# A Perspective Survey on Deep Transfer Learning for Fault Diagnosis in Industrial Scenarios: Theories, Applications and Challenges

Weihua Li [a, b], Ruyi Huang [a, c], Jipu Li [c], Yixiao Liao [c], Zhuyun Chen [c], Guolin He [c], Ruqiang Yan [d], Konstantinos Gryllias [e]

[a] Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 510640, China

[b] PazhouLab, Guangdong Artificial Intelligence and Digital Economy Laboratory, Guangzhou 510335, China

[c] School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510640, China

[d] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China

[e] Department of Mechanical Engineering, KU Leuven, Celestijnenlaan 300, Leuven 3001, Belgium

***Abstract***—Deep Transfer Learning (DTL) is a new paradigm of machine learning, which can not only leverage the advantages of Deep Learning (DL) in feature representation, but also benefit from the superiority of Transfer Learning (TL) in knowledge transfer. As a result, DTL techniques can make DL-based fault diagnosis methods more reliable, robust and applicable, and they have been widely developed and investigated in the field of Intelligent Fault Diagnosis (IFD). Although several systematic and valuable review articles have been published on the topic of IFD, they summarized relevant research only from an algorithm perspective and overlooked practical applications in industry scenarios. Furthermore, comprehensive review on DTL-based IFD methods is still lacking. From this insight, it is particularly important and more necessary to comprehensively survey the relevant publications of DTL-based IFD with the goal of helping readers to conveniently understand the current state-of-the-art techniques and to quickly design an effective solution for solving IFD problems in practice. First, theoretical backgrounds of DTL are briefly introduced to present how the transfer learning techniques can be integrated with deep learning models. Then, major applications of DTL and their recent developments in the field of IFD are detailed and discussed. More importantly, suggestions on how to select DTL algorithms for IFD in practical applications, and some future challenges and research trends are shared. Finally, conclusions of this survey are given. As last, we have reason to believe that the works done in this article can provide convenience and inspiration for the researchers who want to devote his/her efforts in the progress and advance of IFD.

***Keywords***—Fault Diagnosis, Deep Learning, Transfer Learning, Domain Adaptation, Deep Transfer Learning

**Abbreviations (Abbr.)**

| **Abbr.** | Terminology | **Abbr.** | Terminology |
|---|---|---|---|
| **Adagrad** | Adaptive Gradient | **IMS** | Intelligent Maintenance Systems |
| **Adam** | Adaptive Moment Estimation | **JDA** | Joint Distribution Adaptation |
| **ADDA** | Adversarial Discriminative Domain Adaptation | ***k*-NN** | k-Nearest Neighbors |
| **AI** | Artificial Intelligence | **KL** | Kullback-Leibler |
| **ANN** | Artificial Neural Network | **LSTM** | Long short-term memory |
| **CNs** | Capsule Networks | **MMD** | Maximum Mean Discrepancy |
| **CWRU** | Case Western Reserve University | **MVD** | Maximum Variance Discrepancy |
| **CMD** | Central Moment Discrepancy | **MAE** | Mean Absolute Error |
| **CMS** | Condition Monitoring System | **OSDA** | Open Set Domain Adaptation |
| **CMMD** | Conditional Maximum Mean Discrepancy | **PSO** | Particle Swarm Optimization |
| **CNNs** | Convolutional Neural Networks | **PK-MMD** | polynomial kernel induced MMD |
| **CORAL** | Correlation Alignment | **PHM** | Prognostics and Health Management |
| **DACN** | Deep Adversarial Capsule Network | **RNNs** | Recurrent Neural Networks |
| **DADAN** | Deep Adversarial Domain Adaptation Network | **RKHS** | Reproducing Kernel Hilbert Space |
| **DBNs** | Deep Belief Networks | **RMSE** | Root Mean Square Error |
| **DBM** | Deep Boltzmann Machines | **SAN** | Selective Adversarial Network |
| **DDCNN** | Deep Decoupling Convolutional Neural Network | **SAE** | Sparse Auto-Encoder |
| **DL** | Deep Learning | **SGD** | Stochastic Gradient Descent |
| **DNNs** | Deep Neural Networks | **SVM** | Support Vector Machine |
| **DTL** | Deep Transfer Learning | **TrAdaBoost** | Transfer Adaptive Boosting |
| **DANN** | Domain Adversarial Neural Network | **TCA** | Transfer Component Analysis |
| **GK-MMD** | Gaussian kernel induced MMD | **TL** | Transfer Learning |
| **GANs** | Generative Adversarial Networks | **TCNN** | Transferable Convolutional Neural Network |
| **GRL** | Gradient Reversal Layer | **WATN** | Weighted Adversarial Transfer Network |
| **GNNs** | Graph Neural Networks | **WCs** | Working Conditions |
| **IFD** | Intelligent Fault Diagnosis | **WD** | Wasserstein Distances |

## 1. Introduction

Powerfully driven by advanced computing, sensing, measuring and communicating technologies, the manufacturing industry is characterized by an irresistible trend from automatic to digital and to intelligent, and it has embraced the new era of the fourth industrial revolution (Industry 4.0), whose ultimate goal is to make precise self-perception, to enable autonomous decision-making, and to realize intelligent networking for machines during the process of manufacturing [1]-[3]. Industrial equipment (IE), one of the most crucial carriers for manufacturing industry in such trend and revolution, has been devoting itself to generating economic benefits such as quality improvement, efficiency enhancement, energy conservation and cost reduction. Meanwhile, the IE is typically asked to accomplish the herculean tasks that often have harsh operating environment and need providing long-term services [4]-[6]. To ensure the safety and reliability of the industrial environment, the health status of IE has to be monitored and diagnosed in time, which can reduce equipment downtime, formulate scheduled maintenance, increase economic benefits, and avoid tragic catastrophes [7], [8]. Because of the complexity and dynamicity associated with the manufacturing processes, which inevitably leads to degradation, failure and damage, how to precisely make fault diagnosis for IE in time was and remains a great challenge.

In past decades, more and more attention has been paid to timely and precise IFD from academics and industry researchers since it has been listed as a key concern by many governments and organizations. Fortunately, owing to the rapid development of Artificial Intelligence (AI) technologies, especially in deep learning and transfer learning, abundant intelligent algorithms have been developed by researchers and engineers to address various practical problems in industrial scenarios, and have also brought successful breakthroughs for intelligent fault diagnosis (IFD) of IE.

Deep learning, a branch of machine learning in AI, broadly refers to methods that utilize hierarchical architectures, such as Deep Neural Networks (DNNs), Deep Belief Networks (DBNs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) [9]-[13], to learn higher-level representations from raw inputs which are images of pixel data, files of audio data, documents of text data, etc., [14]. On one hand, deep learning technology has been proven to be a promising tool in many applications of manufacturing industry due to its phenomenal advantages in massive data processing, discriminative feature learning and effective pattern recognition, constructing intelligent models by mapping relationships between health conditions of IE and industrial data in an end-to-end way [15]-[18]. On the other hand, deep learning technology has limitations which inhibit its further progress, advance and application in complex real-world scenarios. The ideal and hypothetical application scenarios of deep learning present the following characteristics:

(1) Deep learning requires abundant labeled samples in advance for model training. One of the limitations of deep learning methods is that they learn how to perform tasks through observations. That is to say, deep learning methods heavily rely on large amounts of labeled training data, without which these methods are prone to overfitting and will lack a robust generalization performance.

(2) Deep learning has strict requirements for the distributions between training and testing data. If a deep learning model is trained on data that present distributions discrepancy with the target data, the performance of the model will decrease dramatically and even will not work.

Considering the practical applications in many industrial scenarios, however, it is time-consuming, labor-intensive and even unrealistic to collect sufficient labeled data, especially labeled fault data, because the IE is always kept in a good status with time- or condition-based schedule maintenance. More importantly, it is often the case that the IE operates at harsh, varying and complex environments, which makes the distributions of the data in future testing situations different from that of the data of the pretrained model.

Transfer learning, another branch of machine learning that focuses on learning common knowledge from one or more related but different application scenarios to help AI algorithms to obtain more powerful performance in an application scenario of interest, has been demonstrated as a promising methodology for helping deep learning to overcome the limitations mentioned above [19]. In analogy with the ability of human beings that can leverage only a few examples or previous experience to help tackle unforeseeable problems, transfer learning can endow an AI model with better learning performance even when training data is sparse and limited, and with robust generalization performance from the related but different application scenarios to a new one [20]. However, the traditional machine learning approaches might not be able to learn the discriminative representations in an effective way, which is a major roadblock for fulfilling the potential of transfer learning.

Combining the advantages of deep learning in feature representation and the benefits of transfer learning in knowledge transfer, Deep Transfer Learning (DTL), a new paradigm of machine learning developed in recent years, leverages deep learning technology for transfer learning, which can learn hidden transferable knowledge and capture complex patterns more effectively [21]. DTL would be better preferred in practical application scenarios for manufacturing industry because it can be easier integrated with deep learning models that are widely developed for IFD of IE and can make the deep-learning-based methods more reliable, robust and accessible [22], [23].

The goal of this survey is to offer an in-depth overview of DTL for fault diagnosis in industrial scenarios, which can provide a comprehensive guidance for the readers who want to devote his/her efforts in the progress and advance of IFD. Historically, several systematic review articles have been published on the topic of fault diagnosis. For instance, Jay Lee, the founding director of the National Science Foundation Industry/University Cooperative Research Center (NSF I/UCRC) for Intelligent Maintenance Systems (IMS), conducted a comprehensive overview for Prognostics and Health Management (PHM) of rotary machinery systems from designing PHM methodology to selecting appropriate algorithms and to making accurate diagnosis decision, in 2014 [24]. That literature review placed much emphasis on the traditional fault diagnosis and prognosis algorithms, which cannot reflect the state-of-the-art techniques at present. Chen *et al.* gave a broad comprehensive literature survey of AI algorithms in the fault diagnosis of rotating machinery from the aspect of theory and application, in 2018 [25], which mainly focuses in the following algorithms: k-Nearest Neighbors (k-NN), Naive Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN) and Deep Neural Network (DNN). Yan and Gao summarized the deep learning-based research work published before 2019 for machine health monitoring [26], in which the popular deep learning models, such as Sparse Auto-Encoder (SAE), Deep Belief Network (DBN), Deep Boltzmann Machines (DBM), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have been systematically reviewed and the corresponding data and codes have been opened for replicating the reported results. Lei and Nandi presented a review and roadmap for IFD methods based on machine learning [27], in which the development of IFD methods was divided into three periods: the past (traditional machine learning), present (deep learning) and future (transfer learning). Besides the survey articles mentioned above, there are some other related articles which focus on the deep learning-based [28], transfer learning-based [29], convolutional neural network-based [30], [31], AI-Enabled-based [32] or special components-based [33]-[35] methods for machine fault diagnosis, which will not be enumerated here in detail. Admittedly, as the valuable and systematic scholarly sources on the IFD, these literature reviews have contributed to the development of fault diagnosis from many aspects and guided the researchers towards a clearer future direction [36], [37]. Nevertheless, there are still some aspects that have not yet been comprehensively summarized by the previous literature articles:

(1) The historical reviews mainly concerned IFD on either traditional machine learning, deep learning or transfer learning. As a new promising tool to solve the problems faced by the engineer and researcher in the real industrial scenarios, there is still a lack of systematic review on DTL-based IFD methods.

(2) Throughout the above discussion, it is clear that almost all the reviews categorized the relevant research from the perspective of algorithm and analytic technology, resulting in the difficulty to select appropriate algorithms for engineers in specific industrial applications.

Therefore, it is particularly important and more necessary to overview the relevant publications of DTL with the goal of helping readers to conveniently understand the current state-of-the-art techniques related to IFD and to quickly design an effective solution for some challenges in practice.

To overcome the limitation forementioned, this review article attempts to provide a comprehensive survey on DTL for fault diagnosis in industrial scenarios. First, different from the existing review articles which mainly focused on the IFD methods using either traditional machine learning, deep learning or transfer learning, this review article aims at focusing on the IFD methods using the new paradigm of machine learning, i.e., DTL. Second, in contrast to the existing review articles summarized the related publications from the algorithm perspective, this review article categorizes the DTL-based IFD methods from the perspective of practical industrial scenarios, which could provide suggestions to select appropriate algorithms for engineers in specific industrial applications. Last, the existing review articles mainly covered the related publication before 2020. It is the fact that there are many new articles have been published in recent years since the IFD have attracted lots of attentions from both academics and industry researchers, but, by contrast, this review article has included most of the state-of-the-art DTL techniques before it summitted to the Journal.

The main contributions of this article are outlined as the following three aspects:

(1) Basic concepts and theories of DTL are introduced, including instance-based DTL, model-based DTL and feature-based DTL, which can present a comprehensive overview about the DTL from the algorithm perspective.

(2) Applications of DTL approaches are summarized into four categories from the perspective of practical industrial scenarios, and each category in IFD are detailed, which would be instructive for engineers in specific industrial applications to select the appropriate algorithms.

(3) Future challenges and potential directions of DTL for IFD are concluded, attempting to provide new insights on the future works for potential newcomers and seasoned researchers.

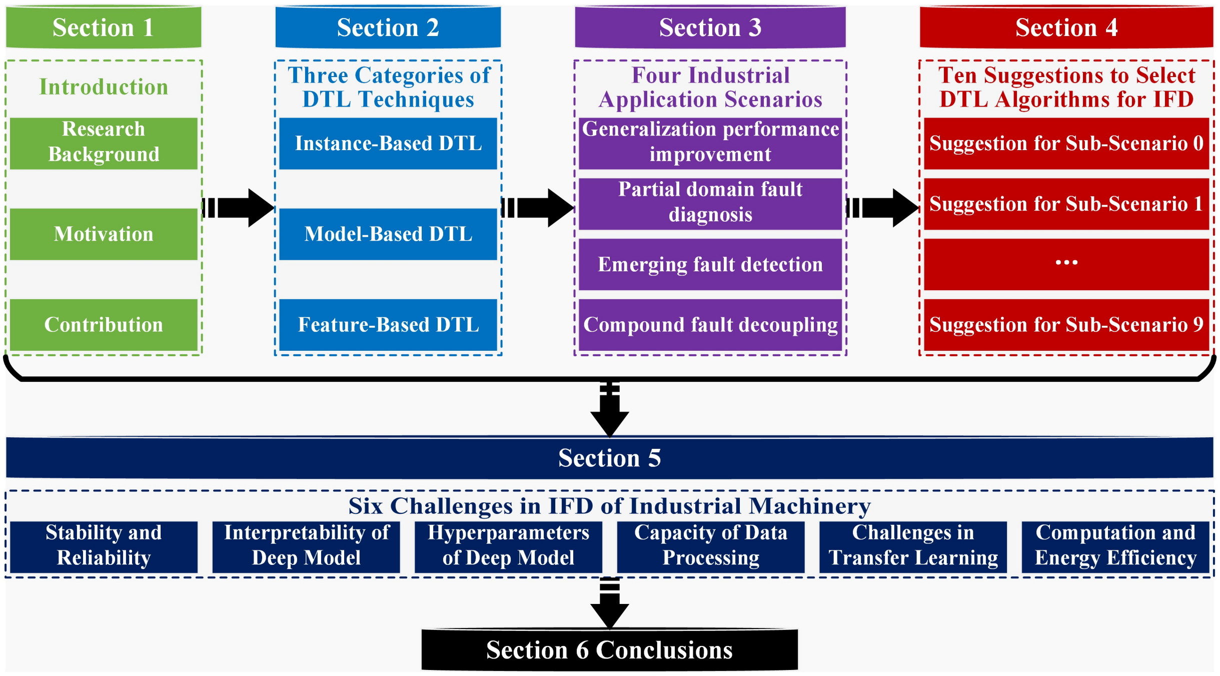

Fig. 1. Flow chart showing the overall logic of this literature review

As show in Fig. 1, the rest of this review article is organized as follows. In Section 2, the theoretical backgrounds of DTL, including basic definition of DTL and three categories of DTL, are briefly introduced to present how the transfer learning techniques can be integrated with deep learning models. Section 3 details the major applications of DTL and its recent developments in the field of IFD from the perspective of practical industrial scenarios, in which four application scenarios are formulated according to the task of fault classification. More importantly, suggestions that on how to select DTL algorithms for IFD in practical industrial applications and some future challenges are shared in Section 4 and 5, respectively. Finally, the conclusions of this survey are given in Section 6.

