdot \boldsymbol{x}_i + b = 0 \tag{23}$$

where the parameters of $H(\boldsymbol{x})$ are $\boldsymbol{w}$ and b. Moreover, to classify the data samples into two classes (the positive one and the negative one), $H(\boldsymbol{x})$ should be subject to

$$y_i H(\boldsymbol{x}_i) = y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i = 1, 2, \ldots, M. \tag{24}$$

As given in Fig. 12, $H'(\boldsymbol{x})$ and $H''(\boldsymbol{x})$ are the two hyperplanes satisfying the constraints in Eq. (24). The distance from $\boldsymbol{x}_i$ to $H(\boldsymbol{x})$ can be calculated as d_i.

$$d_i = \frac{y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|}. \tag{25}$$

Therefore, the margin γ between $H'(\boldsymbol{x})$ and $H''(\boldsymbol{x})$ is $\frac{2}{\|\boldsymbol{w}\|}$. As a result, SVM will find the hyperplane $H(\boldsymbol{x})$ between $H'(\boldsymbol{x})$ and $H''(\boldsymbol{x})$, which can maximize the margin γ by optimizing the objective loss function L.

$$L = \arg\max_{\boldsymbol{w},b}\left\{\min\left(\frac{y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|}\right)\right\} = \arg\max_{\boldsymbol{w},b}\left(\frac{2}{\|\boldsymbol{w}\|}\right). \tag{26}$$

For the convenience of calculation, the loss function L is rewritten as follows:

$$\begin{aligned} &L = \min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t.} y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i = 1, 2, \ldots, M \end{aligned} . \tag{27}$$

5.2.1.2. Applications of SVMs to faults classification. Some researchers utilized SVM and its variants to classify limited fault data [114–119]. For example, a K-means based SVM-tree and SVM-forest were developed in [114], in which the K-means algorithm was introduced to SVM for sensitive samples selection from an imbalanced dataset. The results indicated that the presented network improved the diagnosis performance using a few faulty data samples. Xi et al. [116] proposed a least-squares SVM (LSSVM-CIL) with parameter regularization for the imbalanced fault detection of aircraft engines, in which the size of support vectors was reduced and the representative fault samples were retained using recursive strategy. The experimental results proved that LSSVM-CIL was more effective than related methods in imbalanced fault detection. Based on the traditional SVM, He et al. [118] presented a nonlinear support tensor machine containing dynamic penalty factor (DC-NSTM) for faults identification of machines in the limited faulty samples case. A tensor kernel function was added to the DC-NSTM so that it could process the nonlinear separable problem and improve the overall classification accuracy with small training samples.

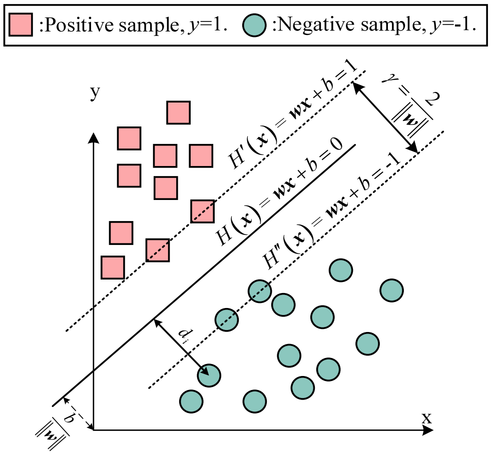


Fig. 12. Illustration of SVM.

Generally, the SVMs based fault classifiers are optimized by minimizing the overall structural risk of training samples [8], so they are more suitable for dealing with the limited fault data compared to deep neural networks, which are optimized by minimizing the empirical risk. However, two drawbacks restrict the applications of SVM. First, the diagnosis accuracy of SVM is sensitive to the setting of kernel parameters. How to choose a set of high-quality kernel parameters is one of the core issues when using the SVM-based fault classifier. Second, although SVM is good at handling small sample problems, it is difficult to fit massive monitoring data. With the development of data acquisition technologies, the monitoring data of machines increases rapidly, which will bring computing challenges to the SVM-based fault classifier.

5.2.2. Cost-sensitive classifier-based methods

5.2.2.1. Introduction to cost-sensitive learning. As a learning paradigm, Cost-sensitive learning [19] will give different misclassification losses to different classes contained in a classification task. Cost-sensitive learning aims at reducing all misclassification costs on the whole dataset. In other words, cost-sensitive learning will give more attention to the samples in the minority classes to improve the overall classification performance on imbalanced datasets.

Given a training dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{M}$ containing M training samples, the ith sample is $\boldsymbol{x}_i$ and the ith sample label is $y_i \in [1, 2, \ldots, K]$. Assume a misclassification loss $C_{u,v}$, which represents the loss or the penalty of misclassifying the sample $\boldsymbol{x}_i$ in class u to class v. For a classification task, the minimum misclassification loss should be achieved when classifying the sample $\boldsymbol{x}_i$ into a class. Specifically, the misclassification loss $L(u|\boldsymbol{x}_i)$ of sample $\boldsymbol{x}_i$ classified into class u can be described as

$$L(u|\boldsymbol{x}_i) = \sum_{v=1}^{K} P(v|\boldsymbol{x}_i)\, C_{u,v} \tag{28}$$

where $P\left(v|\boldsymbol{x}_i\right)$ denotes the probability distribution of classifying the sample $\boldsymbol{x}_i$ into the class v. Moreover, if $u=v$, $C_{u,v}=0$, which is the loss of classifying sample $\boldsymbol{x}_i$ correctly. Therefore, the overall expected misclassification cost on the training dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{M}$ can be described as

$$L\left(C\right)=\sum_{i=1}^{M}\sum_{v=1}^{K}P\left(v|\boldsymbol{x}_i\right)C_{u,v}. \tag{29}$$

Finally, the ideal classifier will make a decision by minimizing the overall expected misclassification cost $L\left(C\right)$.

5.2.2.2. Applications of cost-sensitive classifier to faults classification. In S&I-IFD, how to design and assign the misclassification loss $C_{u,v}$ is the key to the applications of cost-sensitive learning. For an imbalanced dataset, the imbalance ratio is an important index to measure the imbalance degree. For a training dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{M}$, the ith sample is $\boldsymbol{x}_i$ and the ith sample label is $y_i\in[1,2,\ldots,K]$. The imbalance ratio $r_{u,v}$ of the class u to the class v is defined as

$$n\left(k\right)=\sum_{i=1}^{M}1\left\{y_i=k\right\} \tag{30}$$

$$r_{u,v}=\frac{n\left(v\right)}{n\left(u\right)} \tag{31}$$

where $n\left(k\right)$ represents the number of samples in class k and $1\left\{\cdot\right\}$ is an indicator function returning 1 if $y_i=k$ and 0 otherwise. It is a common choice that the misclassification loss $C_{u,v}$ is designed based on the data imbalance ratio, i.e., if $u\neq v$, $C_{u,v}=r_{u,v}$, and if $u=v$, $C_{u,v}=0$. By this design, the classification model will pay more attention to the minority classes to improve the identification accuracy of the minority classes. Many studies have shown the effectiveness of applying the imbalance ratio into the design of the cost-sensitive loss function [94,98,120]. For example, Geng et al. [98] presented a diagnosis scheme using deep residual feature learning, in which the imbalance-weighted cross-entropy (IWCE) was used for imbalanced fault classification. The original cross-entropy (CE) can be described as

$$CE=-\sum_{i=1}^{K}\vec{y}_i\log\hat{P}_i \tag{32}$$

where the class number is K. $\vec{y}_i$ is the one-hot vector representing labels information and $\hat{P}_i$ denotes the output of the softmax classifier. Based on the original CE, IWCE used the data imbalance ratios to weight the minority classes to enhance the samples' influence in the minority classes.

$$IWCE=-\sum_{i=1}^{K}w_i\vec{y}_i\log\hat{P}_i \tag{33}$$

where w_i is a function just related to the data imbalance ratios.