## 2. Theoretical Backgrounds of DTL

In this section, the basic definitions related to DTL are firstly introduced for convenience. The theories of DTL that explain how the transfer learning technologies can leverage the powerful representation ability of deep learning to extract and transfer knowledge, are summarized from the perspective of algorithm to allow readers understand the mechanisms and strategies of DTL approaches.

### *2.1 Basic Definitions of DTL*

According to the book of Transfer Learning [38], some basic definitions related to this survey, such as Domain, Task and Transfer Learning, are listed as follows:

***Domain***, denoted as $\mathcal{D}=\{\mathcal{X},P(X)\}$, consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X=\{x \mid x_i \in \mathcal{X}, i=1,\cdots,N\}$ is a dataset that contains $N$ instances. Generally, different domains are defined based on the fact that there are different feature spaces or different marginal probability distributions between these domains. In the scenarios of machinery fault diagnosis, different working conditions (WCs), locations and machines can be regarded as different domains.

***Task***, denoted as $\mathcal{T}=\{\mathcal{Y},f(\bullet)\}$ when giving a specific domain $\mathcal{D}$, consists of two components: a label space $\mathcal{Y}$ and a mapping function $f(\bullet)$, where $Y=\{y \mid y_i \in \mathcal{Y}, i=1,\cdots,N\}$ is a label set for the corresponding instances in $\mathcal{D}$. The mapping function $f(\bullet)$, also denoted as $f(\boldsymbol{x})=P(y \mid \boldsymbol{x})$, is a non-linear and implicit function that can bridge the relationship between the input instance and the predicted decision, which is expectedly learned from the given datasets. Similarly, different tasks are defined as there are different label spaces between these tasks. Different fault classes and types can be regarded as different tasks.

***Transfer Learning***, given a source domain $\mathcal{D}^S=\{\mathcal{X}^S,P^S(X^S)\}$ with the source task $\mathcal{T}^S=\{\mathcal{Y}^S,f^S(\bullet)\}$ and a target domain $\mathcal{D}^T=\{\mathcal{X}^T,P^T(X^T)\}$ with the target task $\mathcal{T}^T=\{\mathcal{Y}^T,f^T(\bullet)\}$, aims to learn a better mapping function $f^T(\bullet)$ for the target task $\mathcal{T}^T$ with the transferable knowledge gained from the source domain $\mathcal{D}^S$ and task $\mathcal{T}^S$. Contrary to the tradition machine learning and deep learning in which the domain and task between the source and target scenarios are identical (that is, $\mathcal{D}^S=\mathcal{D}^T$ and $\mathcal{T}^S=\mathcal{T}^T$), the transfer learning counters the problems where the domain and/or the task between the source and the target scenarios could be different (i.e., $\mathcal{D}^S \neq \mathcal{D}^T$ and/or $\mathcal{T}^S \neq \mathcal{T}^T$).

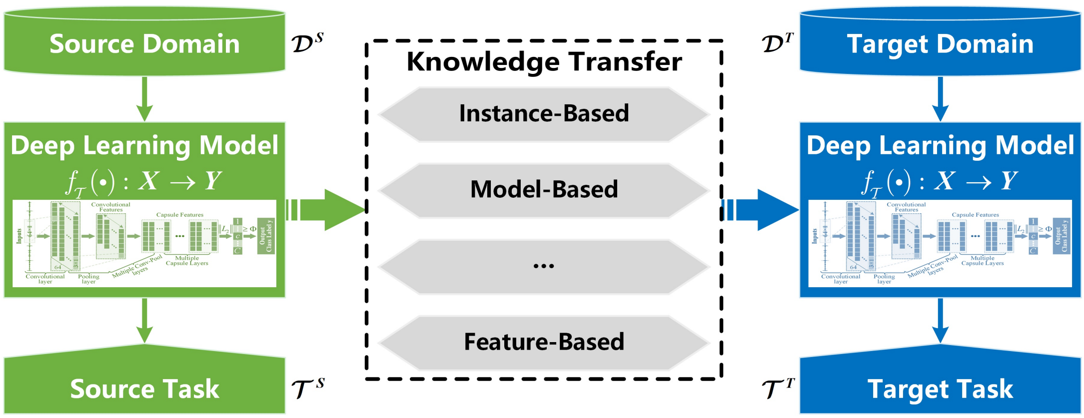


Fig. 2. An illustration of DTL

Based on the definition mentioned above, a definition of deep transfer learning can be formulated as: Given a transfer learning task $f^{S\to T}(\bullet): \boldsymbol{X}^T \to \boldsymbol{Y}^T$ based on [$\mathcal{D}^S$, $\mathcal{D}^T$, $\mathcal{T}^S$, $\mathcal{T}^T$], deep transfer learning aims to learn the mapping function $f^{S\to T}(\bullet)$ by leveraging the powerful deep learning model, that is, deep neural networks, in which the transfer learning technique and the deep learning model can be integrated to a more robust AI method.

### *2.2 Categorization of DTL*

Fig. 2 shows a typical concept of DTL process that is capable of transferring the valuable knowledge by further exploiting the representation learning ability of deep neural networks. The literature on deep learning or transfer learning has gone through a considerable number of iterative updates. In contrast, few literatures focus on deep transfer learning as a new emerging technique. There is no mutual consensus on how to classify the categorization of DTL. According to the survey published by Tan et al. [21], the DTL approaches have been divided into four categories, that is, instances-based, mapping-based, network-based and adversarial-based DTL. However, these types of DTL approaches are associated and inter-related with each other, which makes it difficult to be well-categorized.

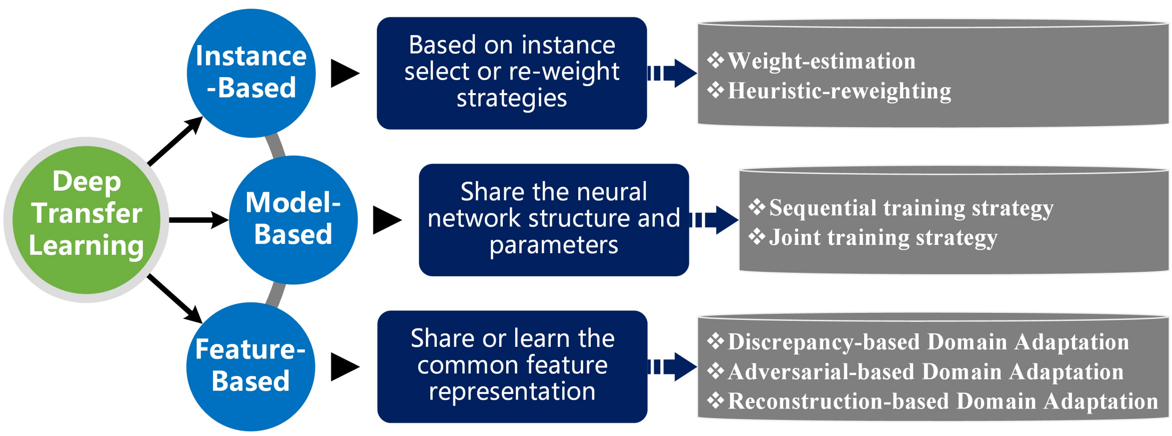


Fig. 3. Categorization of DTL

In this survey, from a viewpoint of the mechanism with which the deep learning model bridge the generalization errors between target and source domains by leveraging the transfer learning techniques, the DTL approaches are divided into three groups: instance-based, model-based and feature-based DTL. The categorization of DTL is illustrated in Fig. 3. Instance-based DTL approaches are typically based on instance select or re-weight strategies. Model-based DTL approaches mainly share the neural network structure and parameters between target and source domains. Feature-based DTL approaches share or learn the common feature representation between target and source domain. In the following parts of this section, the theorical backgrounds of each category of DTL will be introduced from the perspective of algorithm in detail.

### *2.2.1 Instance-based DTL*

Instance-Based DTL aims to train a more precise deep model under a transfer scenario where the difference between source and target domains/tasks only comes from either the different marginal probability distribution, i.e., $P^S(X^S) \neq P^T(X^T)$, or the conditional probability distribution, i.e., $P^S(Y^S \mid X^S) \neq P^T(Y^T \mid X^T)$, which also assumes that the labeled instances in the target domain are too limited to train a satisfied diagnosis model. An intuitive motivation behind instance-based DTL approaches is that directly merging the source data into the target data might deteriorate the performance of the target deep model and result in a negative transfer during the model training because some labeled instances in the source domain are significantly different from the target domain ones. Inspired by such motivation, the goal of the instance-based DTL approaches is to single out the instances in the source domain that are positive for target model training and to augment the target data by adapting the instance weighting strategies. A promising solution in terms of deep learning models is to automatically learn the instances weights of the source domain in the objective function. The general objective function of an instance-based DTL task can be formulated as

$$\mathcal{L} = \frac{1}{C^S} \sum_{i=1}^{N^S} \omega_i \mathcal{R}^S \left( f(\boldsymbol{x}_i^S), y_i^S \right) + \mathcal{R}^* \left( f(\boldsymbol{X}), \boldsymbol{Y} \right) \tag{1}$$

where $\omega_i$ is the weighting coefficient of the corresponding source instance, $C^S = \sum_{i=1}^{N^S} \omega_i$, $\mathcal{R}^S$ represents the risk function of selecting the source instance, and $\mathcal{R}^*$ denotes the second risk function related to the target task or the parameter regularization. The theoretical value of $\omega_i$ is defined as the ratio of the marginal probability distributions between the target domain and the source domain at the input instance $\boldsymbol{x}_i$

$$\omega_i = P^T(\boldsymbol{x}_i) / P^S(\boldsymbol{x}_i). \tag{2}$$

However, it is well known that such ratio is difficult to be directly computed with the marginal probability distribution. In this way, many effective methods have been developed to approximately estimate the aforementioned ratio bypassing the estimation step of the marginal probability distribution.

From the perspective of deep model training, the instance-based DTL can be further divided into the following two subcategories by considering whether the labeled instances are available in the target domain: the weight-estimation and the heuristic-reweighting method.

The weight-estimation method, which mainly focuses on the situation where there is a lack of labeled instances in target domain, convers the instance transfer problem into a weight estimation problem by leveraging the kernel embedding techniques. For instance, based on the theory of Maximum Mean Discrepancy (MMD) between distributions, the weights of source instances can be estimated by matching the means between the reweighted sources instances and the target instances in a Reproducing Kernel Hilbert Space (RKHS) [39], which can be obtained by optimizing the following objective

$$\begin{aligned} &\arg\min_{\boldsymbol{\omega}} \left\| \frac{1}{C^S} \sum_{i=1}^{N^S} \omega_i \Phi^S\left(\boldsymbol{x}_i^S\right) - \frac{1}{N^T} \sum_{j=1}^{N^T} \Phi^T\left(\boldsymbol{x}_j^T\right) \right\|_{\mathcal{H}}^2 \\ &\text{s.t. } \omega_i \geq 0, \ \left| \frac{1}{C^S} \sum_{i=1}^{N^S} \omega_i - 1 \right| \leq \varepsilon \end{aligned} \tag{3}$$

where $\varepsilon$ denotes a positive real number. There are some other tricks to estimate the weights by utilizing the Kullback-Leibler (KL) divergence [40]. With the weight of each source instance being estimated, Eq. 3 can be integrated into the objective function of the target task to learn a deep model. It is worth mentioning that the weight estimation of the source instances can be integrated into the training process of the deep model.

The heuristic-reweighting method, which is suitable for implementing the DTL task when some labeled instances are available in the target domain, aims to identify negative source instances by using instance reweighting strategies in a heuristic way. One of the most popular instance reweighting strategies is the Transfer Adaptive Boosting (TrAdaBoost) algorithm proposed by Dai *et al.* [41], in which the different weighting strategies are applied for the labeled instances in the source-domain and the target-domain to reduce the impact of negative source instances. Similar to the boosting-style algorithms, the weights of the source instances and the target instance can be updated through a lot of iterations, whose updating strategies are described as

$$\omega_i^S = \omega_i^S \left(1 + \sqrt{2 \ln N^S / (N^S + N^T)}\right)^{-\mathcal{R}^S\left(f(\boldsymbol{x}_i^S), y_i^S\right)} \tag{4}$$

$$\begin{cases} \omega_j^T = \omega_j^T \left(\varepsilon / (1 - \varepsilon)\right)^{-\mathcal{R}^T\left(f(\boldsymbol{x}_j^T), y_j^T\right)} \\ \varepsilon = \sum_{j=1}^{N^T} \omega_j^T \mathcal{R}^T\left(f(\boldsymbol{x}_j^T), y_j^T\right) \Big/ \sum_{j=1}^{N^T} \omega_j^T \end{cases} \tag{5}$$

where $i = [1, \cdots, N^S]$, $j = [1, \cdots, N^T]$, and $\varepsilon$ denotes the mean loss of all target domain instances. It should be highlighted here that each iteration will learn a new weak deep model, and therefore ensemble techniques are used to form a final classifier by integrating all weak deep models. Besides the TrAdaBoost algorithm and its variations, the other heuristic-reweighting methods make full use of not only the labeled instances in the source and the target domain but also the unlabeled instances in the target domain. An intuitive solution of these methods is to decompose the objective function into three parts:

$$\begin{aligned} \mathcal{L} = &\frac{1}{C^S}\sum_{i=1}^{N^S}\omega_i \mathcal{R}^S\left(f(\boldsymbol{x}_i^S), y_i^S\right) + \frac{1}{C^{T(L)}}\sum_{j=1}^{N^{T(L)}} \mathcal{R}^{T(L)}\left(f(\boldsymbol{x}_j^{T(L)}), y_j^{T(L)}\right) \\ &+ \frac{1}{C^{T(U)}}\sum_{k=1}^{N^{T(U)}} \gamma_k \mathcal{R}^{T(U)}\left(f(\boldsymbol{x}_k^{T(U)}), \hat{y}_k^{T(U)}\right) \end{aligned} \tag{6}$$

where the superscript of $S$, $T(L)$ and $T(U)$ denotes the labeled source, the labeled target and the unlabeled target, respectively. $C^S = \sum_{i=1}^{N^S}\omega_i$, $C^{T(L)} = N^{T(L)}$, $C^{T(U)} = \sum_{k=1}^{N^{T(U)}}\gamma_k$, $\gamma_k$ denotes the weight for the unlabeled target instance, and $\hat{y}_k^{T(U)} = P^T(y \mid \boldsymbol{x}_k^{T(U)})$ is the true conditional distributions of the unlabeled target instances. Generally, the optimal values of $\omega_i$, $\gamma_k$ and $\hat{y}_k^{T(U)}$ are unknown for computing these loss terms. Therefore, several techniques can be used during the deep model training to learn these parameters in a heuristic way. The typical procedure can be concluded as the following steps:

(1) An auxiliary classifier is firstly trained on the labeled target instances and then used to classify the labeled source and the unlabeled target instances to obtain the predicted probability of each instance.

(2) The labeled source and the unlabeled target instances are ranked based on its predicted probability, respectively.

(3) The $\omega_i$ of top n instances from the labeled source domain that are incorrectly predicted by the auxiliary classifier are set to zero, and the weights of others are set to one.

(4) The top n instances from the unlabeled target domain that have the highest prediction confidence are selected, for which the $\gamma_k$ is set to one and the $\hat{y}_k^{T(U)}$ is assigned to a pseudo label according to its predicted probability. Additionally, for all other instances from the unlabeled target domain, $\gamma_k$=0.

With the steps mentioned above, the whole loss can be calculated with the objective function presented in Eq. 6. Note that the selected labeled source and the unlabeled target instances can be used to train the auxiliary classifier again in the next iteration.

### *2.2.2 Model-based DTL*

Model-Based DTL focuses on the transfer assumption that the tasks between the source and the target domains share some common knowledge in the model level, which means that the transferable knowledge is well embedded into a pretrained source deep model whose structure and parameters are general and helpful for learning a powerful target model. The goal of model-based DTL approaches is to exploit which part of the deep learning model pretrained in the source domain can help improving the model learning process for the target domain. Model-based DTL algorithms are based on the assumption that some labeled instances in the target domain should be available during the target model training. According to the way of training of the target deep model, the model-based DTL can be further divided into two subcategories: sequential training and joint training.