Besides, some researchers have combined the real-time classification results and the data imbalance ratios to design the cost-sensitive loss function because the real-time training results are thought to be able to indicate the updating of parameters [38,39,121]. For instance, Dong et al. [38] adopted a cost-adaptive network structure for imbalanced mechanical data classification, in which the cost-sensitive loss function L was designed as follows

$$L=-\sum_{i=1}^{K}t_i\vec{y}_i\log\hat{P}_i \tag{34}$$

where t_i is a function related to the data imbalance ratios, the evaluation metric G_{mean}, and the Euclidean distance E_d.

$$t_i=r_i*\exp\left(-\frac{G_{mean}}{2}\right)*\exp\left(-\frac{1}{2E_d}\right) \tag{35}$$

$$G_{mean}=\sqrt{\frac{TP}{TP+FN}*\frac{TN}{TN+FP}} \tag{36}$$

$$E_d=\frac{1}{n\left(k\right)}\sqrt{\sum_{i=1}^{n(k)}\left(\vec{y}_i-\hat{P}_i\right)^2} \tag{37}$$

where r_i is the data imbalance ratio. True positive, false positive, true negative and false negative are represented by TP, FP, TN, and FN.

On the whole, cost-sensitive learning pays more attention to the fault samples in the minority classes through misclassification losses assignment, which ensures the fault identification accuracy of the minority fault samples. The output of the cost-sensitive fault classifier is sensitive to the design of the cost-sensitive loss function. Most of the current research achievements set the cost-sensitive loss function based on the data imbalance ratios, which is indeed effective, but how to update it to obtain better results is still worth exploring. In the future, one of the possible solutions is to set the cost-sensitive loss function automatically using the attention mechanism [122], which has been applied in sensitive information selection and adaptive weights assignment successfully.

5.3. Fault classifier design using transfer learning

In this part, fault classifiers are designed with the help of other related datasets. In the transfer learning scenario, some model parameters can be shared by the target and the source domain data [86]. Based on this, scholars use parameter transfer-based approaches to design the classifier. After pre-training with the source domain data, the parameters of the faults classifier are fine-tuned using a few target domain samples. As a result, the fine-tuned fault classifier will be expected to achieve high classification accuracies in the diagnosis tasks.

5.3.1. Parameter transfer-based methods

In parameter transfer-based methods, the parameters of diagnosis models are first pre-trained using sufficient source domain data. After that, the classification layers of the pre-trained models are fine-tuned using a few target domain data. The idea of parameter transfer-based approaches is relatively simple, but it is widely used. For example, Kim et al. [43] and Li et al. [123] constructed parameter transfer-based fault classifiers with deep convolutional neural network (DCNN), the parameters were pre-trained using an existing dataset. After pre-training, the Softmax classifier in the last layer of DCNN was fine-tuned using another small dataset. The fine-tuned Softmax classifier was able to classify data samples in the new dataset. The methods had good diagnosis performance for bearings using small training samples. Similarly, to identify gearbox faults using small training samples, Cao et al. [124] and Wu et al. [125] applied parameters based transfer learning for fault classifier design with DCNN, the experimental results showed that the pre-trained fault classifiers could obtain high accuracy on target gearbox diagnosis tasks after fine-tuning.

Besides, some scholars believe that updating all the model parameters in the fine-tuning stage will be more helpful for accurate fault identification than just updating the classifier layers. Therefore, the fault classifier can be obtained after global parameters fine-tuning in this case. For example, He et al. [112] applied a transfer learning model based on multi-wavelet deep auto-encoder for the gearbox fault classifier design, in which all the model parameters were pre-trained using vibration data from one working condition and fine-tuned with vibration data from another working condition. After fine-tuning, the obtained classifier could achieve high diagnosis accuracy in the new working condition. Similarly, Li et al. [41] and He et al. [42] adopted deep transfer auto-encoders to design the faults classifier of bearings, and the fault classifiers were obtained after fine-tuning with a few target domain samples.

In general, the size of the source domain dataset will influence the classification accuracies of the fault classifiers obtained using parameter transfer-based approaches. The larger the source domain dataset used for pre-training is, the better the performance of the obtained fault classifier is. However, it is difficult to construct an ideal pre-training dataset in practice, which is one of the major problems in the application of parameter transfer-based fault classifier design. If the source domain dataset is not large enough, the fault classifier obtained in this way will have poor diagnosis performance on the target diagnosis tasks.

5.4. Epilog

This section reviews the achievements of dealing with S&I-IFD based on the classifier design strategy. According to whether auxiliary datasets are used or not, fault classifier design-based strategies have two ways. The first is designing fault classifiers using small and imbalanced data directly, such as optimizing SVM or developing cost-sensitive classifiers. This kind of method generally depends on the engineering experience of the researchers, especially the design of cost-sensitive loss functions, so the optimal results are difficult to be achieved. The second is to use auxiliary datasets to pre-train diagnosis models and then fine-tuning the classifier with a few fault data to get the final fault classifier. The performance of the fault classifier obtained in this way depends on the quality of the auxiliary dataset. When the auxiliary dataset is not large enough, the classification ability of the fault classifier is usually not strong enough.

6. Future challenges and possible extensions for S&I-IFD

In the end, we try to discuss future challenges and provide some possible extensions for S&I-IFD based on the existing research achievements.

6.1. How to improve the quality of the augmented samples in S&I-IFD?

Benefiting from new machine learning theories and technologies like GAN and VAE, many existing achievements have proved that the performance of S&I-IFD can be improved by expanding the size of the training samples set using data generation and over-sampling. However, by reviewing these research achievements, it can be found that the existing researches mainly focus on expanding the size of fault data samples and lack attention to the quality of the samples. Specifically, when the size of training samples is too small, the samples generated by generative models are too similar to the real samples, which means the fault information increased by this way is very limited. For the data over-sampling models like SMOTE, the synthesized fault samples have a strong linear relationship with the training samples due to the problem of distributional marginalization [83]. Although these generated samples can expand the size of training samples, it is not clear how much fault information they can provide for the training of the diagnosis models. If they cannot provide more fault information, the low-quality generated samples will have a limited improvement in the diagnosis performance of the intelligent diagnosis models.

In future researches, the authors believe that researchers need to pay attention not only to the size of samples but also to the quality of samples. First, in addition to data generation, data over-sampling, and data reweighting, more different data augmentation ways can be applied [126–129]. For example, Yu et al. [126] tried seven kinds of data augmentation strategies via hand-crafted rules to augment the vibration signals of rolling bearings, including local data reversing, local random reversing, global data reversing, local data zooming, global data zooming, local segment splicing, and noise addition. Compared with other data augmentation strategies such as data generation, these data augmentation methods require less computing resources and less computing time. Moreover, experimental results showed that these data augmentation methods could also improve the diagnosis performance of S&I-IFD significantly. Besides, the existing data augmentation strategies are often tailor-made for each dataset and cannot be easily used in other datasets [17]. To address this, scholars proposed AutoAugment [130], which can automatically learn a data augmentation strategy for neural network. Inspired by this, the fault data samples augmented through AutoAugment may provide a good solution for S&I-IFD. In addition to the data augmentation methods mentioned above, some researchers used semi-supervised learning-based models to select data samples with target labels from a large unlabeled dataset to expand the target dataset directly [131]. In engineering scenarios, the unlabeled monitoring datasets are easier to collect and usually have a larger size than the labeled datasets. Therefore, the use of unlabeled datasets is also helpful to expand the limited target datasets and improve the performance of diagnosis models.

Second, how to establish the samples quality evaluation indexes is also an important issue. In [12] and [52], researchers used the Pearson correlation coefficient to evaluate the similarity of the generated data and the real data. However, the excessive similarity of the generated and the real data will lead to information redundancy, which has a very limited improvement on the generalization ability of the diagnosis models. Therefore, it is not appropriate to evaluate the generated samples' quality only from similarity. From the aspect of data augmentation, it is also significant to establish a relatively objective and reliable evaluation index for the generated samples to improve the diagnosis performance of S&I-IFD.

6.2. How to prevent transfer learning-based approaches from negative transfer in S&I-IFD?

Among the three strategies, transfer learning-based approaches account for a large proportion, so it is an important theory for S&I-IFD. However, when negative transfer occurs, the transfer learning-based models will perform poorly in the case of lacking data samples. Negative transfer refers to the case that the knowledge extracted in the source domain harms the target task [86]. Negative transfer will occur if the distribution discrepancies of the target and the source domain data are too big. For example, when the source domain data is the bearings faulty samples while the target data is the gears faulty samples, the knowledge learned in the bearings faulty samples is meaningless or even has a negative impact to the gears fault diagnosis. In addition, the transferable components between the two domains are the foundation of transfer learning, like data samples, data features, or model parameters. In some cases, although the data distributions in the two domains are similar, negative transfer may also occur when the diagnosis model fails to find the components that can be transferred. For example, the physical structures of motors and generators are similar and their fault data distributions are also similar. However, if the transfer learning-based diagnosis models cannot find the components that can be transferred, the diagnosis knowledge learned from the motor fault data is useless for the generator fault diagnosis.

It is a big challenge for S&I-IFD to avoid the negative transfer. First, to describe the discrepancies of the data distributions in the target and the source domain, reasonable measurement rules need to be developed. In existing researches, most researchers rely on engineering experience to judge the similarity of the data distributions in the two domains, however, it lacks a unified and effective standard. Therefore, developing a distribution similarity metric is worth exploring in future research. Second, to build effective diagnosis models, the idea of transitive transfer learning is worth trying [132,133]. Different from traditional transfer learning methods, which involve only two domains, transitive transfer learning connects multiple related domains and updates the learned knowledge in a transitive manner, which may provide a feasible idea for the construction of general transfer learning-based diagnosis models for S&I-IFD.

6.3. Meta-learning theory and its possible applications in S&I-IFD

Meta-learning, or learning to learn, is an outstanding and new machine learning theory. The purpose of meta-learning is to improve the learning level from data to tasks and enable algorithms to obtain transferable knowledge from multiple tasks [134]. By training on various related tasks with few data, knowledge can be accumulated over several training episodes and used to the new but related task without fine-tuning [135], which makes meta-learning-based methods suitable for dealing with small sample problems.

Generally speaking, meta-learning has three categories: optimization-based methods, model-based methods, and metric-based methods [136]. Among them, optimization-based models aim at the learning of the meta-knowledge, which is the initialization parameters of the network, and then iterate them with a few training samples to get good classifiers. Model-Agnostic Meta-Learning (MAML) [137] is the most famous meta-learning method based on optimization. Model-based methods are good at data-efficient few-shot learning [138]. They can embed the current training dataset into the activation condition and predict the test data based on this condition. Recurrent neural network [139], convolutional neural network [140], and hyper-network [141] are the typical architectures of the model-based meta-learning. Finally, metric-based methods are trained by comparing the training datasets with the validation datasets. Siamese network [142], matching network [143], prototypical network [144], and relation network [145] are typical meta-learning models based on metric. On the whole, meta-learning-based models have two obvious characteristics. The first is that meta-learning-based models are trained through learning the task of "N-way K-shot", where the classes number is N and the training samples number in each class is K. Generally, K is small, which means meta-learning is suitable for the case of lacking fault samples in engineering scenarios. The second is that meta-learning-based models have strong generalization ability. Some models like matching network [146] can perform well in the classification task even containing new class data that have not been trained in the training stage, which means meta-learning is good at dealing with actual problems in engineering scenarios.

It is worth noting that some scholars have tried to apply meta-learning theory to solve the S&I-IFD problem and some preliminary results have been achieved. For example, Chang et al. [20] presented a faults identification scheme for bearings in satellite communication antenna, in which a meta-learning module based on relation network was applied to measure the correlation degree of vibration data so as to realize bearings faults identification using small sample. In [125], a meta-learning framework based on the meta-relation net was presented for machine fault diagnosis. The experimental results showed that this meta-relation net-based model was suitable for fault classification with a few training samples.

At present, intelligent diagnosis models using meta-learning theory have not been deeply developed. The existing research results are mainly based on relation network to build diagnosis models, however, Siamese network, matching network, and prototypical network have not been applied yet. In addition to metric-based approaches, optimization-based and model-based approaches can achieve good results in image classification in the small sample case [138]. How to use them to build intelligent diagnosis models is worthy of further exploration. Overall, meta-learning theory has great potential to solve the problem of S&I-IFD, so it is one of the important directions for future research.

6.4. Zero-shot learning theory and its possible applications in S&I-IFD

Zero-shot learning [147] may bring research breakthroughs in S&I-IFD. Zero-shot learning uses seen data, which has been collected in practice, for training and realizes the recognition of unseen data, which has not been collected. In engineering scenarios, most collected data are under normal conditions, fault data are rare. In extreme cases, researchers cannot obtain fault signals under a certain fault type or under a certain working condition, which means diagnosis models do not have training samples from unseen data classes. In intelligent fault diagnosis, the recognition of the unseen data classes is a quite hard task, which is difficult to accomplish using common diagnosis models. Zero-shot learning is a feasible way to recognize unseen data, which is a valuable direction for further research in S&I-IFD.

Zero-shot learning realizes the recognition of unseen classes by inferring from seen classes to unseen classes [148], which has been applied in image recognition widely. Zero-shot learning mainly includes model embedding [149] and feature generation [150], etc. Through training on seen classes, the model can learn the mapping relationship between the data features and their attributes, while the correlativity between the attributes and the data labels is predefined. Based on the learned mapping relationships between the features and the attributes, the model could infer the attributes of unseen classes in the testing stage and realize the recognition of unseen classes through the correlativity between the attributes and the data labels.

In intelligent fault diagnosis, scholars have begun the preliminary research on the zero-shot data classification. A zero-shot diagnosis model using contractive auto-encoder was presented in [151] to identify machine faults without faulty samples. Feng et al. [152] used a faults description model based on the attribute transfer strategy for the zero-sample fault classification of complex mechanical systems. Lv et al. [153] proposed a conditional adversarial de-noising auto-encoder for machine fault identification without fault data, which generated unseen classes with the hybrid attribute as conditions.