Sequential training establishes the target deep model by pretraining a deep learning model on auxiliary domains which have much richer and larger labeled instances and then fine-tuning the well-trained source model on the target domain which often lacks sufficient labeled instances. Specifically, sequential training-based DTL approaches typically contains two stages. In the first stage, i.e., the pretraining on auxiliary domains, a well-trained source model $\mathcal{F}^S(\bullet;\theta^S)$ has been learned from the source data, which can be defined as

$$\mathcal{F}^S(\bullet;\theta^S) = \arg\min \mathcal{R}^S\left(f^S(\boldsymbol{X}^S;\theta^S),\boldsymbol{Y}^S\right) \tag{7}$$

where $\theta^S = \{\theta_i^S\}_{i=1}^{L^S}$ is the model parameter set of the pretrained source model, $L^S$ denotes the layer number of the source model, $\mathcal{R}^S$ denotes the risk function for the source task. In the second stage, that is, the fine-tuning on the target domain, the target deep model $\mathcal{F}^T(\bullet;\theta^T)$ can be obtained by freezing some components of the well-trained source model and fine-tuning the rest components with the target domain data, or by reusing all the parameters of the well-trained source model to initialize the target deep model and retraining the whole target model with the target domain data. The processes of this stage can be formulated as

$$\begin{aligned}&\mathcal{F}^T(\bullet;\theta^T) = \arg\min \mathcal{R}^T\left(f^T(\boldsymbol{X}^T;\theta^T),\boldsymbol{Y}^T\right)\\&\text{s.t. } \theta^T \text{ initialized/frozen with } \theta^*\end{aligned} \tag{8}$$

where $\theta^* = \{\theta_i^S, i \in [1,\cdots,L^S]\}$ is a subset of $\theta^S$ learned in the first stage, $\theta^T$ denotes the model parameter set expectedly learned in the second stage, $\mathcal{R}^T$ denotes the risk function for the target task. It is worth mentioning that the higher-level layers are prone to learn the task-specific representations and the lower-level layers are able to capture general representations in a deep learning model. Therefore, it is a classical fine-tuning strategy to freeze *n* lower-level layers learned from auxiliary domains and retrain the higher-level layer with limited target domain data.

Joint training tries to implement the source and the target tasks simultaneously. Different from the multi-task learning approaches which equally optimize the performance over all tasks, joint training-based DTL approaches focus on improving the performance of the target task by leveraging common knowledge from the source task. More specifically, there are two ways to joint training target task with source task. The first one is hard parameter sharing which shares the hidden layers directly while keeping the task-specific layers independently. The second one is soft parameter sharing which simply change the weight coefficient for the source and the target tasks or add regularization terms in the risk function. The processes of the soft parameter sharing can be defined as

$$\begin{aligned}\mathcal{F}^{T}(\bullet;\theta^{T}) = \arg\min[\alpha\mathcal{R}^{S}\left(f^{S}(\boldsymbol{X}^{S};\theta^{S}),\boldsymbol{Y}^{S}\right) + \beta\mathcal{R}^{T}\left(f^{T}(\boldsymbol{X}^{T};\theta^{T}),\boldsymbol{Y}^{T}\right) \\ + \gamma\mathcal{R}^{R}\left(f(\boldsymbol{X}^{S},\boldsymbol{X}^{T};\theta^{S},\theta^{T})\right)]\end{aligned} \tag{9}$$

where $\mathcal{R}^{S}$, $\mathcal{R}^{T}$ and $\mathcal{R}^{R}$ are the risk functions of the source task, the target task and regularization terms, respectively; and $\alpha$, $\beta$ and $\gamma$ are the weight coefficients for the corresponding task.

### *2.2.3 Feature-based DTL*

Feature-Based DTL endows deep models with the ability to transfer knowledge by learning the common representations in the feature space level, rather than in the instances or the model level, which further relaxes the assumption in the instance-based DTL transfer learning scenario to allow the differences of feature spaces to exist in the source and target domains. An intuitive solution behind feature-based DTL approaches is to learn the mapping function as a bridge to convert the raw data in source and target domains from the different feature spaces to a common latent feature space, where the difference between domains can be reduced and the deep feature representations that are discriminative for the main learning task and indiscriminate with respect to the shift between different domains can be obtained. With these good representations, the performance of deep models can be significantly improved in accomplishing the target task.

From a broader perspective, feature-based DTL approaches intuitively covers two transfer styles without or with adaptation to target domain. The approaches without adaptation firstly extract the lower-level representations by using a pretrained source model, and then directly take the extracted representations as inputs for the target model, which are suitable and effective only when the target domain is closely related to the source domain. The approaches with adaptation adapt the feature representations across different domains through domain adaptation techniques, which can obtain a well performing model even if there is a shift or gap between source and target domains. Since the approaches without adaptation are easily implemented and their assumption may be too strong to be satisfied in most practical transfer scenarios, the following part mainly focus on the approaches with adaptation.

A crucial problem of feature-based DTL with domain adaptation in learning domain invariant features is how to estimate and learn representation invariance between source and target domains. The ways of constructing representation invariance measures generally include three strategies: leveraging criterions based on the discrepancy to the reduce difference of distribution, adding domain discriminative architectures to encourage the domain confusion through the adversarial mechanism, and combining the data reconstruction as an auxiliary task to help improving representations invariance. Therefore, the feature-based DTL with domain adaptation approaches can be further summarized into the following three subcategories.

The first subcategory is discrepancy-based domain adaptation, which aims to align the feature distribution shift and to improve the ability of learning transferable representations by reducing the discrepancy based on distance metrics or criterions defined between corresponding-level representations of the given source and target domains. The criterions that are proven to be successful for discrepancy-based domain adaptation include MMD [39], KL divergence [40], multiple kernels MMD (MK-MMD) [42], Correlation Alignment (CORAL) [43] and Wasserstein Distances (WD, also known as Earth-Mover distance) [44], among others. Taking the most commonly one, MMD, as an example, and given two domain representations $\boldsymbol{h}^S$ (source) and $\boldsymbol{h}^T$ (target), the criterion based on MMD can be empirically estimated as follows:

$$MMD(\boldsymbol{h}^S,\boldsymbol{h}^T)=\left\|\frac{1}{N_S}\sum_{i=1}^{N_S}\varphi(\boldsymbol{h}_i^S)-\frac{1}{N_T}\sum_{j=1}^{N_T}\varphi(\boldsymbol{h}_j^T)\right\|_{\mathcal{H}}^2 \tag{10}$$

Another common criterion is the Wasserstein Distances. The criterion based on the WD can be expressed as:

$$WD(\boldsymbol{h}^S,\boldsymbol{h}^T)=\inf_{\gamma\in\Gamma(\boldsymbol{h}^S,\boldsymbol{h}^T)}\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma}\left[\|\boldsymbol{x}-\boldsymbol{y}\|\right]=\sup_{\|f\|_L\le 1}\left(\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{h}^S}\left[f(\boldsymbol{x})\right]-\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{h}^T}\left[f(\boldsymbol{x})\right]\right) \tag{11}$$

The more details about WD can be found in [44].

In the process of model training, the deep neural network can be optimized by minimizing the classification loss on the labeled instance, $\mathcal{R}_C(\boldsymbol{X}_L,\boldsymbol{Y}_L)$, while the domain invariant representations are measured by one/multiple adaptation layer(s) with such criterion. The objective function of discrepancy-based domain adaptation is formulated as

$$\mathcal{L}=\mathcal{R}_C(\boldsymbol{X}_L,\boldsymbol{Y}_L)+\sum_{i=1}^{L^A}\lambda_i MMD(\boldsymbol{h}_i^S,\boldsymbol{h}_i^T) \tag{12}$$

where $L^A$ denotes the number of adaptation layers and the coefficient $\lambda_i$ is a penalty parameter for the *i*-th adaptation layer.

The second subcategory is the adversarial-based domain adaptation, which is inspired by the Generative Adversarial Networks (GANs) [45] and seeks to endow the deep neural network with the ability of learning domain-invariant representations. The GAN is typically composed of two components, that is, a generator (G) that generates fake data from noise and a discriminator (D) that distinguishes whether an instance is real or generated, which can be optimized by iteratively training D to maximize correct assignment of (real, fake) labels and training G to minimize the differences of real and generated data to confuse the discriminator:

$$\min_G\max_D\mathcal{L}(D,G)=\mathbb{E}_{\boldsymbol{x}\sim P_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{z}\sim P_z(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))] \tag{13}$$

In adversarial-based domain adaptation, such adversarial mechanism has been introduced into the deep neural network to ensure that the characteristics resulting from the difference of diverse domains cannot be distinguished. In light of whether to generate synthetic data, the historical adversarial-based domain adaptation approaches can be summarized as generative or non-generative adaptation model.

The generative adaptation model focuses on generating new data that are similar to the real data of the target domain by directly using GANs. More specifically, in the generative adaptation model, the generator $G(\boldsymbol{x}^S, \boldsymbol{z})$ generates an adapted instance $\boldsymbol{x}^G$ taking a source instance $\boldsymbol{x}^S$ and a noise vector $\boldsymbol{z}$ as inputs, and the discriminator tries to distinguish between the generated instances $\boldsymbol{x}^G$ (fake) and the target instances $\boldsymbol{x}^T$ (real). It is worth noting that, in contrast to the standard GANs in which the input of the generator is only a noise vector, the generative adaptation model's generator takes both a noise vector and a source instance as inputs. The generative adaptation model and its variants could be divided into two types from the perspective of neural network structure. The first one has two stages during model training: (1) generates the synthetic instances to augment the training dataset; (2) trains an extra classifier with both real and generated instances. The second one is usually augmented with a task-specific classifier (T) apart from the G and D, such that the goal of the generative adaptation model is to alternatively optimize the following minimax objective:

$$\min_{G,T} \max_{D} \mathcal{L}(G,D,T) = \alpha \mathcal{L}_{adv}(G,D) + \beta \mathcal{L}_{task}(G,T) \tag{14}$$

where $\alpha$ and $\beta$ are coefficients of the corresponding loss, $\mathcal{L}_{adv}$ and $\mathcal{L}_{task}$ denotes the adversarial loss and task loss, respectively.

The non-generative adaptation model pays more attention to learn the domain-invariant representations, rather than generating new data, by introducing the minimax loss or the domain-confusion loss into the deep model which typically consists of three parts: the feature extractor (instead of the generator), the domain discriminator and the task-specific classifier. One of the promising solutions in implementing the non-generative adaptation is to introduce a special Gradient Reversal Layer (GRL) between the feature extractor and the domain discriminator, which first was introduced in the Domain Adversarial Neural Network (DANN) [46]. DANN ensures that the representations learned from different domains are as closer as possible by maximizing the domain confusion loss through the GRL. The GRL function as an identity transformation during the forward propagation, while during the backward propagation it receives the gradient from the subsequent layer and reverses the sign of the gradient before delivering to the preceding layer:

$$\begin{cases} GRL(\boldsymbol{h}) = \boldsymbol{h}, & \text{forward propagation} \\ \dfrac{dGRL}{\boldsymbol{h}} = -\alpha \boldsymbol{I}, & \text{backward propagation} \end{cases} \tag{15}$$

With such GRL, the parameters of the feature extractor and the domain discriminator can be globally optimized and simultaneously updated. Another promising solution is to splits the optimization into two independent objectives: the parameter of the feature extractor $\theta_{FE}$ and the parameter of the domain discriminator $\theta_D$, and to perform iterative updates for the two objectives given the fixed parameters from the previous iteration:

$$\min_{\theta_D} \mathcal{L}_D(\boldsymbol{x}^S, \boldsymbol{x}^T, \theta_{FE}; \theta_D) = -\mathcal{R}^D\left(\boldsymbol{h}_D, y_D\right) \tag{16}$$

$$\min_{\theta_{FE}} \mathcal{L}_{conf}(\boldsymbol{x}^S, \boldsymbol{x}^T, \theta_D; \theta_{FE}) = -\mathcal{R}^{conf}\left(\boldsymbol{h}_D\right) \tag{17}$$

where $\boldsymbol{h}_D = D(FE(\boldsymbol{x}; \theta_{FE}); \theta_D)$ denotes the output of the domain discriminator, $\mathcal{R}^D$ is the risk function for the domain classification where the Cross-entropy loss function is commonly used, and $\mathcal{R}^{conf}$ is the risk function for the domain confusion where the probability density function of a uniform distribution is adapted based on the cross entropy between the predicted domain labels. Thus, the deep model can be optimized by adversarial training through minimizing the Eq. 15 only for updating $\theta_D$ and minimizing the Eq. 16 for updating $\theta_{FE}$, which can ensure that the learned representations is domain invariant.

The third subcategory is the reconstruction-based domain adaptation, which combines the auto-encoder neural networks with a task-specific classifier to jointly optimize a private encoder that captures domain-specific representations and a shared encoder that learns common representations between the domains. The reconstruction-based domain adaptation model integrates a shared decoder which learns to reconstruct the input instances with a reconstruction loss by taking both the private and the common representations as inputs. The reconstruction losses that have widely used in DTL are the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). The task-specific classifier is trained on the common representations learned by the shared encoder, which will be able to generalize across domains better since its inputs have been separated from the representations that are special to each domain.

Note that, besides the categories of DTL mentioned above, there exist many hybrid methods to build a DTL model using several of the aforementioned techniques simultaneously. The core idea of the hybrid methods is that the domain-invariant knowledge between source and target domain can be learned in any two or more of the levels, that is the instance-, model- and feature-level. Since the main definitions and theories of hybrid methods are almost the same with those mentioned above, this survey will not enumerate them in detail.

## 3 Formulations and Applications of DTL for Fault Diagnosis in Industrial Scenario

In the past few years, scientific researchers and engineers from both academic and industrial communities have already brought many impressive achievements and successful real-world application across a lot of DTL algorithms in implementing complex tasks. Examples include object recognition and detection based on image data collected in different conditions [47], speech recognition based on audio data sampled from different speakers [48], text classification and translation based on document data written in different languages [49], etc. Compared to the vast literature focused on the application in the field of computer vision and natural language processing, few surveys focus on the relative developments of DTL in industrial scenarios for the task of fault diagnosis. Therefore, in this section, the literature historically published in addressing the fault diagnosis problems with the DTL approaches is systematically reviewed, including the problem formulation of DTL for fault diagnosis in industrial scenarios and its main applications of each scenario.

### *3.1 Problem Formulation of DTL for Fault Diagnosis*

Within industrial scenarios, there exist many exact problems that have attracted considerable attentions and much emphasis has been placed on solving such problems. Understanding what type of problems have been faced with IFD and how to solve them is of great significance for researchers and engineers to correctly understand the reasons we survey this topic from the perspective of practical industrial scenarios, and to formulate the pattern of the DTL for fault diagnosis.

In the phase of current manufacturing industry, the major problems encountered in applying intelligent methods for machines are summarized as follows:

(1) The deep models learned from the given training data are not robust enough to be generalized from one application to a new or similar one, so it is difficult to deal with the uncertainty caused by the varying environment during machines working. For instance, the WCs of machines are various during long-term operation, and the health status is also declining with the degradation of crucial components. However, the generalization performance of deep models is insufficient in the face of changeable WCs and diversified data.

(2) Considering the fast upgrading and updating of the manufacturing products, the deep models also require periodic updates for the performance improvement. However, it is hard to collect and annotate the training data from scratch for the application of new products while reusing the labeled historical data collected and accumulated from the old products is relatively easy.

(3) The deep models learn how to make a fault diagnosis through the observations of given labeled data, so they encounter difficulties to recognize unknown patterns or faults. In order to step into the real industrial applications, it is a significant function that the fault diagnosis models can automatically detect a new anomaly since the unseen faults inevitably occur during the long-term services of the complex mechanical equipment.

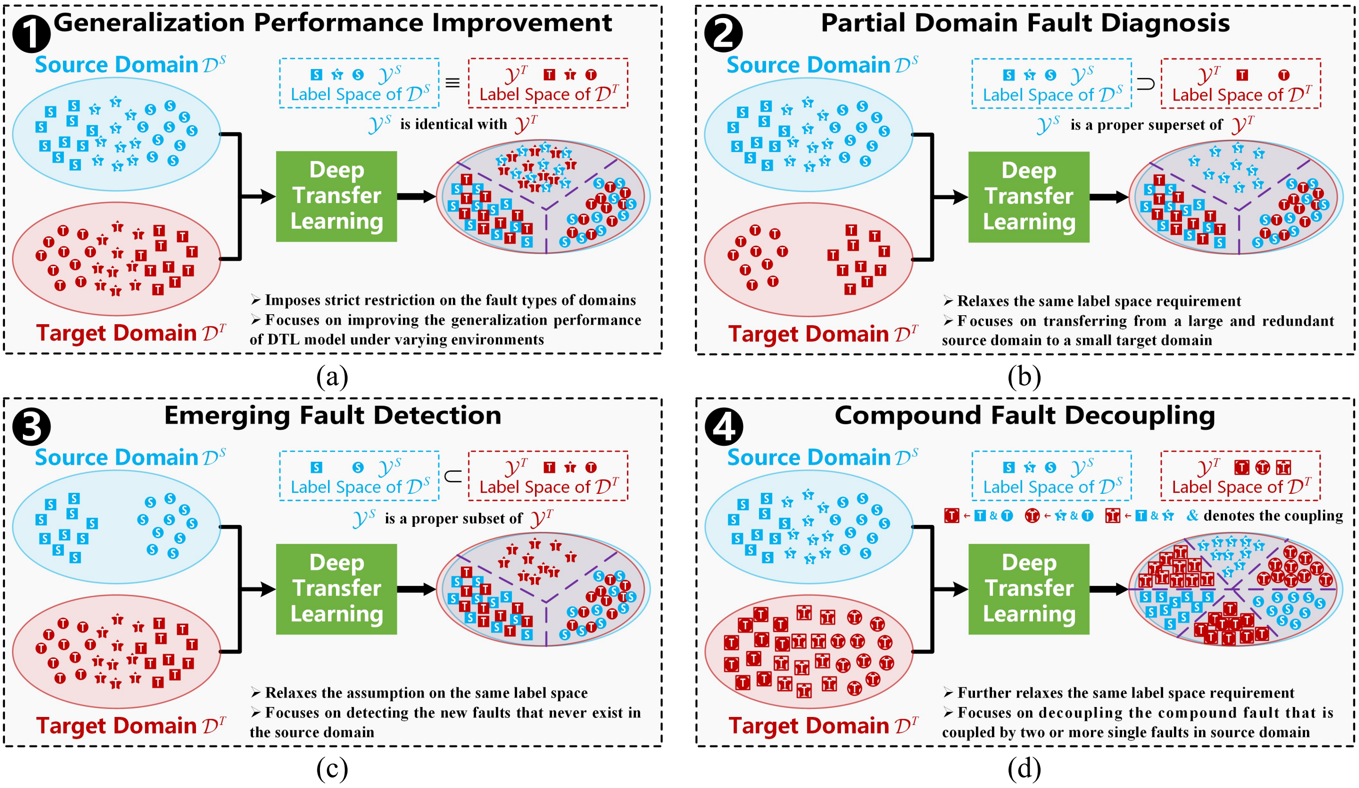


Fig. 4. Illustration of the four application scenarios for DTL (a) Generalization Performance Improvement, (b) Partial Domain Fault Diagnosis, (c) Emerging Fault Detection, (d) Compound Fault Decoupling

(4) The vast majority of researchers and engineers are concerned about the improvement of precision and accuracy for classifying the different faults. The compound fault, as a primary failure leading to expensive maintenance costs and tragic catastrophes in industrial scenarios, often emerges and evolves when multiple crucial components are simultaneously degraded or even broken. However, few time and effort have been paid to investigate the task of decoupling compound faults in an intelligent manner.

Aiming at solving the four problems mentioned above, the industrial application of DTL can be defined as four scenarios: generalization performance improvement, partial domain fault diagnosis, emerging fault detection, and compound fault decoupling, respectively. As descripted in the previous section and shown in Fig. 4, given a DTL task defined by $f^{S\to T}(\bullet): \boldsymbol{X}^T \to \boldsymbol{Y}^T$ based on [$\mathcal{D}^S$, $\mathcal{D}^T$, $\mathcal{T}^S$, $\mathcal{T}^T$], the four application scenarios can be formulated from the perspective of fault classification as follows:

***Generalization performance improvement***: In this scenario, the label space of target domain is identical with the label space of the source domain, that is, $\mathcal{Y}^T \equiv \mathcal{Y}^S$, which imposes strict restriction on the fault types of domains and mainly focuses on improving the generalization performance of DTL model under varying environments. Such scenario is called as generalization performance improvement.

***Partial domain fault diagnosis***: In this scenario, the label space of target domain is a proper subset of the label space of the source domain, that is, $\mathcal{Y}^T \subset \mathcal{Y}^S$, which relaxes the same label space requirement and mainly focuses on transferring knowledge from a large-scale but redundant source domain to an unknown small-scale target domain. Such scenario is referred to as partial domain fault diagnosis.

***Emerging fault detection***: In this scenario, the label space of the target domain is a proper superset of the label space of the source domain, that is, $\mathcal{Y}^T \supset \mathcal{Y}^S$, which also relaxes the assumption on the same label space and mainly focuses on detecting the new faults that never exist in the source domain. Such scenario is known as emerging fault detection.

***Compound fault decoupling***: In this scenario, the label space of the target domain is different from the label space of the source domain, but each fault in target domain is coupled by multiple single faults in the source domain. More specifically, a fault in the target domain $y_i^T$ is a compound fault which is coupled by two or more single faults in the source domain $y_j^S \& \cdots \& y_k^S$. Such scenario is defined as compound fault decoupling.

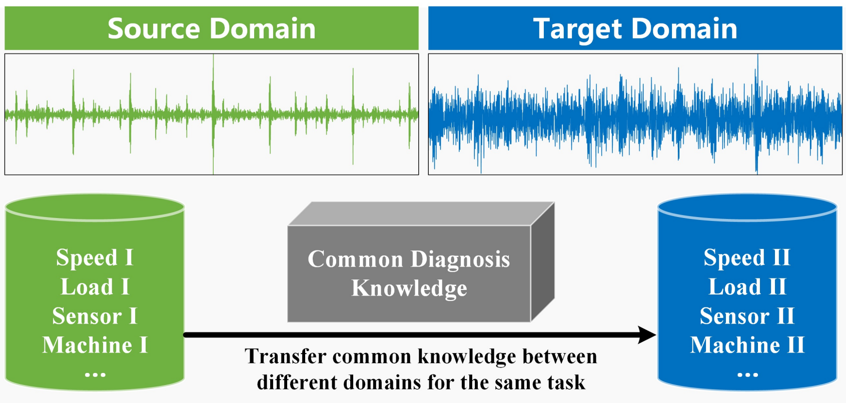


Fig. 5. Illustration of the motivation behind the scenario of generalization performance improvement

### *3.2 Generalization Performance Improvement*

#### 3.2.1 Motivations and goals

As illustrated in Fig. 5, the motivation behind this scenario is that, in the real-world application, if the common knowledge, which does not contain the uncertainty information caused by varying environments, can be learned with limited source data for a specific task, a deep model with satisfactory generalization performance can be obtained for the same task even when it faces a new environment. Thus, as depicted in Fig. 4 (a), the ultimate goal in this scenario is to learn a robust deep model that should be able to implement the objective task under varying environments. Table I concludes the current solutions using the DTL-based fault diagnosis approaches for the generalization performance improvement from three application scenarios, that is varying WCs, across different machines, and other scenarios.

TABLE I Solutions for Generalization Performance Improvement

| Application Scenarios | Categorization of DTL | References | Common algorithms used |
|---|---|---|---|
| Varying Working Conditions | Instance-based | Zhang *et al.* [50] , Shen *et al.* [51], Song *et al.* [52], Pan *et al.* [53] | Weight-estimation with MVD/MMD, Heuristic-reweighting with pseudo-label, TrAdaBoost |
| | Model-based | Shao *et al.* [54], Zhou *et al.* [55], Lu *et al.* [56], Zhang *et al.* [57], Hasan *et al.* [58], [59], Han *et al.* [60], He *et al.* [61], [62], Zhao *et al.* [63], Du *et al.* [64], Chen *et al.* [65], Li *et al.* [66], Wang *et al.* [67], Shao *et al.* [68], Cao *et al.* [69] | Sequential training (DL+Fine-tune), Joint training (Y-Net) |
| | Feature-based | Lu *et al.* [70], Li *et al.* [71], Tong *et al.* [72], [73] Zhang *et al.* [74], Xiao *et al.* [75], An *et al.* [76], Li *et al.* [77], Azamfar *et al.* [78], [79], Zhang *et al.* [80], Zhu *et al.* [81], Singh *et al.* [82], Li *et al.* [83], Han *et al.* [84], [85], Wei *et al.* [86], Wang *et al.* [87], Shen *et al.* [88], Li *et al.* [89], Wu *et al.* [90], Yang *et al.* [91], Qian *et al.* [92], [93], Wang *et al.* [94], An *et al.* [95], Xiong *et al.* [96], Li *et al.* [97], Bao *et al.* [98], Xu *et al.* [99] Huang *et al.* [100] | Discrepancy-based (MMD, CMMD, MK-MMD, KL Divergence, CORAL, CMD, MCD) |
| | | Li *et al.* [101], Zheng *et al.* [102], Liang *et al.* [103], [104], Tao *et al.* [105], Shao *et al.* [106], Guo *et al.* [107], Shi *et al.* [108], Han *et al.* [109], Jiao *et al.* [110], Shao *et al.* [111], Chai *et al.* [112], Chen *et al.* [113], Li *et al.* [114], Mao *et al.* [115], Liu *et al.* [116], Li *et al.* [117], [118], Zhang *et al.* [119], Li *et al.* [120], Zhang *et al.* [121], Jiao *et al.* [122], [123], Cheng *et al.* [124], Wang *et al.* [125], Zou *et al.* [126], Han *et al.* [127], Wang *et al.* [128], Yu *et al.* [129], She *et al.* [130], Ragab *et al.* [131], Liao *et al.* [132] | Adversarial-based (GAN, DATN, DANN, GRL, ADDA, W-GAN) |
| | | Pang *et al.* [133], Liu *et al.* [134], [135], Wen *et al.* [136], Wan *et al.* [137], Tang *et al.* [138] | Reconstruction-based (SAE+TL) |
| Across Different Machines | Instance-based | Zheng *et al.* [144], Yang *et al.* [145], Wu *et al.* [146] | Instance-based discriminative loss |
| | Model-based | Wang *et al.* [147], Shao *et al.* [148], Li *et al.* [149], He *et al.* [150], Chen *et al.* [151] | VGG-19/SAE+Fine-tune |
| | Feature-based | Guo *et al.* [152], Yang *et al.* [139], [153], Li *et al.* [154], Wu *et al.* [155], Zheng *et al.* [156], Lv *et al.* [157], Zhao *et al.* [158], Li *et al.* [159], Tan *et al.* [160], Chen *et al.* [161], Zhang *et al.* [162], [163], [164] | Discrepancy-based (MMD, PK-MMD, MCD, Mutual information) |
| | | Wang, *et al.* [165], Feng *et al.* [166], Zhu *et al.* [167], Liao *et al.* [168] | Adversarial-based (GAN, DATN) |
| | | Lu. *et al.* [169] | Reconstruction-based (SAE+MMD) |
| Others (Imbalanced instances, Across sensors, etc.) | Instance-based | Xiao, *et al.* [170] | TrAdaBoost |
| | Model-based | Li *et al.* [171], Kim *et al.* [172], He, *et al.* [173] | CNN/SAE+Fine-tune |
| | Feature-based | Zhang *et al.* [174], Zou *et al.* [175], Zhang *et al.* [176], Li *et al.* [177], Zareapoor *et al.* [178], Zhang *et al.* [179], Li *et al.* [180], Li *et al.* [181], Siahpour *et al.* [182], Pandhare *et al.* [183], Qin *et al.* [184], Wu *et al.* [185] | GAN and its variants, DANN, One-shot learning, Unsupervised parallel data alignment |

### 3.2.2 Solutions for varying WCs

One of the main factors leading to distribution shift between training and testing data is that the WCs of IE are complex as the frequently changing of speeds, loads or operations. Therefore, in this case, a lot of solutions based on DTL have been investigated for enhancing the generalization performance of deep models which can effectively deal with the uncertainty caused by varying WCs during the long-term services of machines.

***Instance-based DTL solutions:*** Combining the maximum variance discrepancy (MVD) and the maximum mean discrepancy (MMD), Zhang *et al.* [50] proposed a weight-estimation method for bearing fault diagnosis to calculate the adaptation matrix between the source and target instances, which is used to reweight and down-weight the source instances that are negative for the target model training. As the most popular instance reweighting strategy, the fast TrAdaBoost algorithm was introduced by Shen *et al.* [51] as an instance reweighting strategy that can weaken the weight of the low-quality instances and enhance the weight of high-quality instances through iteratively update, which successfully employed to enhance the generalization performance for the fault diagnosis model of a gearbox operating under varying working conditions. Song *et al.* [52] proposed a retraining strategy-based domain adaption network (DAN-R) for IFD, which annotates the unlabeled instances in the target domain with pseudo-labels and then retrains the classification network using both training instances and pseudo-labeled testing instances. According to these instance-based DTL solutions [50]-[53], it can be found that instances-based approaches are effective and applicable for the application scenario of varying WCs. However, the performance of these methods discussed above are depended to some extent on the number or the quality of target instances, and they may have difficulty to tackle the problems in more challenging but complex scenarios which have significant discrepancy between source and target domain.

***Model-based DTL solutions:*** One of the major model-based DTL solutions for varying WCs is to pretrain a deep model (such as VGG-16 [54], VGG-19 [55] and AlexNet [56]) on source WCs and sequentially fine-tune it using the labeled instances in the target working condition [57]-[68]. For instance, Han *et al.* [60] finetuned a well-trained CNN with three transfer learning strategies at different levels of the CNN architecture: (1) just finetuning the classifier, (2) just finetuning the feature extractor, and (3) finetuning the whole CNN model, where the characteristics of each strategy are discussed and compared. The experimental validations show that the proposed transfer strategies can effectively transfer the useful features of the well-trained CNN for the target task and achieve the highest accuracy for the generalization problem of WCs. Another method is to jointly implement the source and target tasks in a deep model with multi-branches. Cao *et al.* [69] developed a multi-branch deep model, named Y-Net, to transfer knowledge for the fault diagnosis of planetary gearboxes, which consists of three components: two convolutional classification networks (one for the source task and another for the target task, and sharing weights with each other) and a reconstruction network. Compared with other solutions training a model from scratch, model-based DTL solutions benefit from the faster convergence rate and the reduction of the risk of overfitting. Additionally, these solutions have its inherent limitation that the fine-tune algorithm heavily relies on the dependence of labeled training data in the target scenario.

***Feature-based DTL solutions:*** For the application scenario of varying WCs, the feature learned by deep models is expected to be speed-insensitive and load-insensitive. Generally, the more insensitive to working conditions the features are, the better the generalization performance of deep model will be. Researchers, therefore, have been placed many efforts on how to learn universal features under varying WCs from the following three aspects.

From the aspect of discrepancy-based domain adaptation, a variety of criteria, such as MMD [70]-[83], Conditional Maximum Mean Discrepancy (CMMD) [84]-[88], MK-MMD [89]-[91], KL Divergence [92], [93], CORAL [94], [95], Central Moment Discrepancy (CMD) [96], [97], Maximum Classifier Discrepancy (MCD) [98] and others [99], [100] have been widely introduced into the objective function to measure the features discrepancy between the source and the target WCs. Specifically, Lu *et al.* [70] proposed a Deep neural network for domain Adaptation in Fault Diagnosis, named DAFD. The DAFD introduced the MMD term into the objective function of deep neural network for reducing the distribution discrepancy between different WCs in 2017, which was the first time that the transfer learning technique, i.e., the domain adaptation, was applied to train the deep model in the field of IFD. Aiming at minimizing the discrepancy of marginal and conditional distributions simultaneously, a deep transfer network (DTN) with joint distribution adaptation (JDA) was proposed by Han *et al.* [84] through the integration of marginal MMD and conditional MMD. Experiments carried on three practical industry datasets show that, comparing with the traditional deep learning- and transfer learning-based methods, the DTN with JDA achieves state-of-the-art diagnosis results regarding the application scenario of diverse operating WCs. In contrast to constructing a single layer with linear MMD in deep model, a multilayer domain adaptation (MLDA) method was proposed by Yang *et al.* [91]. The MLDA that matches the shift in both marginal and conditional distributions across WCs by adding MK-MMD and pseudo-label learning in multiple adaptation layers, can effectively extract working-condition-insensitive features for bearing fault diagnosis. Apart from using the MMD and its variations, Qian *et al.* [92], [93] utilized the KL Divergence to align the first and higher order moment discrepancies, Wang *et al.* [94] and An *et al.* [95] adopted CORAL to minimize the distribution gap between source and target WCs by aligning the second-order statistics, and Li *et al.* [96] employed CMD to reduce the discrepancy between different working conditions.