At present, the research on intelligent diagnosis using zero-shot learning theory has obtained preliminary achievements in the perspective of data attributes description and data features generation. In machine faults identification, the attributes of machine monitoring data are related to the monitoring object and the data type. For example, due to the difference of fault form and fault mechanism, the attributes of induction motor monitoring data are different from that of generator monitoring data. Moreover, for some complex equipment, such as aero-engines, their monitoring data include pressure data, temperature data, flow data, vibration data, and so on. These different types of monitoring data have different data attributes. Therefore, how to effectively describe data attributes according to different monitoring objects and data types is one of the key research directions in the future, which is of great value to the application of zero-shot learning-based diagnosis models. In addition, how to learn and generate general data features is an important basis for zero-shot learning. These existing research results are mainly based on auto-encoder to generate data features in unseen classes [151, 153]. In the future, how to use other models such as GAN [154] to achieve feature learning and generation is a necessary research direction. Generally speaking, zero-shot learning theory has a strong application value for machine fault diagnosis under small sample conditions. Although there have been some preliminary research results, we think it still has a broad research space. Therefore, how to design effective diagnosis models based on zero-shot learning is an important research direction for S&I-IFD in the future.

7. Conclusions

S&I-IFD has attracted the attention of scholars for a long time. In this paper, we review the research achievements on S&I-IFD, which can be classified into three categories: data augmentation-based strategy, feature learning-based strategy, and classifier design-based strategy. Specifically, data augmentation-based strategy improves the diagnosis performance on small &imbalanced data by generating, over-sampling, or reweighting the training data samples. Feature learning-based strategy learns the fault features from small &imbalanced data using regularized neural networks or feature adaptation. Classifier design-based strategy achieves high diagnosis accuracy by designing the fault classifier suitable for small &imbalanced data classification.

For future research, how to enhance the augmented samples' quality is a problem that needs to be paid more attention to. Besides, how to prevent transfer learning-based diagnosis schemes from the negative transfer is a challenge for further applications in engineering scenarios. Finally, meta-learning theory and zero-shot learning theory have great potential in dealing with the S&I-IFD problem, which may bring research breakthroughs in the future.

Declaration of competing interest

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

Acknowledgments

The authors would like to sincerely thank all the anonymous reviewers for the valuable comments that greatly helped to improve the manuscript.

This research is supported financially by the National Natural Science Foundation of China (No. 91960106, No. 51875436, No. U1933101, No. 61633001, No. 51421004, No. 51965013), China Postdoctoral Science Foundation (No. 2020T130509, No. 2018M631145) and Shaanxi Natural Science Foundation, China (No. 2019JM-041).

References

[1] Pan J, Zi Y, Chen J, Zhou Z, Wang B. LiftingNet: A Novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification. IEEE Trans Ind Electron 2018;65:4973–82. http://dx.doi.org/10.1109/TIE.2017.2767540.

[2] Jiang W, Zhou J, Liu H, Shan Y. A multi-step progressive fault diagnosis method for rolling element bearing based on energy entropy theory and hybrid ensemble auto-encoder. ISA Trans 2019;87:235–50. http://dx.doi.org/10.1016/j.isatra.2018.11.044.

[3] Xiang Z, Zhang X, Zhang W, Xia X. Fault diagnosis of rolling bearing under fluctuating speed and variable load based on TCO spectrum and stacking auto-encoder. Meas J Int Meas Confed 2019;138:162–74. http://dx.doi.org/10.1016/j.measurement.2019.01.063.

[4] Zhang K, Chen J, Zhang T, Zhou Z. A compact convolutional neural network augmented with multiscale feature extraction of acquired monitoring data for mechanical intelligent fault diagnosis. J Manuf Syst 2020. http://dx.doi.org/10.1016/j.jmsy.2020.04.016.

[5] Chang Y, Chen J, Qu C, Pan T. Intelligent fault diagnosis of wind turbines via a deep learning network using parallel convolution layers with multi-scale kernels. Renew Energy 2020. http://dx.doi.org/10.1016/j.renene.2020.02.004.

[6] Pan T, Chen J, Zhou Z, Wang C, He S. A novel deep learning network via multiscale inner product with locally connected feature extraction for intelligent fault detection. IEEE Trans Ind Informatics 2019. http://dx.doi.org/10.1109/tii.2019.2896665.

[7] Pan T, Chen J, Pan J, Zhou Z. A deep learning network via shunt-wound restricted Boltzmann machines using raw data for fault detection. IEEE Trans Instrum Meas 2020. http://dx.doi.org/10.1109/TIM.2019.2953436.

[8] Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK. Applications of machine learning to machine fault diagnosis: A review and roadmap. Mech Syst Signal Process 2020;138:106587. http://dx.doi.org/10.1016/j.ymssp.2019.106587.

[9] Hashmi MB, Majid MAA, Lemma TA. Combined effect of inlet air cooling and fouling on performance of variable geometry industrial gas turbines. Alexandria Eng J 2020. http://dx.doi.org/10.1016/j.aej.2020.04.050.

[10] Martin-Diaz I, Morinigo-Sotelo D, Duque-Perez O, De Romero-Troncoso RJ. Early fault detection in induction motors using adaboost with imbalanced small data and optimized sampling. IEEE Trans Ind Appl 2017;53:3066–75. http://dx.doi.org/10.1109/TIA.2016.2618756.

[11] Gao L, Ren Z, Tang W, Wang H, Chen P. Intelligent gearbox diagnosis methods based on SVM, wavelet lifting and RBR. Sensors 2010;10:4602–21. http://dx.doi.org/10.3390/s100504602.

[12] Zhang T, Chen J, Li F, Pan T. A small sample focused intelligent fault diagnosis scheme of machines via multi-modules learning with gradient penalized generative adversarial networks, vol. 0046. 2020, http://dx.doi.org/10.1109/TIE.2020.3028821.

[13] Xiao D, Huang Y, Qin C, Liu Z, Li Y, Liu C. Transfer learning with convolutional neural networks for small sample size problem in machinery fault diagnosis. Proc Inst Mech Eng Part C J Mech Eng Sci 2019;233:5131–43. http://dx.doi.org/10.1177/0954406219840381.

[14] Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX. Deep learning and its applications to machine health monitoring. Mech Syst Signal Process 2019. http://dx.doi.org/10.1016/j.ymssp.2018.05.050.

[15] Gangsar P, Tiwari R. Signal based condition monitoring techniques for fault detection and diagnosis of induction motors: A state-of-the-art review. Mech Syst Signal Process 2020. http://dx.doi.org/10.1016/j.ymssp.2020.106908.

[16] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. Expert Syst Appl 2017;73:220–39. http://dx.doi.org/10.1016/j.eswa.2016.12.035.

[17] Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: A survey on few-shot learning. ACM Comput Surv 2020;53. http://dx.doi.org/10.1145/3386252.

[18] Sun Y, Wong AKC, Kamel MS. Classification of imbalanced data: A review. Int J Pattern Recognit Artif Intell 2009;23:687–719. http://dx.doi.org/10.1142/S0218001409007326.

[19] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009. http://dx.doi.org/10.1109/TKDE.2008.239.

[20] Chang YH, Chen JL, He SL. Intelligent fault diagnosis of satellite communication antenna via a novel meta-learning network combining with attention mechanism. J Phys Conf Ser 2020;1510. http://dx.doi.org/10.1088/1742-6596/1510/1/012026.

[21] Pan T, Chen J, Qu C, Zhou Z. A method for mechanical fault recognition with unseen classes via unsupervised convolutional adversarial auto-encoder. Meas Sci Technol 2020. http://dx.doi.org/10.1088/1361-6501/abb38.

[22] Govindan K, Jepsen MB. ELECTRE: A comprehensive literature review on methodologies and applications. Eur J Oper Res 2016;250:1–29. http://dx.doi.org/10.1016/j.ejor.2015.07.019.