From the aspect of adversarial-based domain adaptation, the mechanism of GAN has been employed to help the deep model to learn task-sensitive but domain-insensitive features for the target tasks in the generative or non-generative adaptation ways. In the case of generative adaptation, deep GANs and its variants have been exploited to generate different types of data, such as frequency domain data [101], [102] and time-frequency domain data [103]-[105], with the help of available source data, and then these generated and real data are used to train an extra deep model, achieving reliable diagnosis results when testing data in target WCs are not available during model training. Shao *et al.* [106], Guo *et al.* [107] and Shi *et al.* [108] added an auxiliary classifier into the GAN, rather than training an extra classifier, to fully utilize the label information, hence, the enhanced models achieved higher diagnosis accuracy with few training data. However, one particular challenge for generative adaptation is the difficulty in evaluating the quality of the generated data with effective metrics quantitatively. In the case of non-generative adaptation, motivated by GAN, several deep DA frameworks such as the Domain Adversarial Transfer Network (DATN) [109]-[115], the DANN [116]-[119], the Adversarial Discriminative Domain Adaptation (ADDA) [120], the Wasserstein GAN (W-GAN) [121]-[128] and others [129], [130], have been developed and applied for fault diagnosis of machines under varying WCs. For instance, Chen *et al.* [113] exploited the discriminator of GANs as a domain classifier that performs binary domain classification and introduced a domain confusion loss, that is, the inverted label loss, to encourage the source and the target distributions to be a uniform distribution as close as possible. Liu *et al.* [116] utilized the DANN that integrates a GRL into the standard GAN to construct the deep model for bearing fault diagnosis, which largely enhances the generalization performance of diagnosis model under different speeds and loads. Following the principle of ADDA, a knowledge mapping-based adversarial domain adaptation (KMADA) method was proposed by Li and Shen *et al.* [120], which ensures that the feature space mapping from the target domain data can be updated until it is indistinguishable with the feature space mapping from the real source data. The KMADA achieved strong diagnosis results on an experimental bearing dataset and a locomotive bearing dataset. Wang *et al.* [128] proposed a Deep Adversarial Domain Adaptation Network (DADAN) for fault diagnosis of bearings and hard disk datasets collected from real-case data center, which employed a discriminator to measure the empirical Wasserstein distance between two domains instead of using a discriminator to classify the domain label. In addition to the methods mentioned above, some explorations have been proposed to deal with the problems of multiple target domains [131] and domain generalization [132]. Liao *et al.* [132] developed a deep semi-supervised domain generalization network to deal with a challenging diagnosis scenario where the well-trained model can generalize to an unseen working condition.

From the aspect of reconstruction-based domain adaptation, there are also some applications that utilized encoder–decoder reconstruction to enhance the generalization capability of diagnosis model under different WCs [133]-[138]. Wen *et al.* [136] constructed a three-layer SAE to learn the common representations from the raw data of different WCs in a reconstruction manner. An IFD method based on an autoencoder with adaptive transfer learning was proposed by Tang *et al.* [138], which use a shared encoder to learn transferable features using the reconstruction loss of RMSE and the adaptation loss of MMD.

### *3.2.3* Solutions for across different machines

Compared with the application scenario of varying WCs where the data used for model training and testing are both measured on the identical machine under different speeds, loads or operations, the main difference in the application scenario of across different machines is that those data are measured on related but different machines and suffer from more complicated factors, such as different mechanical structures, diverse material and various sizes. Such factors inevitably lead to more significant distribution shift between the training and testing data than the application scenario of varying WCs. Therefore, it is a more challenging task to transfer diagnosis knowledge across different machines. There are three typical applications of across different machines: transfer from laboratory to industry [139], transfer from simulation to reality [140], [141], and transfer from past to future [142], [143]. Once the bridge of transfer knowledge across different machines can be built, it will not only largely eliminate the dependency of the fault data collected from the target machine but also potentially reduce the economic cost spent for the maintenance of the target machine. With such demand increasing exponentially, fortunately, some solutions have been developed and investigated for further improving the practicability and generalization performance of diagnosis model.

***Instance-based DTL solutions:*** A Deep Domain Generalization Network for Fault Diagnosis (DDGFD), optimized with an instance-based discriminative loss, was proposed by Zheng *et al.* [144], aiming to explore the more challenging but practical across different machines scenarios where only normal samples are available in the dataset of the target machine. Yang *et al.* [145] developed a metric, named Optimal Transport-embedded Similarity Measure, for analyzing the transferability of diagnostic knowledge across machines, in which cluster-conditional distributions are explored to assign cluster labels for the target instances. Wu *et al.* [146] proposed a hybrid DTL method that combines the instance- and feature-transfer learning techniques to solve the diagnosis problem of bearings when sufficient labeled fault data in the practical engineering is lacking, which was validated in the application scenario of transfer from the Case Western Reserve University (CWRU) dataset to a locomotive bearing dataset collected in real industry.

***Model-based DTL solutions:*** Taking the rolling bearing fault diagnosis as a case study, Wang and Gao [147] adapted the VGG-19 network as the backbone model that was pretrained on non-manufacturing data, and then was finetuned on manufacturing machine for transferring common latent features among different machines. A Novel Stacked Transfer Auto-encoder (NSTAE), optimized using Particle Swarm Optimization (PSO), was proposed by Shao *et al.* [148] and was applied for IFD based on bearing and gear data collected from different rotating machines. Unlike the previous methods which focus on selecting the backbone model [149], [150], Chen *et al.* [151]

proposed a novel model-based DTL strategy for training a Transferable Convolutional Neural Network (TCNN), which exploits the knowledge learned from different source machines to improve the generalization performance of the target task. Its core idea is that the layers and the parameters of the pretrained TCNN are firstly subdivided into several blocks, and then each block is finetuned in reverse order. With respect to the model transfer, such strategy is suitable not only for the CNNs but also for other deep models such as DBN, SAE and Long short-term memory (LSTM). The model-based DTL solutions, especially the fine-tune algorithm, are comparatively easy to implement in the scenario of across different machines. But their performance would decrease dramatically if the labeled instances are insufficient or unavailable.

***Feature-based DTL solutions:*** For feature-based DTL, the discrepancy-based domain adaptation is still one of the most popular and promising solutions for fault diagnosis in the scenario across different machines and brings successful breakthroughs compared with traditional DL methods [152]-[164]. For example, Lei *et al.* have proposed several IFD methods based on discrepancy-based domain adaptation for transferring knowledge from laboratory to real industrial bearings [139], [152], [153]. A feature-based transfer neural network (FTNN) was proposed in [139] to learn transferable representation by combining multi-layer domain adaptation and pseudo label learning. In FTNN, a domain-shared CNN was trained by simultaneously minimizing three discrepancies: the classification discrepancy of the labeled instances in the source domain, the classification discrepancy of the unlabeled instances in the target domain with the help of the pseudo label learning, and the multilayer MMD discrepancy of the learned representations between across domains. In [153], a distance metric named polynomial kernel induced MMD (PK-MMD) was proposed to overcome the weakness of the Gaussian kernel induced MMD (GK-MMD). The experimental results showed that the PK-MMD based DTL method can not only improve the computation efficiency but also can achieve better performance for IFD in the across different machines scenario compared with other algorithms such as the Transfer Component Analysis (TCA), the DAFD, and the GK-MMD-based method. Meanwhile, Tan *et al.* [160] proposed a deep coupled joint distribution adaptation network (DCJDAN) to reduce the domain discrepancy between artificial and real damages, which has been validated on the dataset provided by Konstruktions- und Antriebstechnik (KAt) Bearing Data Center, Paderborn University. In addition, there are a few published methods which provide other solutions to solve the problems in across different machines scenarios by exploring the adversarial-based [165]-[168] and reconstruction-based domain adaptation techniques [169], which may not be as popular as the methods based on the discrepancy-based domain adaptation.

### 3.2.4 Solutions for other scenarios

Besides varying WCs and across different machines, there are other application scenarios as well, including across sensors and imbalanced instances. As for IFD methods, the locations, types and sampling frequency of sensors, as well as the number of training instances of each class, result in a huge distribution diversity between realistic industrial data. Many impressive studies have been applied to the application scenarios of imbalanced instances [170]-[180] and across different sensors [181]-[185], and paid much attentions to the investigation of how to improve the generalization performance of IFD models.

For the scenario of imbalanced instances, Zareapoor *et al.* [178] proposed a Minority oversampling Generative Adversarial Network (MoGAN) to deal with the problems where the number of each fault class are imbalanced during model training. The MoGAN converts the imbalanced problem into the balanced scenario by generating the minority instances through the GAN, which provides a potential solution for the scenarios where some labeled data are available in the target domain but they are not enough to train a satisfactory model. A one-shot learning method for fault diagnosis of 3D printers was proposed by Li *et al.* [180], which only requires one instance of each fault condition to accomplish the model training. Another scenario encountered in real industry is across different sensors. Prof. Jay Lee and his group have proposed several solutions for transferring diagnosis knowledge across sensors at different locations [181]-[185]. The proposed solutions are based on the unsupervised parallel data which are utilized to align the conditional distribution of the different health conditions. The experimental results showed that such solutions are promising to transfer common knowledge between the data from different locations of machines, and they can further improve the generalization performance of deep models in practical industry applications. Similarly, aiming at transferring the diagnosis model from one sensor to another, Qin *et al.* [184] designed a new transfer strategy for domain adaption, called Multi-Scale Transfer Voting Mechanism (MSTVM), which combines multi-scale feature learning and plurality voting operation techniques. The MSTVM can be used to the traditional domain adaption models, and the model's performance will be well improved.

### *3.3 Partial Domain Fault Diagnosis*

#### 3.3.1 Motivations and goals

As illustrated in Fig. 6, the strong motivation behind this scenario is that, under the industrial big data environment, it is a promising solution to utilize the labeled historical data and the open-source industrial data which are collected from related scenarios, for training a diagnosis model that can transfer knowledge from large-scale but redundant source domain to unknown small-scale target domain. The challenges for partial domain fault diagnosis are due to the following two factors:

(1) Label space information of the target domain is lacking. In the trend of Industry 4.0, a large amount of monitoring data can be collected and stored for the target scenario. However, it is

expensive and unrealistic to annotate these large amounts of data, therefore the numbers and the types of faults are unknown.

(2) Outlier source faults may lead to negative transfer. From the viewpoint of big data, the large-scale but redundant source dataset is diverse enough to subsume all fault classes of the small-scale target dataset.

Thus, directly transferring between the entire source and target domains as the popular DTL methods is not an optimal and effective solution for the partial transfer scenario.

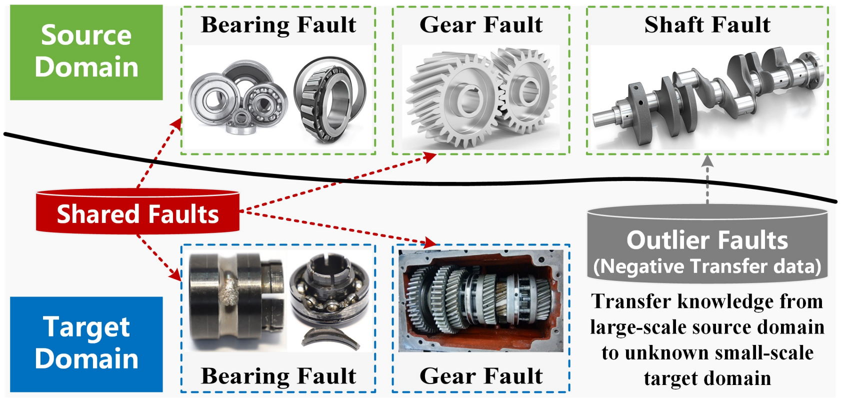


Fig. 6. Illustration of the motivation behind the scenario of partial domain fault diagnosis

As illustrated in Fig. 4 (b), the key goal in this scenario is to build an effective diagnosis model for partial domain fault diagnosis by aligning the distribution of source and target domains in the scope of the shared label space and singling out the outlier source data in the scope of the private label space. The DTL-based solutions developed for the partial domain fault diagnosis in recent years have been summarized in Table II. According to the experiments presented in the publications, these solutions not only have the capability to promote the positive transfer of the relevant data and to alleviate negative transfer of irrelevant data, but also can address the practical and challenging issues under the industrial big data environment.

TABLE II Solutions for Partial Domain Fault Diagnosis

| Application Scenarios | Categorization of DTL | References | Common algorithms used |
|---|---|---|---|
| Partial Domain Fault Diagnosis | Instance-based | Jiao *et al.* [187], Li *et al.* [188], Li *et al.* [189], Liu *et al.* [190] | Class weight-estimation strategy |
| | Feature-based | Li *et al.* [191], Han *et al.* [192], Deng *et al.* [193], Yang *et al.* [194], Wang *et al.* [195] | SAN, GAN+Attention/PK-MMD |

### 3.3.2 Solutions for partial domain fault diagnosis

As mentioned before, the assumption behind this scenario is that the label information of target data is unknown. Up to now, model-based DTL approaches are hardly applied to the problems in the scenario of partial domain fault diagnosis because they inherently rely on the label information of target instances. Therefore, in this case, most current DTL solutions have been developed on the basis of the instance-based and feature-based DTL methods.

***Instance-based DTL solutions:*** An intuitive solution to transfer knowledge from large-scale source dataset to small-scale target dataset is to select out the outlier instances in sources domain that are negative for building target model. Such an idea can be implemented by adapting the instance-level or class-level weighting strategies during the process of model training. Aiming at transferring knowledge from a large-scale dataset to a small-scale dataset (e.g., from ImageNet to Caltech-256), a Selective Adversarial Network (SAN) was firstly proposed by Cao *et al.* [186] in the Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR), 2018, for partial transfer learning. Inspired by the SAN, several class-level weighting methods have been proposed in the field of IFD [187]-[190]. For example, Jiao *et al.* [187] proposed a classifier inconsistency-based domain adaptation network (CIDA) for unsupervised partial domain fault diagnosis of planetary gearbox. The CIDA estimates the label space of target domain by calculating class weights through classifier inconsistency loss and selects out the source instances beyond the shared label space of source and target domains according to the class weights. The experimental results showed that the CIDA can implement the partial transfer diagnosis task from a working condition (containing all fault classes) to a target working condition (only containing a part of fault classes), and its performance is superior than that of the other popular DTL methods. Similarly, a Weighted Adversarial Transfer Network (WATN) was proposed by Li *et al.* [189] for partial domain fault diagnosis across different machines. In WATN, an auxiliary classifier is introduced to automatically learn the weight of each source instance, which can weight the contributions of each training instance to both feature learning and domain confusion. As a result, the role of irrelevant source instances can be effectively weakened during the knowledge transferring. However, these instance-based DTL methods are depended to some extent on the prediction distribution of the instances in the target domain.

***Feature-based DTL solutions:*** Besides the weighting mechanism described above, feature-based DTL solution have also been developed with promising results for partial domain fault diagnosis [191]-[195]. One example is that, inspired by GAN, Li and Zhang proposed an IFD method to address the partial domain adaptation problem by combining the techniques like conditional data alignment and unsupervised prediction consistency. Conditional data alignment is implemented by minimizing the distribution discrepancy between source and target domains through MMD. Unsupervised prediction consistency is achieved when the same prediction results of target domain data can be obtained after finishing the adversarial learning between multiple classification modules and the discriminator [191]. Similar application can be found in [192], which has been validated on a wind turbine fault dataset and achieved superior performance under different transfer scenarios than other traditional transfer learning methods. In addition, a double-layer attention based generative adversarial network (DA-GAN) was proposed by Deng *et al.* [193] for partial domain fault diagnosis of bearings, which aims to solve the problem, “where to transfer”, since the label space of target domain is unknown. In DA-GAN, the attention mechanism is introduced into two layers, one for domain attention and another for sample attention, which can provide guidance for the model to focus which fault classes should be shared or singled out. Yang *et al.* [194] further extended the partial domain fault diagnosis to a more practical and challenging setting where the instances imbalanced between fault classes, exist in the target domain, and proposed a deep partial transfer learning network (DPTL-Net) to selectively transfer diagnosis knowledge for planet gearbox. In DPTL-Net, a domain discriminator is employed to automatically learn domain asymmetry factors via adversarial learning, which can be utilized to weight the PK-MMD. The domain adaptation based on weighted PK-MMD can focus on the distribution discrepancy of source instances in the shared fault classes and filter out the instances in the outlier classes.

With the literatures surveyed above, the instance-based and feature-based DTL solutions have made significant breakthroughs to partial domain fault diagnosis, which can function as a bridge between the large-scale source domain to unknown small-scale target domain for the diagnosis knowledge transfer. However, it is obviously inappropriate to take all the labeled data as the source domain. Therefore, according to the characteristic of the target domain data, how to select the labeled source instances and determine the range of source domain from numerous low-quality industrial data is a challenging problem, which is ignored by the researcher as so far.

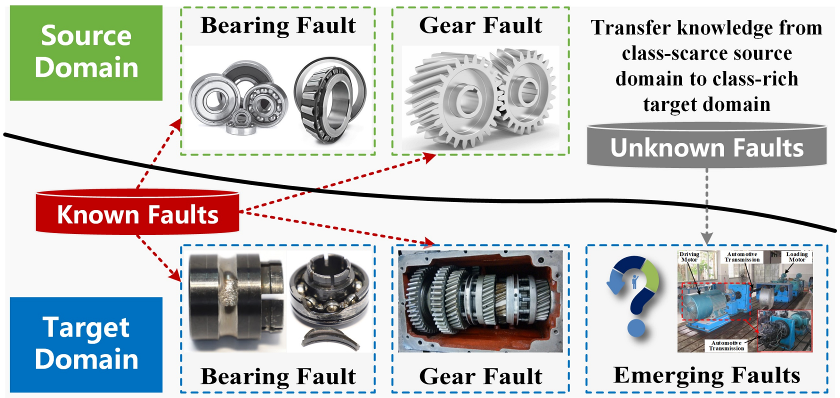


Fig. 7. An illustration of the motivation behind the scenario of emerging fault detection

### *3.4 Emerging Fault Detection*

#### 3.4.1 Motivations and goals

As illustrated in Fig. 7, the critical motivation behind this scenario is that, in practical industry applications, if IFD model can detect the unknown faults which are absent in the labeled source dataset and annotate these faults with correct labels, the IFD model will be able to precisely monitor the health conditions of machines and to continually expand its diagnosis knowledge. In the process of the emerging fault detection, the following two factors that should be taken fully into account are:

(1) Any knowledge about the faults is lacking. The unknown faults are emerging fault classes which newly occur in the target application scenario. More importantly, the unknown faults

never exist in the source domain. It is a challenging task to separate the known and unknown fault classes in an unsupervised manner.

(2) The emerging fault classes may also jeopardize the knowledge alignment between the source and target domains due to the absence of emerging faults in the source domain. In other words, negative transfer will happen if the distribution of the target domain is directly matched with that of the whole source domain.

Different from the partial domain fault diagnosis where knowledge is transferred from class-rich source domains to class-scarce target domain, the emerging fault detection aims at transferring diagnosis knowledge from class-scarce source domain to class-rich target domain. That is, as depicted in Fig. 4 (c), the main goal in this scenario is simultaneously to recognize the emerging faults as "unknown fault" classes and to classify the shared faults of two domains into the correct fault classes. Generally, unpredicted faults are prone to occur since the machines typically operate in complex and uncertain environments during long-term service. Such problem seriously restricts the practical application of the DTL-based methods. Consequently, it is an urgent demand for IFD methods to recognize emerging faults in practical engineering applications. However, there are only a few studies focusing on the emerging fault detection, for which Table III summarizes the current DL-based and DT-based solutions.

TABLE III Solutions for Emerging Fault Detection

| Application Scenarios | Categorization | References | Common algorithms used |
|---|---|---|---|
| Emerging Fault Detection | DL-based | Zhang *et al.* [196], Wang *et al.* [197], Feng *et al.* [198] | Similarity metric |
| | DTL-based | Li *et al.* [199], [200], Wang *et al.* [201], Zhang *et al.* [202], [203], Yang *et al.* [204], Li *et al.* [205], Yu *et al.* [206] | Open set domain adaptation |

### 3.4.2 Solutions for emerging fault detection

Detecting new faults during the testing scenario is one of the key steps for IFD methods when implementing the task of emerging fault detection. In terms of similarity metric learning, several DL-based solutions are established without transfer learning techniques to detecting the emerging faults [196]-[198]. For instance, Zhang *et al.* [196] proposed an emerging new labels method based on SAE (ENL-SAE) for detecting the emerging fault conditions of gearbox. The ENL-SAE forms a prior distribution of known faults with the Gaussian Distribution by utilizing the features extracted by SAE from the training samples, which can be employed to identify the unknown instances whose distribution deviate from the prior distribution of the known faults. These unknown instances are annotated with a new label as the emerging fault and used to retraining the diagnosis model. Simulation and realistic experimental results showed that the ENL-SAE can effectively recognize new faults and improve its practicality. Similarly, a deep metric learning (DML) model was proposed by Wang *et al.* [197], which has capability to classify the new fault by retrieving similarities. In DML, the raw data of each instance are firstly mapped into cosine space, and then the cosine similarity is used to retrieve the most similar fault. The methods mentioned above break through the limitations of the traditional intelligent algorithms owing to the capability of emerging fault detection [198]. However, an obvious bottleneck behind these methods is that they cannot deal with the diagnosis task under complex application scenarios where the distribution shift exists between the training and testing data. Furthermore, some of them rely on a few labeled instances of new faults.

Attempting to further break through the bottleneck mentioned above, the DTL-based solutions have been greatly developed for emerging fault detection in real industry application [199]-[206]. Inspired by the idea of Open Set Domain Adaptation (OSDA) [207], [208], Li *et al.* proposed an IFD method, called Deep Adversarial Transfer Learning Network (DATLN), for detecting the emerging faults of bearings and gearboxes [199], [200], which offered a highly successful attempts on this challenging diagnosis task. The DATLN consists of two components: a feature extractor and a classifier, which are trained by adversarial training. The feature extractor extracts features from input data, and the classifier outputs $K$+1 dimension probability, where $K$ represents the number of known faults in source domain and the $K$+1 th of the classifier output indicates the probability of the unknown fault. On the one hand, the feature extractor aligns the features extracted from the source and target domains, which can deceive the classifier. On the other hand, the classifier can build a decision boundary to recognize the unknown fault in the target domain. The experiments carried on bearing and gearbox datasets showed that the DATLN can not only align the distribution discrepancy between the different domain in the scope of the shared faults, but also can detect the emerging fault with high accuracy. Wang *et al.* [201] proposed a Deep Prototypical Networks based on DA (DPDAN), in which a prototypical layer was applied to learn the prototypes of each fault class and the classification is implemented by finding the nearest class prototype. The DPDAN is another attempt to address the problem where the fault classes of the target scenarios are partially overlapped with that of the source scenarios. Besides the aforementioned feature-based DTL methods, an OSDA method based on Instance-Level Weighted Adversarial Learning was proposed by Zhang *et al.* [202] and applied for IFD of machinery. The instance-level weighted mechanism is introduced to reflect the similarities of testing instances with known faults, therefore, the unknown faults, as well as the known faults, can be effectively identified. Admittedly, these methods are promising for the emerging fault detection and largely improve the applicability of IFD algorithms in the practical engineering. Nevertheless, a major limitation of them is that it can only detect all unknown faults as one category even if there exist multiple emerging faults.

To overcome such limitation, Li *et al.* [205] further extended the DATLN method to a Two-Stage Transfer Adversarial Network (TSTAN) for IFD of rotating machinery with multiple emerging faults. In the first stage, a DTL model is trained by the adversarial learning strategy and employed to single out the unknown fault instances as outliers from the known ones. In the second stage, an unsupervised convolutional SAE with silhouette coefficient is built to further recognize the number of the emerging faults. The TSTAN was validated on two OSDA scenarios: two and three new faults exist in the target domain respectively, and it achieved the highest diagnosis accuracy for the emerging fault detection compared with other state-of-the-art methods. To move one step forward, Yu *et al.* [206] proposed an open set fault diagnosis (OSFD) method with bilateral weighted adversarial networks (BWAN) and extreme value theory for the application scenario where the source and target domains share partially fault classes but hold its private fault classes at the same time. Such assumption is more in accordance with the case of practical engineering in industry. The experimental results on the CWRU and the Traction Motor Bearing (TMB) Dataset illustrate the superior performance of the proposed OSFD approach for emerging fault detection.

With the literature surveyed above, several excellent applications, have been witnessed the last years, addressing the challenging task of emerging fault detection for practical engineering. In terms of the more complex diagnosis task such as the across machines and sensors, however, there are still few or no solutions for the emerging fault detection. More efforts should be placed on these aspects.

## *3.5 Compound Fault Decoupling*

### 3.5.1 Motivations and goals

As illustrated in Fig. 8, the intuitive motivation behind this scenario is that, with the development of intelligent technology, the IFD model should certainly be endowed with the ability to decouple the compound fault in an intelligent manner by only leveraging upon the diagnosis knowledge learned from the data of the corresponding single faults. Such motivation, that is, intelligent compound fault decoupling, is inspired by the phenomenon that human beings are capable of separating the overlapping entities into multiple individual entities easily. As shown in the upper part of Fig. 8, taking the overlapping digits as a concrete example, humans can rapidly capture the key characteristics about each individual digit and can recognize multiple digits in the image even if digits overlap.

However, such an "easy" task is difficult for the majority of IFD algorithms. The challenges for intelligent compound fault decoupling mainly came from the following aspects:

(1) A compound fault occurs unpredictably when multiple key parts and components present defects or even damage at the same time. The monitoring signals become more complex since the fault characteristics of each component are coupled and exerted influence reciprocally, which dramatically increases the difficulties of IFD.

(2) The completeness of compound fault data within the training dataset is hard to be ensured. The practical challenge that is hardly avoidable is that it is difficult and unrealistic to

accumulate single-fault data in industrial applications, let alone to completely collect all types of compound fault data.

(3) The traditional classifier that utilizes the Softmax as the activation function of the last fully connected layer only outputs one label for a testing instance, which inherently determines that the compound fault is simplify regarded as an independent fault pattern for classification and the relationship between the compound fault and its corresponding single faults is ignored.

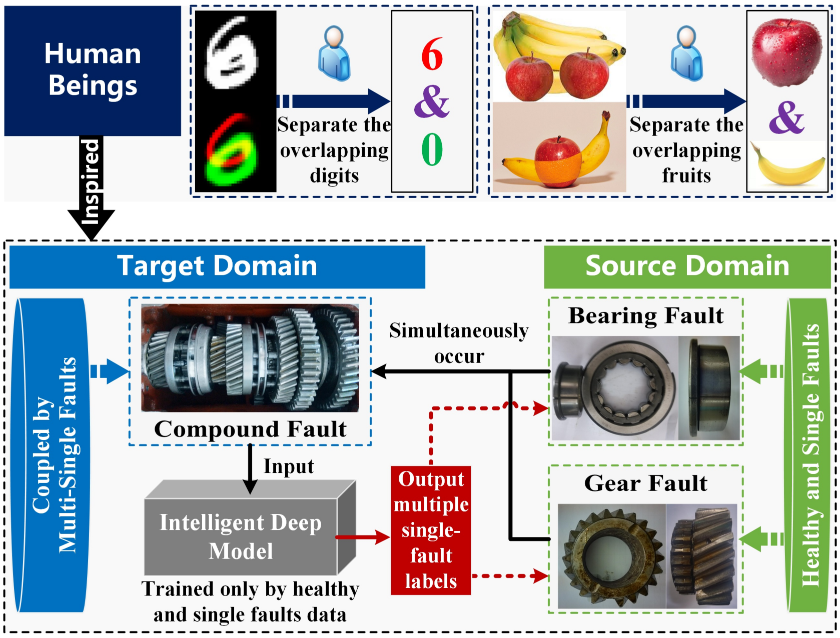


Fig. 8. Illustration of the motivation behind the scenario of compound fault decoupling

Based on the core idea of transfer learning, as illustrated in Fig. 4 (d), the goal here is to develop an IFD model for compound fault decoupling which can learn and capture useful fault characteristics from only the single-fault instances (source domain) and transfer the learned knowledge to help in making a right decoupling of compound fault instances (target domain). Following this insight, as summarized in Table IV, several successful attempts have been made for intelligent compound fault decoupling to imitate the learning ability of humans.

TABLE IV Solutions for Compound Fault Decoupling

| Application Scenarios | Categorization | | References | Common algorithms used |
|---|---|---|---|---|
| Compound Fault Decoupling | Supervised | DL-based | Huang *et al.* [222], Liang *et al.* [223], Jin *et al.* [224] | DNN with multi-label classifier |
| | Unsupervised | DL-based | Huang *et al.* [226], [227], Dibaj *et al.* [228], Xing *et al.* [229] | CNs, Triple probabilistic terms, Zero-shot learning |
| | | DTL-based | Huang *et al.* [230], [231] | TCN, DACN |

### 3.5.2 Solutions for compound fault decoupling

Compound fault diagnosis was and remains a challenging but practical task in the field of fault diagnosis. Before the widely application of IFD, the traditional methods for compound fault diagnosis generally extract the fault characteristic frequencies of each single fault from the monitoring signals of compound fault to make an accurate diagnosis by utilizing the advanced signal processing algorithms [209]-[212]. For example, a compound fault diagnosis method based on multiple enhanced space decomposition was developed by Li *et al.* [211], which can extract the characteristic features of gear defect and bearing fault simultaneously. Cui *et al.* [212] proposed a method based on the Maximum Entropy Deconvolution Adjusted (MEDA) and Adapted Dictionary-free Orthogonal Matching Pursuit (ADOMP) to isolate the compound fault coupled by the gear and bearing faults. Although these solutions can be used to monitor the health states of IE, they heavily rely on the empirical knowledge and the engineering experience of experts, which is a major obstacle for its wide application in industry.

Benefitted from the advantages of DL in representation learning and pattern recognition, some phenomenal solutions have been proposed and applied for the compound fault diagnosis [213]-[221]. For example, in [213], [214] and [215], several DBN-based IFD methods were proposed and applied to diagnose the compound faults of machinery, which mainly focus on enhancing the structure of DBN to improve the performance of diagnosis model. Shao *et al.* [216] developed a multisensory fusion strategy using a stacked wavelet AE structure with a Morlet wavelet function and applied to the collaborative fault diagnosis of planetary gearbox with compound fault. Combining with other techniques, such as adaptive separation, Euclidean matrix sample entropy and adversarial learning, CNN were developed and enhanced for intelligent compound fault diagnosis in many fields [217]-[221].

It can be seen from the publications mentioned above that most of these solutions lose sight of an importance aspect that the compound fault is anything but an individual pattern when it comes to the corresponding single faults. It is inappropriate to simply regard the compound fault as an independent fault class for fault classification. To overcome the shortcoming mentioned above, an intelligent compound fault diagnosis framework based on Deep CNN with multiple-label classifier (DCNN-MLC) was proposed by Huang *et al.* [222] and validated on a gearbox dataset. The core idea of DCNN-MLC is that the sigmoid function, which can transform the output value of each neuron into [0, 1], is employed to substitute the Softmax as the activation function of the last fully connected layer. As a result, the MLC can output single or multiple labels for a testing instance by priorly setting a confidential threshold. The DCNN-MLC is trained with the single faults and compound faults instances, which can decouple the compound fault in a supervised manner by outputting multiple labels. Such an idea has been further investigated and applied for the compound fault diagnosis of gearboxes and bearings [223], [224]. The diagnosis model with MLC is effective for compound fault decoupling by having the ability of outputting multiple labels. However, these models heavily rely on the completeness of compound fault data, suffering setbacks when the labeled data of compound faults are incomplete or even unavailable.

Aiming at eliminating the dependance of completeness of compound fault data, scientific researchers proposed several DL-based solutions for compound fault decoupling in an unsupervised manner, in which the diagnosis model is only trained on the healthy and single faults instances and then can be used to diagnose the compound fault instances [226]-[229]. For instance, inspired by the Capsule Networks (CNs), a Deep Decoupling Convolutional Neural Network (DDCNN) was proposed by Huang *et al.* [226] and applied for intelligent compound fault decoupling of an automobile transmission. In DDCNN, a decoupling classifier is constructed with two capsule layers, rather than a fully connected layer, and is optimized by an agreement-based dynamic routing algorithm, which can decouple the compound fault via outputting multiple labels. The DDCNN is a first successful effort to realize the intelligent compound fault diagnosis by transferring the knowledge learned from the data of healthy and single faults in the scenario that the compound fault data are unavailable during model training. To achieve a common goal, a similar attempt has been investigated by Dibaj *et al.* [228]. The main idea of the method proposed in [228] is that the CNN is trained without the compound fault data, and triple probabilistic conditions are used to restrict the output label of the classifier by judging whether the acquired probabilities of each neuron satisfy these conditions. Thus, the untrained compound fault can be recognized in an intelligent manner. A label description space embedded model for intelligent fault diagnosis (LDS-IFD) was proposed by Xing *et al.* [229] to recognize the compound faults just using the single-faults data during the model training, which is validated by two datasets collected from bearing and planetary gearbox. Admittedly, these solutions have brought successful breakthroughs in intelligent compound fault diagnosis because they eliminate an important problem: the dependance of the completeness of the compound fault data. Nevertheless, the methods mentioned above still lack a robust generalization performance when they encounter a varying and harsh environment, restricting its further practical application in industry.