[23] Yang J, Xie G, Yang Y. An improved ensemble fusion autoencoder model for fault diagnosis from imbalanced and incomplete data. Control Eng Pract 2020;98. http://dx.doi.org/10.1016/j.conengprac.2020.104358.

[24] Zhao X, Jia M, Lin M. Deep Laplacian auto-encoder and its application into imbalanced fault diagnosis of rotating machinery. Meas J Int Meas Confed 2020;152:107320. http://dx.doi.org/10.1016/j.measurement.2019.107320.

[25] Dixit S, Verma NK. Intelligent condition based monitoring of rotary machines with few samples. IEEE Sens J 2020;1748:1. http://dx.doi.org/10.1109/jsen.2020.3008177.

[26] Liu J, Qu F, Hong X, Zhang H. A small-sample wind turbine fault detection. IEEE Trans Ind Informatics 2019;15:3877–88.

[27] Liu Q, Ma G, Cheng C. Data fusion generative adversarial network for multi-class imbalanced fault diagnosis of rotating machinery. IEEE Access 2020;8:70111–24. http://dx.doi.org/10.1109/ACCESS.2020.2986356.

[28] Hang Q, Yang J, Xing L. Diagnosis of rolling bearing based on classification for high dimensional unbalanced data. IEEE Access 2019;7:79159–72. http://dx.doi.org/10.1109/ACCESS.2019.2919406.

[29] Shen F, Chen C, Yan R, Gao RX. Bearing fault diagnosis based on SVD feature extraction and transfer learning classification. In: Proc. 2015 progn. syst. heal. manag. conf.. 2016, http://dx.doi.org/10.1109/PHM.2015.7380088.

[30] Zeng Y, Wu X, Chen J. Bearing fault diagnosis with denoising autoencoders in few labeled sample case. In: 2020 5th IEEE int conf big data anal.. 2020, p. 349–53. http://dx.doi.org/10.1109/ICBDA49040.2020.9101321.

[31] Saufi SR, Bin ZA, Leong MS, Lim MH. Gearbox fault diagnosis using a deep learning model with limited data sample. IEEE Trans Ind Informatics 2020;16:6263–71. http://dx.doi.org/10.1109/TII.2020.2967822.

[32] Li Q, Tang B, Deng L, Wu Y, Wang Y. Deep balanced domain adaptation neural networks for fault diagnosis of planetary gearboxes with limited labeled data. Meas J Int Meas Confed 2020;156:107570. http://dx.doi.org/10.1016/j.measurement.2020.107570.

[33] Yang B, Lei Y, Jia F, Xing S. A transfer learning method for intelligent fault diagnosis from laboratory machines to real-case machines. In: Proc. - 2018 int. conf. sensing, diagnostics, progn. control. 2019, http://dx.doi.org/10.1109/SDPC.2018.8664814.

[34] Li X, Zhang W, Ding Q, Sun JQ. Multi-layer domain adaptation method for rolling bearing fault diagnosis. Signal Process 2019. http://dx.doi.org/10.1016/j.sigpro.2018.12.005.

[35] Chen F, Tang B, Chen R. A novel fault diagnosis model for gearbox based on wavelet support vector machine with immune genetic algorithm. Meas J Int Meas Confed 2013;46:220–32. http://dx.doi.org/10.1016/j.measurement.2012.06.009.

[36] Deng S, Lin SY, Chang WL. Application of multiclass support vector machines for fault diagnosis of field air defense gun. Expert Syst Appl 2011;38:6007–13. http://dx.doi.org/10.1016/j.eswa.2010.11.020.

[37] Chen F, Tang B, Song T, Li L. Multi-fault diagnosis study on roller bearing based on multi-kernel support vector machine with chaotic particle swarm optimization. Meas J Int Meas Confed 2014;47:576–90. http://dx.doi.org/10.1016/j.measurement.2013.08.021.

[38] Dong X, Gao H, Guo L, Li K, Duan A. Deep cost adaptive convolutional network: A classification method for imbalanced mechanical data. IEEE Access 2020;8:71486–96. http://dx.doi.org/10.1109/ACCESS.2020.2986419.

[39] Zhang C, Tan KC, Li H, Hong GS. A cost-sensitive deep belief network for imbalanced classification. IEEE Trans Neural Networks Learn Syst 2019;30:109–22. http://dx.doi.org/10.1109/TNNLS.2018.2832648.

[40] Peng P, Zhang W, Zhang Y, Xu Y, Wang H, Zhang H. Cost sensitive active learning using bidirectional gated recurrent neural networks for imbalanced fault diagnosis. Neurocomputing 2020;407:232–45. http://dx.doi.org/10.1016/j.neucom.2020.04.075.

[41] Li X, Jiang H, Zhao K, Wang R. A deep transfer nonnegativity-constraint sparse autoencoder for rolling bearing fault diagnosis with few labeled data. IEEE Access 2019;7:91216–24. http://dx.doi.org/10.1109/ACCESS.2019.2926234.

[42] He Z, Shao H, Zhang X, Cheng J, Yang Y. Improved deep transfer autoencoder for fault diagnosis of gearbox under variable working conditions with small training samples. IEEE Access 2019;7:115368–77. http://dx.doi.org/10.1109/access.2019.2936243.

[43] Kim H, Youn BD. A new parameter repurposing method for parameter transfer with small dataset and its application in fault diagnosis of rolling element bearings. IEEE Access 2019;7:46917–30. http://dx.doi.org/10.1109/ACCESS.2019.2906273.

[44] Chen J, Chang Y, Qu C, Zhang M, Li F, Pan J. Intelligent impulse finder: A boosting multi-kernel learning network using raw data for mechanical fault identification in big data era. ISA Trans 2020. http://dx.doi.org/10.1016/j.isatra.2020.07.039.

[45] Yu Y, Tang B, Lin R, Han S, Tang T, Chen M. CWGAN: Conditional wasserstein generative adversarial nets for fault data generation. In: IEEE int conf robot biomimetics. 2019, p. 2713–8. http://dx.doi.org/10.1109/ROBIO49542.2019.8961501.

[46] Pan T, Chen J, Xie J, Zhou Z, He S. Deep feature generating network: A new method for intelligent fault detection of mechanical systems under class imbalance. IEEE Trans Ind Informatics 2020;3203:1. http://dx.doi.org/10.1109/tii.2020.3030967.

[47] Wu Z, Lin W, Fu B, Guo J, Ji Y, Pecht M. A local adaptive minority selection and oversampling method for class-imbalanced fault diagnostics in industrial systems. IEEE Trans Reliab 2019;1–12. http://dx.doi.org/10.1109/TR.2019.2942049.

[48] Zhang Y, Li X, Gao L, Wang L, Wen L. Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning. J Manuf Syst 2018;48:34–50. http://dx.doi.org/10.1016/j.jmsy.2018.04.005.

[49] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inf Process Syst 2014.

[50] Kingma DP, Welling M. Auto-encoding variational bayes. In: 2nd int. conf. learn. represent. ICLR 2014 - conf. track proc. 2014.

[51] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: An overview. IEEE Signal Process Mag 2018. http://dx.doi.org/10.1109/MSP.2017.2765202.

[52] Shao S, Wang P, Yan R. Generative adversarial networks for data augmentation in machine fault diagnosis. Comput Ind 2019;106:85–93. http://dx.doi.org/10.1016/j.compind.2019.01.001.

[53] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: 4th int. conf. learn. represent. ICLR 2016 - conf. track proc. 2016.