With the help of transfer learning techniques, the DTL-based solutions for intelligent compound fault decoupling have been attracted increasing attention and application in recent years. The compound fault diagnosis models are getting more generalizable and accessible under varying WCs [230], [231]. Huang *et al.* [230] further proposed a Transferable CN (TCN) for decoupling compound fault of rotating machinery under varying WCs. The TCN is a variant of DDCNN, which can reduce the distribution discrepancy between the source and the target domains by introducing the MMD into the last layer of the feature extractor and the decoupling classifier, respectively. The experimental results demonstrated that the TCN outperforms the DDCNN for the compound fault decoupling under varying WCs. To improve the practicality of diagnosis model, Huang *et al.* [231] further relaxed the assumption on training data by considering that the data cannot be obtained in advance for some special and extreme WCs, and proposed a Deep Adversarial Capsule Network (DACN) which embeds the domain generalization task into the intelligent compound fault diagnosis task. The DACN consists of three parts: the feature extractor (FE), the decoupling classifier (DC) and the multidomain classifier (MC), which is designed for representation learning, compound fault decoupling and multidomain adaptation, respectively. Using the single fault data collected under multiple WCs, the adversarial training strategy is employed to train the DACN. The comprehensive experiments carried on an automobile transmission demonstrates that the DACN is endowed with the ability to decouple the compound fault in an intelligent manner, as well as the ability of strong generalization performance across unseen working condition.

Through the literature surveyed above, the current solutions for intelligent compound fault decoupling have to some extent addressed the two problems: the dependence of data completeness and the lack of robust generalization performance. However, it seems that few studies focus on the more complex industrial scenarios, e.g., the compound fault coupled with three or more single faults, which might be more in accordance with the practical application in industry.

## 4 Suggestions to Select DTL Algorithms for IFD in Industry Applications

After the comprehensive literature survey in Section III, the recent development of DTL approaches in the field of IFD is systematically presented and discussed from the perspective of different industrial application scenarios. To provide a constructive guide for the readers who want to solve the practical industry problems via using DTL-based IFD methods, in this section, the general procedure of IFD based on DTL is concluded, as well as the suggestions to select DTL algorithms for IFD in industry applications.

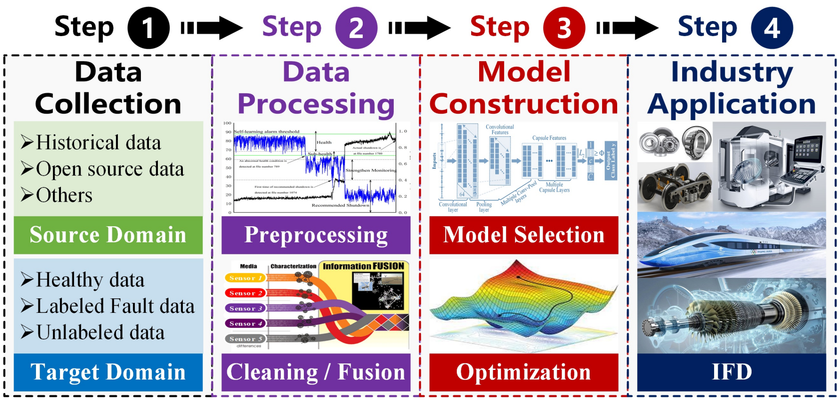


Fig. 9. General procedure of IFD based on DTL

### *4.1 General Procedure of IFD Based on DTL*

As demonstrated in Fig. 9, the general procedure of IFD based on DTL includes four crucial steps: data collection, data processing, model construction, and industry application. Following these steps, a practical IFD project can be implemented in industry applications.

***Step 1: Data Collection.*** In a systematic DTL approach to apply IFD methodologies for a specific task, the first step mostly focuses in collecting the available data from the source and the target domain. Before a DTL algorithm is utilized to accomplish the specific task, it is absolutely necessary and extremely beneficial to familiarize the characteristics of the collected monitoring data and the information of the interested equipment in terms of the key components, WCs, service intensity and all other important physical attributes. In other words, no matter what type of data, such as vibration, electrical and acoustic emission signals, can be collected, the quantity and the quality of data will be fundamental for the subsequent steps in developing an effective solution with dependable diagnosis accuracy via combining appropriate algorithms. As mentioned before, one of the advantages of DTL is that the labeled data in the related but different domains can be used to help training the target model. Therefore, the characteristics of data in the source domain largely affect the performance of the target model. Generally, there are mainly two ways to collect the source domain data. The first way is to use the labeled historical data, collected from similar machines, while the second one is to select similar data from open-source industrial big data. The public datasets, which have been provided by the PHM data challenges that have been held by the PHM Society since 2008, are real data collected from practical industry scenarios. All the datasets are fully opened to all researchers and covered the diagnostics and prognostics tasks in many industry fields, and can be downloaded by the website of PHM Society [1].

***Step 2: Data Processing.*** Contrary to the data collected from laboratory experiments, real industrial big data typically have four main characteristics: large volume, low value density, multi-source and heterogeneous data structure, and monitoring data stream [27]. Therefore, data processing is one of the key steps for improving the performance of the IFD model. Essentially, for an intelligent learning process, garbage data in inevitably leads to garbage results out. There is no one absolute way to prescribe the exact steps in data processing because the process would be better to combine some background information in the specifical scenario. Data cleaning, normalization and data fusion are popular and effective techniques for the processing of original industrial data [232], [233], which can remove errors and inconsistencies and improve the quality of the data that will be used to train the target model.

***Step 3: Model Construction.*** Along with the continuous progress in manufacturing industry, many advanced algorithms have been introduced, developed and benchmarked to implement the diagnostic and prognostic tasks in a supervised or unsupervised manner. Different algorithms are able to handle different problems depending on its adaptability. Therefore, a crucial step for developing an effective solution for IFD in industry scenarios is to select and adopt the most appropriate algorithms, based on available data and target tasks. The suggestions to select the DTL algorithms for IFD will be

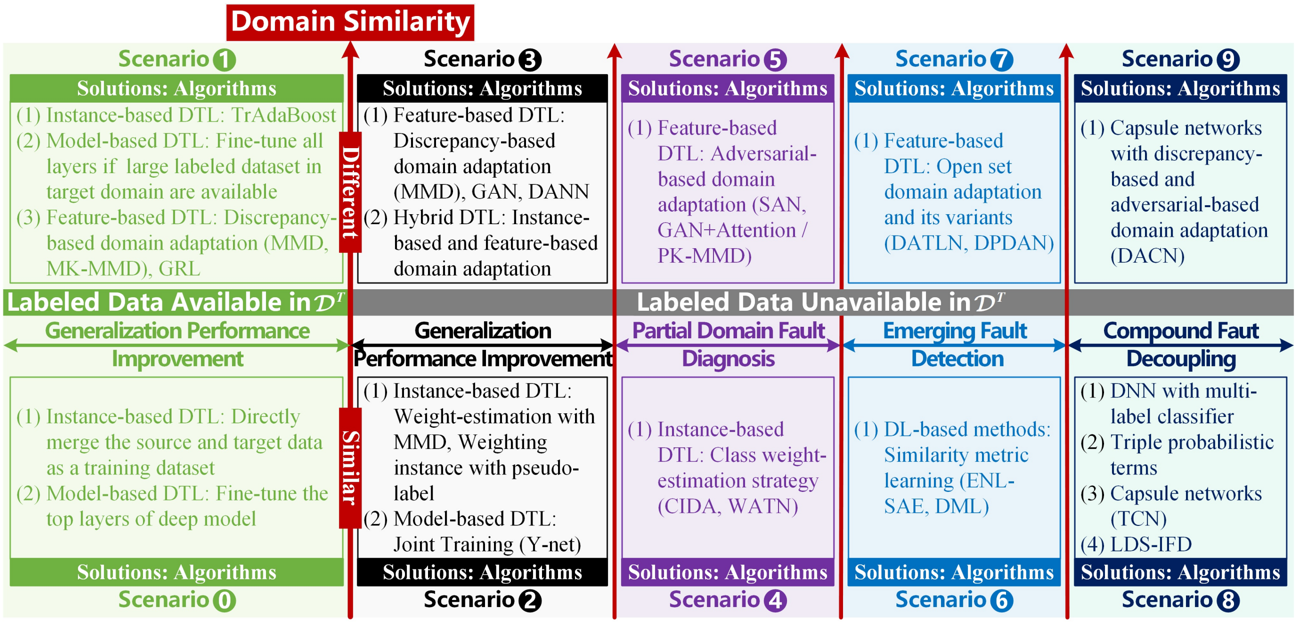


Fig. 10. Suggestions to select appropriate algorithms for practical industry applications

detailly introduced in Section IV, Part B. Once the DTL algorithm is determined, the target model can be optimized according to the source and target data via using gradient-based optimizers, such as Stochastic Gradient Descent (SGD) [234], Adaptive Gradient (Adagrad) [235] and Adaptive Moment Estimation (Adam) [236].

***Step 4: Industry Applications.*** After the diagnosis model has been constructed and optimized by feeding with the data, it can be ready for further application to monitor the health states of target equipment. In this step, it is important to use an Internet of Things (IoT) platform to support the IFD system to convey the useful information to the engineers through visualization tools.

### *4.2 Suggestions to Select DTL Algorithms for IFD*

After the general procedure for IFD methods has been systematically introduced, this subsection will offer some guidance and suggestions to select DTL algorithms according to specific scenarios of the industry applications.

It is an acknowledged truth that there is no general algorithm regarding to the IFD in industry application. At the beginning of selecting DTL algorithms, there are two factors which should be first considered. The first factor is to consider the circumstance of whether the labeled data are available in the target domain, while another one is to evaluate the similarity between the source and the target domain. As illustrated in Fig. 10, the corresponding algorithm selection strategies are provided according to the above two factors and the different industry application.

As for the generalization performance improvement, the application scenario can be further divided into four sub-scenarios: the scenario 0 to 3, in which the appropriate algorithms can be selected by considering the following suggestions.

***Scenario 0:*** Labeled data are available in the target domain, and the source and target domains are similar. Any DTL algorithms may work well in this situation. But, the most efficient and optimal option will be the model-based DTL algorithms, more specifically, the fine-tuning strategy. Since the labeled data are available, the target model can be trained in a supervised manner. Furthermore, since the gap between the source and target domain is small, the knowledge learned from the source domain will be also suitable to the target domain. therefore, it should be enough to directly merge the source and the target data as a training dataset or to fine-tune the top layers of the pre-trained deep model.

***Scenario 1:*** Labeled data are available in the target domain, and the source and target domains are different. In this situation, intuitively, instance-based DTL algorithms, e.g., TrAdaBoost, can be used to single out the similar instances in the source domain to augment the training dataset for the target task. However, such algorithms will be unsuccessful if the data are largely different between the source and the target domains. If the labeled data is sufficient for target model training, fortunately, another solution is to fine-tune all the layers of the pre-trained model. Further, if the labeled data is insufficient, it would be promising solutions to select the feature-based DTL algorithms, such as the discrepancy-based and non-generative domain adaptation.

***Scenario 2:*** Labeled data are unavailable in the target domain, and the source and target domains are similar. Implementing the target task in this situation will be a little bit more difficult than that of in the Scenario 0 due to the fact that the target instances are not annotated. However, since the instances in source and target are similar, the instance-based DTL algorithms, such as the weight-estimation based on kernel embedding techniques and the heuristic weighting strategy, would be a good choice to select out the positive instances in the source domain to help training the target model.

***Scenario 3:*** Labeled data are unavailable in the target domain, and the source and target domains are different. In this situation, the model-based and the instance-based DTL algorithms can hardly improve the generalization performance of the deep model because the label information of the target instances is lacking and the gap between the source and the target domain is large. Therefore, this situation will lead the engineer to the feature-based DTL algorithms (discrepancy-based and adversarial-based domain adaptation). The hybrid DTL algorithms which combine the instances-based and the feature-based domain adaptation will also be a promising tool in this scenario.

As for the other three application scenarios, their basic assumption is that the labeled data in the target domain are unavailable. Therefore, the main factor that should be considered for selecting algorithms is the domain similarity. Each application scenario can be further divided into two sub-scenarios, that is, scenario 4 & 5 (Partial Domain Fault Diagnosis), scenario 6 & 7 (Emerging Fault Detection), and scenario 8 & 9 (Compound Fault Decoupling).

***Scenario 4:*** Source and target domains are similar. It is important to use the similar data as the source dataset for partial domain fault diagnosis. For example, data collected from similar working conditions or same machines are perhaps the best option. As a result, the instance-based DTL algorithm, e.g., the class weight-estimation strategy, is recommended to single out the instances in the shared classes, and then used to train the target model.

***Scenario 5:*** Source and target domains are different. In this situation, since the similar source data are difficult to be collected, it is a potential solution to use the different but related data collected from related industry applications. Considering the demand for reducing the domain discrepancy and avoiding the negative transfer, the feature-based DTL algorithms, especially the adversarial-based domain adaptation (SAN or GAN+PK-MMD), should be given priority.

***Scenario 6:*** Source and target domains are similar. As for emerging fault detection, if the instances in the target domain are similar to those in the source domain, an effective method would be to apply the traditional DL-based methods that detect the new faults by calculating the similarity metric between the testing and the labeled instances.

***Scenario 7:*** Source and target domains are different. In the practical industry application, it is more common that the domain shift exists between the source and the target domains. Therefore, the OSDA algorithm and its variants would be more practical and effective to address the problems of emerging fault detection.

***Scenario 8:*** Source and target domains are similar. Even under an identical working condition and the same machine, it is a challenging task to intelligently decouple the compound fault via using a target model that just trained by single fault instances. If the labeled compound fault data is available, the DNN with MLC can be trained in a supervised manner and further applied for compound fault detection. Otherwise, the deep model can be trained only using the normal and single fault data, and then a rule (e.g., Triple probabilistic terms) can be used to restrict the outputting labels of classifier. From the results shown in literature, the capsule network is the best choice for compound fault decoupling.

***Scenario 9:*** Source and target domains are different. Since the domain shift is introduced with the varying environments, the DNN-based algorithms perform not well in this situation. Up to now, an effective and promising solution is to combine the capsule networks with the feature-based DTL algorithms, such as the discrepancy-based and the adversarial-based domain adaptation.

## 5 Future Challenges and Trends in IFD of Industrial Machinery

An obvious conclusion can be drawn from the comprehensive survey and discussion that, despite the fact that IFD algorithms based on DTL have made successful breakthroughs in many industry applications, there is still a long way to go until it is widely adopted in practical manufacturing industry systems. This is mainly because the performance of DTL algorithms lags far behind the requirements of manufacturing industry which places more emphasis on stability, standardization, accuracy, and repeatability. Before the IFD technologies can be fully embraced and applied in real world industry systems, the researchers in the related field should put significant effort into overcoming the following challenges.

### *5.1 Stability and Reliability*

Historically, the generalization performance of IFD model has been significant improved by leveraging upon transfer learning techniques. However, the current IFD methods based on DTL could only accomplish the well-defined transfer tasks that often have restrictions on WCs, machines, and other hypotheses, which lead to the fact that the IFD model is not yet robust enough in dealing with uncertain circumstances. For a trained IFD model, an uncertain change in the input could cause a large change in the output [237]. Furthermore, most IFD algorithms published in the papers had not been verified as reproducible [238] due to the complexity of model training process and the numerous hyperparameters. In fact, there are many uncertain deviations caused by human or non-human factors during the long-term service of IE, and such deviations will directly affect the robustness, generalization performance and reproducibility of IFD algorithms, resulting in their low stability and reliability in practical industry scenarios. Therefore, it is and remains a challenging task to improve the stability and reliability of IFD algorithms for the technology to truly be applied in practical manufacturing industry systems, which requires further breakthroughs in not only the improvement of generalization performance, but also the reproducibility of diagnosis results.