[54] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: 34th int. conf. mach. learn. 2017.

[55] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of wasserstein GANs. Adv Neural Inf Process Syst 2017.

[56] Mirza M, Osindero S. Conditional generative adversarial nets. 2014, p. 1–7.

[57] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans. In: 34th int. conf. mach. learn. 2017.

[58] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved Techniques for Training GANs. Adv Neural Inf Process Syst 2016.

[59] Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. Adv Neural Inf Process Syst 2016.

[60] Yin H, Li Z, Zuo J, Liu H, Yang K, Li F. Wasserstein generative adversarial network and convolutional neural network (WG-CNN) for bearing fault diagnosis. Math Probl Eng 2020;2020. http://dx.doi.org/10.1155/2020/2604191.

[61] Gao X, Deng F, Yue X. Data augmentation in fault diagnosis based on the wasserstein generative adversarial network with gradient penalty. Neurocomputing 2020;396:487–94. http://dx.doi.org/10.1016/j.neucom.2018.10.109.

[62] Zhang W, Li X, Jia XD, Ma H, Luo Z, Li X. Machinery fault diagnosis with imbalanced data using deep generative adversarial networks. Meas J Int Meas Confed 2020;152:152. http://dx.doi.org/10.1016/j.measurement.2019.107377.

[63] Zhang T, Chen J, Xie J, T. Pan. SASLN: Signals augmented self-taught learning networks for mechanical fault diagnosis under small sample condition, vol. 9456. 2020, http://dx.doi.org/10.1109/TIM.2020.3043098.

[64] Wu J, Zhao Z, Sun C, Yan R, Chen X. Ss-infogan for class-imbalance classification of bearing faults. Procedia Manuf 2020;49:99–104. http://dx.doi.org/10.1016/j.promfg.2020.07.003.

[65] Wang Z, Wang J, Wang Y. An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition. Neurocomputing 2018;310:213–22. http://dx.doi.org/10.1016/j.neucom.2018.05.024.

[66] Zou L, Li Y, Xu F. An adversarial denoising convolutional neural network for fault diagnosis of rotating machinery under noisy environment and limited sample size case. Neurocomputing 2020;407:105–20. http://dx.doi.org/10.1016/j.neucom.2020.04.074.

[67] Wang J, Li S, Han B, An Z, Bao H, Ji S. Generalization of deep neural networks for imbalanced fault classification of machinery using generative adversarial networks. IEEE Access 2019;7:111168–80. http://dx.doi.org/10.1109/access.2019.2924003.

[68] Ding Y, Ma L, Ma J, Wang C, Lu C. A generative adversarial network-based intelligent fault diagnosis method for rotating machinery under small sample size conditions. IEEE Access 2019;7:149736–49. http://dx.doi.org/10.1109/ACCESS.2019.2947194.

[69] Mao W, Liu Y, Ding L, Li Y. Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: A comparative study. IEEE Access 2019;7:9515–30. http://dx.doi.org/10.1109/ACCESS.2018.2890693.

[70] Ren Wang Y, Dong SunG, Jin Q. Imbalanced sample fault diagnosis of rotating machinery using conditional variational auto-encoder generative adversarial network. Appl Soft Comput J 2020;92:106333. http://dx.doi.org/10.1016/j.asoc.2020.106333.

[71] Zheng T, Song L, Guo B, Liang H, Guo L. An efficient method based on conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearing. In: 2019 progn syst heal manag conf PHM-Qingdao, vol. 2019. 2019, http://dx.doi.org/10.1109/PHM-Qingdao46334.2019.8942906.

[72] Zheng T, Song L, Wang J, Teng W, Xu X, Ma C. Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings. Meas J Int Meas Confed 2020;158:107741. http://dx.doi.org/10.1016/j.measurement.2020.107741.

[73] Li Z, Zheng T, Wang Y, Cao Z, Guo Z, Fu H. A novel method for imbalanced fault diagnosis of rotating machinery based on generative adversarial networks. IEEE Trans Instrum Meas 2020;9456:1. http://dx.doi.org/10.1109/tim.2020.3009343.

[74] Zhou F, Yang S, Fujita H, Chen D, Wen C. Deep learning fault diagnosis method based on global optimization GAN for unbalanced data. Knowl-Based Syst 2020;187:104837. http://dx.doi.org/10.1016/j.knosys.2019.07.008.

[75] Cabrera D, Sancho F, Long J, Sanchez RV, Zhang S, Cerrada M, et al. Generative adversarial networks selection approach for extremely imbalanced fault diagnosis of reciprocating machinery. IEEE Access 2019;7:70643–53. http://dx.doi.org/10.1109/ACCESS.2019.2917604.

[76] Liang P, Deng C, Wu J, Yang Z, Zhu J, Zhang Z. Single and simultaneous fault diagnosis of gearbox via a semi-supervised and high-accuracy adversarial learning framework. Knowl-Based Syst 2020;198:105895. http://dx.doi.org/10.1016/j.knosys.2020.105895.

[77] Pan T, Chen J, Xie J, Chang Y, Zhou Z. Intelligent fault identification for industrial automation system via multi-scale convolutional generative adversarial network with partially labeled samples. ISA Trans 2020. http://dx.doi.org/10.1016/j.isatra.2020.01.014.

[78] Zhao D, Liu S, Gu D, Sun X, Wang L, Wei Y, et al. Enhanced data-driven fault diagnosis for machines with small and unbalanced data based on variational auto-encoder. Meas Sci Technol 2019. http://dx.doi.org/10.1088/1361-6501/ab55f8.

[79] Larsen ABL, Sønderby SK, Larochelle H, Winther O. Autoencoding beyond pixels using a learned similarity metric. In: 33rd int conf mach learn ICML 2016, vol. 4; 2016, pp. 2341–2349.

[80] Huang H, Yu PS, Wang C. An introduction to image synthesis with generative adversarial nets. 2018, p. 1–17, ArXiv.

[81] Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Inf Sci (Ny) 2019;505:32–64. http://dx.doi.org/10.1016/j.ins.2019.07.070.

[82] Wu Z, Lin W, Ji Y. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. IEEE Access 2018;6:8394–402. http://dx.doi.org/10.1109/ACCESS.2018.2807121.

[83] Soltanzadeh P, Hashemzadeh M. RCSMOTE: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. Inf Sci (Ny) 2021;542:92–111. http://dx.doi.org/10.1016/j.ins.2020.07.014.

[84] Wu Z, Jiang H, Lu T, Zhao K. A deep transfer maximum classifier discrepancy method for rolling bearing fault diagnosis under few labeled data. Knowl-Based Syst 2020;196:105814. http://dx.doi.org/10.1016/j.knosys.2020.105814.

[85] Hoang DT, Kang HJ. A bearing fault diagnosis method using transfer learning and Dempster-Shafer evidence theory. In: ACM int conf proceeding ser. 2019, p. 33–8. http://dx.doi.org/10.1145/3388218.3388220.

[86] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010. http://dx.doi.org/10.1109/TKDE.2009.191.

[87] Dai W, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: ACM int. conf. proceeding ser. 2007, http://dx.doi.org/10.1145/1273496.1273521.

[88] Zhang A, Li S, Cui Y, Yang W, Dong R, Hu J. Limited data rolling bearing fault diagnosis with few-shot learning. IEEE Access 2019;7:110895–904. http://dx.doi.org/10.1109/access.2019.2934233.

[89] Hu Y, Gao J, Zhou Q, Fan Z. Bearing fault diagnosis based on deep semisupervised small sample classifier. In: 2019 progn syst heal manag conf PHM-Qingdao 2019. 2019, http://dx.doi.org/10.1109/PHM-Qingdao46334.2019.8943025.

[90] T Wang, J Wang, Y Wu, X Sheng. A fault diagnosis model based on weighted extension neural network for turbo-generator sets on small samples with noise. Chinese J Aeronaut 2020. http://dx.doi.org/10.1016/j.cja.2020.06.024.

[91] Dong L, S LIU, H ZHANG. A method of anomaly detection and fault diagnosis with online adaptive learning under small training samples. Pattern Recognit 2017;64:374–85. http://dx.doi.org/10.1016/j.patcog.2016.11.026.

[92] Jiao J, Zhao M, Lin J, Liang K. A comprehensive review on convolutional neural network in machine fault diagnosis. Neurocomputing 2020. http://dx.doi.org/10.1016/j.neucom.2020.07.088.

[93] Zhao B, Zhang X, Li H, Yang Z. Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions. Knowl-Based Syst 2020;199:105971. http://dx.doi.org/10.1016/j.knosys.2020.105971.

[94] Jia F, Lei Y, Lu N, Xing S. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. Mech Syst Signal Process 2018;110:349–67. http://dx.doi.org/10.1016/j.ymssp.2018.03.025.

[95] Ren Z, Zhu Y, Yan K, Chen K, Kang W, Yue Y, et al. A novel model with the ability of few-shot learning and quick updating for intelligent fault diagnosis. Mech Syst Signal Process 2020;138. http://dx.doi.org/10.1016/j.ymssp.2019.106608.

[96] Jia F, Li S, Zuo H, Shen J. Deep neural network ensemble for the intelligent fault diagnosis of machines under imbalanced data. IEEE Access 2020;8:120974–82. http://dx.doi.org/10.1109/ACCESS.2020.3006895.

[97] Xu K, Li S, Jiang X, An Z, Wang J, Yu T. A renewable fusion fault diagnosis network for the variable speed conditions under unbalanced samples. Neurocomputing 2020;379:12–29. http://dx.doi.org/10.1016/j.neucom.2019.08.099.

[98] Geng Y, Wang Z, Jia L, Qin Y, Chen X. Bogie fault diagnosis under variable operating conditions based on fast kurtogram and deep residual learning towards imbalanced data. Meas J Int Meas Confed 2020;166. http://dx.doi.org/10.1016/j.measurement.2020.108191.

[99] Liu S, Sun Y, Zhang L. A novel fault diagnosis method based on noise-assisted MEMD and functional neural fuzzy network for rolling element bearings. IEEE Access 2018;6:27048–68. http://dx.doi.org/10.1109/ACCESS.2018.2833851.

[100] Qian W, Li S. A novel class imbalance-robust network for bearing fault diagnosis utilizing raw vibration signals. Meas J Int Meas Confed 2020;156:107567. http://dx.doi.org/10.1016/j.measurement.2020.107567.

[101] Pan SJ, Tsang IW, Kwok JT, Yang Q. Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 2011. http://dx.doi.org/10.1109/TNN.2010.2091281.

[102] Long M, Wang J, Ding G, Sun J, Yu PS. Transfer feature learning with joint distribution adaptation. In: Proc. IEEE int. conf. comput. vis. 2013, http://dx.doi.org/10.1109/ICCV.2013.274.

[103] Han T, Liu C, Yang W, Jiang D. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. Knowl-Based Syst 2019;165:474–87. http://dx.doi.org/10.1016/j.knosys.2018.12.019.

[104] Chen C, Li Z, Yang J, Liang B. A cross domain feature extraction method based on transfer component analysis for rolling bearing fault diagnosis. In: Proc. 29th Chinese control decis. conf.. 2017, http://dx.doi.org/10.1109/CCDC.2017.7978168.

[105] Xie J, Zhang L, Duan L, Wang J. On cross-domain feature fusion in gearbox fault diagnosis under various operating conditions based on transfer component analysis. In: 2016 IEEE int. conf. progn. heal. manag. 2016, http://dx.doi.org/10.1109/ICPHM.2016.7542845.

[106] Duan L, Xie J, Wang K, Wang J. Gearbox diagnosis based on auxiliary monitoring datasets of different working conditions. Zhendong Yu Chongji/J Vib Shock 2017. http://dx.doi.org/10.13465/j.cnki.jvs.2017.10.017.

[107] Han T, Liu C, Yang W, Jiang D. Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application. ISA Trans 2020. http://dx.doi.org/10.1016/j.isatra.2019.08.012.

[108] Qian W, Li S, Yi P, Zhang K. A novel transfer learning method for robust fault diagnosis of rotating machines under variable working conditions. Meas J Int Meas Confed 2019. http://dx.doi.org/10.1016/j.measurement.2019.02.073.

[109] Qian W, Li S, Jiang X. Deep transfer network for rotating machine fault analysis. Pattern Recognit 2019;96. http://dx.doi.org/10.1016/j.patcog.2019.106993.

[110] Zhang Z, Chen H, Li S, An Z. Unsupervised domain adaptation via enhanced transfer joint matching for bearing fault diagnosis. Meas J Int Meas Confed 2020;165:108071. http://dx.doi.org/10.1016/j.measurement.2020.108071.

[111] Cheng C, Zhou B, Ma G, Wu D, Yuan Y. Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data. Neurocomputing 2020;409:35–45. http://dx.doi.org/10.1016/j.neucom.2020.05.040.

[112] He Z, Shao H, Wang P, Lin J, Cheng J, Yang Y. Deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearbox with few target training samples. Knowl-Based Syst 2020;191:105313. http://dx.doi.org/10.1016/j.knosys.2019.105313.

[113] Zhang R, Liu Y. Research on development and application of support vector machine - Transformer fault diagnosis. In: ACM Int Conf Proceeding Ser. 2018, p. 262–8. http://dx.doi.org/10.1145/3305275.3305328.

[114] Chen G, Ge Z. SVM-tree and SVM-forest algorithms for imbalanced fault classification in industrial processes. IFAC J Syst Control 2019;8:100052. http://dx.doi.org/10.1016/j.ifacsc.2019.100052.

[115] Wagner C, Saalmann P, Hellingrath B. Machine condition monitoring and fault diagnostics with imbalanced data sets based on the KDD process. IFAC-PapersOnLine 2016;49:296–301. http://dx.doi.org/10.1016/j.ifacol.2016.11.151.

[116] Xi PP, Zhao YP, Wang PX, Li ZQ, Pan YT, Song FQ. Least squares support vector machine for class imbalance learning and their applications to fault detection of aircraft engine. Aerosp Sci Technol 2019;84:56–74. http://dx.doi.org/10.1016/j.ast.2018.08.042.

[117] Malik H, Mishra S. Proximal support vector machine (PSVM) based imbalance fault diagnosis of wind turbine using generator current signals. Energy Procedia 2016;90:593–603. http://dx.doi.org/10.1016/j.egypro.2016.11.228.

[118] He Z, Shao H, Cheng J, Zhao X, Yang Y. Support tensor machine with dynamic penalty factors and its application to the fault diagnosis of rotating machinery with unbalanced data. Mech Syst Signal Process 2020;141:106441. http://dx.doi.org/10.1016/j.ymssp.2019.106441.

[119] Duan L, Xie M, Bai T, Wang J. A new support vector data description method for machinery fault diagnosis with unbalanced datasets. Expert Syst Appl 2016;64:239–46. http://dx.doi.org/10.1016/j.eswa.2016.07.039.

[120] Mathew J, Pang CK, Luo M, Leong WH. Classification of imbalanced data by oversampling in kernel space of support vector machines. IEEE Trans Neural Networks Learn Syst 2018;29:4065–76. http://dx.doi.org/10.1109/TNNLS.2017.2751612.

[121] Duan A, Guo L, Gao H, Wu X, Dong X. Deep focus parallel convolutional neural network for imbalanced classification of machinery fault diagnostics. IEEE Trans Instrum Meas 2020;9456:1. http://dx.doi.org/10.1109/tim.2020.2998233.

[122] Chaudhari S, Polatkan G, Ramanath R, Mithal V. An attentive survey of attention models. 2019, ArXiv.

[123] Li F, Chen J, Pan J, Pan T. Cross-domain learning in rotating machinery fault diagnosis under various operating conditions based on parameter transfer. Meas Sci Technol 2020. http://dx.doi.org/10.1088/1361-6501/ab6ade.

[124] Cao P, Zhang S, Tang J. Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning. IEEE Access 2018;6:26241–53. http://dx.doi.org/10.1109/ACCESS.2018.2837621.

[125] Wu J, Zhao Z, Sun C, Yan R, Chen X. Few-shot transfer learning for intelligent fault diagnosis of machine. Meas J Int Meas Confed 2020;166:108202. http://dx.doi.org/10.1016/j.measurement.2020.108202.

[126] Yu K, Lin TR, Ma H, Li X, Li X. A multi-stage semi-supervised learning approach for intelligent fault diagnosis of rolling bearing using data augmentation and metric learning. Mech Syst Signal Process 2021;146. http://dx.doi.org/10.1016/j.ymssp.2020.107043.

[127] Li X, Zhang W, Ding Q, Sun JQ. Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation. J Intell Manuf 2020;31:433–52. http://dx.doi.org/10.1007/s10845-018-1456-1.

[128] Lv H, Chen J, Zhang T, Hou R, Pan T, Zhou Z. SDA: Regularization with cut-flip and mix-normal for machinery fault diagnosis under small dataset. ISA Trans 2020. http://dx.doi.org/10.1016/j.isatra.2020.11.005.

[129] Han S, Oh J, Jeong J. Bearing fault detection with data augmentation based on 2-d CNN and 1-d CNN. In: ACM int conf proceeding ser. 2020, p. 20–3. http://dx.doi.org/10.1145/3421537.3421546.

[130] Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. Autoaugment: Learning augmentation strategies from data. In: Proc. IEEE comput. soc. conf. comput. vis. pattern recognit. 2019, http://dx.doi.org/10.1109/CVPR.2019.00020.

[131] Tao X, Ren C, Li Q, Guo W, Liu R, He Q, et al. Bearing defect diagnosis based on semi-supervised kernel local Fisher discriminant analysis using pseudo labels. ISA Trans 2020. http://dx.doi.org/10.1016/j.isatra.2020.10.033.

[132] B Tan, Y Song, E Zhong, Q. Yang. Transitive transfer learning. In: Proc. ACM SIGKDD int. conf. knowl. discov. data min. 2015, http://dx.doi.org/10.1145/2783258.2783295.

[133] Tan B, Zhang Y, Pan SJ, Yang Q. Distant domain transfer learning. In: 31st AAAI conf. artif. intell. 2017.

[134] Mai S, Hu H, Xu J. Attentive matching network for few-shot learning. Comput Vis Image Underst 2019;187:102781. http://dx.doi.org/10.1016/j.cviu.2019.07.001.

[135] Ali AR, Gabrys B, Budka M. Cross-domain meta-learning for time-series forecasting. Procedia Comput Sci 2018;126:9–18. http://dx.doi.org/10.1016/j.procS.2018.07.204.

[136] Lee Y, Choi S. Gradient-based meta-learning with learned layerwise metric and subspace. In: 35th int. conf. mach. learn. 2018.

[137] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: 34th int. conf. mach. learn. 2017.

[138] Hospedales T, Antoniou A, Micaelli P, Storkey A. Meta-learning in neural networks: A survey. 2020, p. 1–20, ArXiv.

[139] Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: Proc. 5th int. conf. learn. represent. 2017.

[140] Mishra N, Rohaninejad M, Chen X, Abbeel P. A simple neural attentive meta-learner. In: 6th int. conf. learn. represent. ICLR 2018 - Conf. Track Proc. 2018.

[141] Qiao S, Liu C, Shen W, Yuille A. Few-shot image recognition by predicting parameters from activations. In: Proc. IEEE comput. soc. conf. comput. vis. pattern recognit. 2018, http://dx.doi.org/10.1109/CVPR.2018.00755.

[142] van der Spoel E, Rozing MP, Houwing-Duistermaat JJ, Eline Slagboom P, Beekman M, de Craen AJM, et al. Siamese neural networks for one-shot image recognition. In: ICML - deep learn work 2015.

[143] Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. Adv. Neural Inf Process Syst 2016.

[144] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. Adv Neural Inf Process Syst 2017.

[145] Sung F, Yang Y, Zhang L, Xiang T, Torr PHS, Hospedales TM. Learning to compare: Relation network for few-shot learning. In: Proc. IEEE comput. soc. conf. comput. vis. pattern recognit. 2018, http://dx.doi.org/10.1109/CVPR.2018.00131.

[146] Zhang K, Chen J, Zhang T, He S, Pan T, Zhou Z. Intelligent fault diagnosis of mechanical equipment under varying working condition via iterative matching network augmented with selective signal reuse strategy. J Manuf Syst 2020;57:400–15. http://dx.doi.org/10.1016/j.jmsy.2020.10.007.

[147] Lampert CH, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE comput. soc. conf. comput. vis. pattern recognit. work. CVPR work. 2009. 2009, http://dx.doi.org/10.1109/CVPRW.2009.5206594.

[148] Romera-Paredes B, Torr PHS. An embarrassingly simple approach to zero-shot learning. In: 32nd int. conf. mach. learn. 2015.

[149] Changpinyo S, Chao WL, Sha F. Predicting visual exemplars of unseen classes for zero-shot learning. In: Proc. IEEE int. conf. comput. vis. 2017, http://dx.doi.org/10.1109/ICCV.2017.376.

[150] Xian Y, Lorenz T, Schiele B, Akata Z. Feature generating networks for zero-shot learning. In: Proc. IEEE comput. soc. conf. comput. vis. pattern recognit. 2018, http://dx.doi.org/10.1109/CVPR.2018.00581.

[151] Gao Y, Gao L, Li X, Zheng Y. A zero-shot learning method for fault diagnosis under unknown working loads. J Intell Manuf 2020. http://dx.doi.org/10.1007/s10845-019-01485-w.

[152] Feng L, Zhao C. Fault description based attribute transfer for zero-sample industrial fault diagnosis. IEEE Trans Ind Informatics 2020. http://dx.doi.org/10.1109/tii.2020.2988208.

[153] Lv H, Chen J, Pan T, Zhou Z. Hybrid attribute conditional adversarial denoising autoencoder for zero-shot classification of mechanical intelligent fault diagnosis. Appl Soft Comput J 2020. http://dx.doi.org/10.1016/j.asoc.2020.106577.

[154] Xian Y, Lorenz T, Schiele B, Akata Z. Feature generating networks for zero-shot learning. In: Proc. IEEE comput. soc. conf. comput. vis. pattern recognit. 2018, http://dx.doi.org/10.1109/CVPR.2018.00581.