### *5.2 Interpretability of Deep Model*

Although DTL-based IFD methods have made phenomenal achievements in mechanical fault diagnostics and prognostics, an acknowledged limitation is that these methods have been perceived as black box techniques and are not interpretable, which does not provide a convincing insight into how and why they can make the final decision [239]. This issue may not only put in doubt the credibility of the decision itself, but also lacks compelling evidence to convince companies or industry that these techniques will work repeatedly. Applications in industry have strict requirements for safety and accuracy, and need to explain the reasonableness of the prediction decisions. As a result, the application of the DTL-based IFD methods in manufacturing industry are very limited. In recent years, fortunately, the theory of interpretable machine learning has captured increasing attention from the academic researchers. One way to make intelligent algorithms interpretable is to use only interpretable models, such as Naïve Bayes Classifier and K-Nearest Neighbors, which typically have the limitation that the performance of these model is inferior to other intelligent models [241]. Therefore, in-depth theoretical research should be placed putting more emphasis on opening the "black box" and increasing the transparency of IFD model. Besides the theoretical research, another potential research trend in recent years is to combine the IFD model with a physical/statistical model which is supported by rigorous theory. With the help of the domain knowledge in the physical/statistical model, the "black box" of the IFD model can be partly opened, and it would be easily understood how decisions are reached step-by-step.

### *5.3 Hyperparameters of Deep Model*

Generally, the architecture and hyperparameters of deep model significantly impact on the performance of DTL-based IFD methods. Therefore, it is a crucial step to select the hyperparameters during designing an effective solution with DTL-based IFD algorithms. However, there are no industry consensus on what the ways of selecting hyper-parameters works best. The hyper-parameters are typically selected in most publications via manual setting and experimental validation based on the grid search technique, which is a time-consuming way to ensure the model achieves the optimum performance. In the future, automatic machine learning might be an effective solution to solve such problem [242].

### *5.4 Capacity of Data Processing*

With regard to industry data, the challenges facing IFD right now mainly comes from the following aspects:

***(1) Data Quality.*** The performance of IFD models still depend heavily on the quantity and the quality of historical instances in the source domain, and annotating the industry data requires more engineering experience. In practical industry, it is often the case that, with more smart sensors embedded in machines and advances of measurement technology, large volume of monitoring data can be easily accumulated, but there have problems in data quality, such as lacking correct maintenance records, missing key parameters related to target components, existing misalignment of different variables, and coupling with strong background noise. Andrew Ng, a famous professor in Stanford, points out that the AI systems equals the integration of code (model/algorithm) and data, where the 80/20 rule for the data processing vs model training might be the right balance to achieve success. Therefore, it is necessary to monitor and improve the data quality before developing the IFD solution in practical application.

***(2) Imbalanced Data.*** It is a common case that, in the era of big industry data, the monitoring data of each health state are imbalanced. For ensuring the security and efficiency of production, the IE typically works under healthy conditions. As a result, the fault instances have a much lower chance of appearing than the healthy instances. This makes the data whether in source or target domain having an imbalanced distribution, which in turn makes the IFD model tending to learn biased decision boundaries that have a poorer diagnosis performance over the fault classes compared to the healthy class. Despite the fact that some publications have been focused on the problem of imbalanced data, it is difficult for the proposed solutions to deal with the imbalanced problems in more complex and uncertain industry environment. Therefore, to endow IFD algorithms with the ability to learn the discriminative representation from an extreme imbalanced dataset, more efforts would be necessary to simulate the knowledge transfer process in which humans can correctly guess that an object may belong to the class which share some physical characteristics, instead of brutally training the IFD model with "big data". Following this insight, the few-shot or zero-shot learning, which is inspired by the phenomenon that human beings can learn a new object with only a few instances or even without any instances, are the promising research trends for solving such issue in practical industry application of IFD.

***(3) Heterogeneous Data.*** The industry factory is a typical multiple source heterogeneous data environment. For instance, in wind farms, there are large amount of multiple source heterogeneous data, such as the high-frequency data (current, acoustic emission and vibration signals) and low-frequency data (environmental index, working condition information and control parameters), have been collected from the Supervisory Control and Data Acquisition (SCADA) system and the Condition Monitoring System (CMS). However, as surveyed in the previous sections, the majority of DTL-based IFD methods focus on cases where instances in source and target domains are homogeneous data (e.g., vibration data). The obvious limitation of existing IFD methods is that, if the target sensor malfunctions unexpectedly, the CMS will be out of operation, which in turn could lead to serious catastrophes. Since the multi-source heterogeneous data can provide different information for the same health states of machine, it is possible to transfer diagnosis knowledge from one sensor data to another ones, which may greatly improve the stability and reliability of IFD algorithms. Furthermore, up to now, few studies focus on the heterogeneous transfer learning in the field of IFD. Therefore, heterogeneous transfer learning between multiple sensors would also be one of the future research trends that more attention should be paid.

***(4) Data Privacy and Protection***. In the era of digital and intelligence, industry data is one of the most important assets a company has. For that reason alone, data privacy and protection should be a top priority for any company. It is difficult to reach an agreement and share labeled data among different companies and factories, which in turn results in data fragmentation and isolation. As a result, such restriction poses significant obstacles for the applications of the IFD algorithms in the practical industry. Therefore, how to solve the problem of data fragmentation and isolation while considering and complying with the restriction of data privacy and protection is one of the major challenges for the IFD algorithms to truly accommodate a wider range of application in practical industry. One potential research trend for addressing the above issues is to combine the federated learning and DTL to build and train an effective and accurate IFD model [243].

### *5.5 Challenges in Transfer Learning*

To design an effective DTL-based algorithms, there are still several key challenges should be placed more efforts on.

***(1) Identifying the Appropriate Source Domain***

The previous survey on DTL-based IFD algorithms elaborates several ways to transfer source domain knowledge for practical industrial applications, however, identifying an appropriate source domain is still a challenging problem due to the challenges caused by big industrial data. For example, for many industrial applications, it is difficult to find an appropriate source domain that includes sufficient training instances annotated with precise label information for implementing target tasks. Even worse, it may be unrealistic to find any failure data from similar or related industrial application. With the rapid development of digital technology, such as Digital Twins, one promising way is to utilize the simulation or generation techniques to generate training data as the source domain in such a scenario. In addition, transferring the knowledge from multiple source domains has been attracted more and more attentions recently.

***(2) Avoiding Negative Transfer***

Once the source domain is determined, avoiding negative transfer is also a challenging problem during building a DTL model. As illustrated in Section 3, although there are several tricks have been proposed for avoiding negative transfer, it should be highlighted that negative transfer still needs further systematic investigation. One of effective measures to improve the performance of the DTL-based IFD model in industrial scenarios is to transfer only the common knowledge that can contribute to the target learning task and to avoid negative transfer at the same time. For example, developing an accurate “distance” metrics between the domains might be a feasible solution for avoiding negative transfer since the existing metrics used in feature-based DTL are not powerful enough in developing a perfect transfer learning application.

***(3) Assessing Transferability***

Assessing the transferability across domains in quantitative is another challenging problem during developing a DTL-based IFD method in industrial scenario. However, as so far, there is still few publications focusing on assessing transferability between the source and target domains mathematically. We confident that assessing transferability across domain will be a significantly important research trend in the future, which will enhance the performance of the DTL-based methods and further boost the application of DTL in industrial scenarios.

### *5.6 Computation and Energy Efficiency*

According to the aforementioned literature survey, it is generally the case that the DTL-based IFD methods suffer from the high requirement of computational source and speed. The inefficient and large computation in deep model has hindered the successful application of IFD methods in real-time data analytics. However, the capability of real-time monitoring is fundamental to PHM systems, which can improve the security of machinal systems, identify potential faults as soon as they occur, allow for early maintenance, and avoid systems failures. Therefore, the real-time IFD algorithms should be encouraged to be investigated to ensure real-time decision-making for monitoring the incipient damages or unexpected faults [244]. Techniques, including efficient neural network compression, incremental learning and deep reinforcement learning [245], are potential research directions to facilitate the real-time ability of DTL based IFD algorithms.

## 6 Conclusions

In this survey article, the theory and strategies of DTL methods have been summarized from the algorithm perspective, which gives the basic definitions related to DTL and explain how the TL technologies can help improving the performance of DL model. The state-of-the-art applications of DTL-based IFD approaches have also been overviewed from the perspective of practical industrial applications, in which the four major application scenarios: generalization performance improvement, partial domain fault diagnosis, emerging fault detection, and compound fault decoupling, are formulated and fully discussed. Thereafter, the suggestions for the selection of DTL algorithms for a new IFD project have been detailed, as well as the future challenges and potential trends. This review article not only leads readers to easily understand the current state-of-the-art DTL techniques related to IFD and to quickly design an effective solution for solving IFD problems in practice, but also provides the main challenges facing IFD until it has wide adoption in practical manufacturing industry systems, as well as the future research trends, for researchers and scholars.

## Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1702400, in part by National Natural Science Foundation of China under Grant 51875208, in part by Key-Area Research and Development Program of Guangdong Province under Grant 2019B010154002. The last author acknowledges the support of Flanders Make and VLAIO in the frames of DGTwin Prediction project, the support of Research Foundation - Flanders (FWO) for its support through the ROBUSTIFY research grant no. S006119N and the Flemish Government under the “Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen” programme.

## References


[1] J. Lee, Industrial AI, Singapore:Springer, 2020.

[2] J. Wang, L. Ye, RX. Gao, C. Li, and L. Zhang, "Digital twin for rotating machinery fault diagnosis in smart manufacturing," *International Journal of Production Research,* vol.57, no.12, pp: 3920-3934, 2019.

[3] J. Wan, X. Li, H. -N. Dai, A. Kusiak, M. Martínez-García and D. Li, "Artificial-Intelligence-Driven Customized Manufacturing Factory: Key Technologies, Applications, and Challenges," *Proceedings of the IEEE*, doi: 10.1109/JPROC.2020.3034808.

[4] R. Yan, R. X. Gao and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," *Signal Process.*, vol. 96, pp. 1-15, Mar. 2014.

[5] X. Yang, K. Ding, and G. He, "Phenomenon-model-based AM-FM vibration mechanism of faulty spur gear," *Mech. Syst. Signal Process.*, vol. 134, Dec. 2019, Art. no. 106366.

[6] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, 2018.

[7] R. Huang, J. Li, S. Wang, G. Li and W. Li, "A Robust Weight-Shared Capsule Network for Intelligent Machinery Fault Diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6466-6475, Oct. 2020.

[8] Datong Liu, Yuchen Song, Lyu Li, Haitao Liao, Yu Peng. On-line life cycle health assessment for lithium-ion battery in electric vehicles, *Journal of Cleaner Production (JCLP)*, 199: 1050-1065, 2018.

[9] B. Zhang, S. Zhang, and W. Li, "Bearing performance degradation assessment using long short-term memory recurrent network," *Comput. Ind.*, vol. 106, pp. 14–29, Apr. 2019.

[10] Z. Chen, K. Gryllias, and W. Li, "Mechanical fault diagnosis using convolutional neural networks and extreme learning machine," *Mech. Syst. Signal Process.*, vol. 133, Nov. 2019, Art. no. 106272.

[11] X. Hong, B. Zhang, Y. Liu, H. Qi and W. Li, "Deep-learning-based guided wave detection for liquid-level state in porcelain bushing type terminal," *Structural Control and Health Monitoring*, vol. 28, no. 1, Jan. 2021, e2651, DOI: https://doi.org/10.1002/stc.2651.

[12] Li X, Yang Y, Shao H, et al. Symplectic weighted sparse support matrix machine for gear fault diagnosis[J]. Measurement, 2021, 168: 108392.

[13] Z. Chen, J. Xu, T. Peng and C. Yang, "Graph Convolutional Network-Based Method for Fault Diagnosis Using a Hybrid of Measurement and Prior Knowledge," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2021.3059002.

[14] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, May 2015.

[15] S. Khan, and T. Yairi, "A review on the application of deep learning in system health management," *Mech. Syst. Signal Process.*, vol. 107, pp. 241–265, 2018.

[16] Li Q, Ding X, He Q, et al. Manifold Sensing-Based Convolution Sparse Self-Learning for Defective Bearing Morphological Feature Extraction[J]. IEEE Trans. Ind. Informatics, 2021, 17(5): 3069-3078.

[17] S. Zhang, Z. Sun, C. Li, D. Cabrera, J. Long and Y. Bai, "Deep Hybrid State Network With Feature Reinforcement for Intelligent Fault Diagnosis of Delta 3-D Printers," IEEE Transactions on Industrial Informatics, vol. 16, no. 2, pp. 779-789, Feb. 2020.

[18] S. Zhang, Z. Sun, M. Wang, J. Long, Y. Bai and C. Li, "Deep Fuzzy Echo State Networks for Machinery Fault Diagnosis," IEEE Transactions on Fuzzy Systems, vol. 28, no. 7, pp. 1205-1218, July 2020.

[19] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[20] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43-76, Jan. 2021.

[21] C. Tan, F. Sun and T. Kong, "A survey on deep transfer learning," *Proc. Int. Conf. Artif. Neural Netw.*, pp. 270-279, 2018.

[22] C. Li et al., “Fusing convolutional generative adversarial encoders for 3D printer fault detection with only normal condition signals,” *Mech. Syst. Signal Process.*, vol. 147, Jan. 2021, Art. no. 107108

[23] S. Zhang, X. Duan, C. Li, M. Liang, “Pre-classified reservoir computing for the fault diagnosis of 3D printers,” *Mech. Syst. Signal Process.*, vol. 146, Jan. 2021, Art. no. 106961

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

[25] R. Liu, B. Yang, E. Zio and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vol. 108, pp. 33-47, Aug. 2018.

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

[27] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020.

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

[29] R. Yan, F. Shen, C. Sun and X. Chen, “Knowledge Transfer for Rotary Machine Fault Diagnosis," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8374-8393, Aug. 2020.

[30] J. Y. Jiao, M. Zhao, J. Lin and K. X. Liang, "A comprehensive review on convolutional neural network in machine fault diagnosis," *Neurocomputing*, vol. 417, pp. 36-63, 2020.

[31] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, ‘‘1D convolutional neural networks and applications: A survey,’’ *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107398.

[32] Zhao Z, Wu J, Li T, et al. Challenges and Opportunities of AI-Enabled Monitoring, Diagnosis & Prognosis: A Review[J]. Chinese Journal of Mechanical Engineering, 2021, 34(1): 1-29.

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

[34] S. Lu, R. Yan, Y. Liu and Q. Wang, "Tacholess Speed Estimation in Order Tracking: A Review With Application to Rotating Machine Fault Diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 7, pp. 2315-2332, July 2019.

[35] A. Stetco, F. Dinmohammadi, X. Zhao, V. Robu, D. Flynn, M. Barnes, et al., "Machine learning methods for wind turbine condition monitoring: A review," *Renew. Energ.*, vol. 133, pp. 620-635, Apr. 2019.

[36] C. Li, S. Zhang, Y. Qin and E. Estupinan, "A systematic review of deep transfer learning for machinery fault diagnosis," *Neurocomputing*, vol. 407, pp. 121-135, Sep. 2020.

[37] S. Zhang, Z. Sun, J. Long, C. Li, and Y. Bai, “Dynamic condition monitoring for 3D printers by using error fusion of multiple sparse autoencoders,” Comput. Ind., vol. 105, pp. 164–176, Feb. 2019.

[38] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, Transfer Learning. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[39] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf and G. R. Lanckriet, “Hilbert space Embeddings and metrics on probability measures,” *J. Mach. Learning Res.*, vol. 99, pp. 1517–1561, 2010.

[40] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, “Direct importance estimation for covariate shift adaptation,” *Ann. Inst. Stat. Math.*, vol. 60, no. 4, pp. 699–746, Dec. 2008.

[41] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” *in Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 193–200.

[42] M. Long, Y. Cao, J. Wang, and M. I. Jordan, “Learning transferable features with deep adaptation networks,” *in Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[43] B. Sun and K. Saenko, “Deep Coral: Correlation alignment for deep domain adaptation,” *in Proc. European Conf. Computer Vision*, 2016, pp. 443–450.

[44] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.

[45] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[46] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” *in Proc. ICML*, 2015, pp. 1180–1189.

[47] H. Chen, Y. Li and D. Su, "Discriminative Cross-Modal Transfer Learning and Densely Cross-Level Feedback Fusion for RGB-D Salient Object Detection," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4808-4820, Nov. 2020.

[48] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, vol. 7, pp. 19143-19165, 2019.

[49] D. Wang et al., "Softly Associative Transfer Learning for Cross-Domain Classification," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4709-4721, Nov. 2020.

[50] Z. Zhang, H. Chen, S. Li and Z. An, "Unsupervised domain adaptation via enhanced transfer joint matching for bearing fault diagnosis," *Measurement*, vol. 165, Dec. 2020.

[51] F. Shen, R. Langari, and R. Yan. "Exploring Sample/Feature Hybrid Transfer for Gear Fault Diagnosis Under Varying Working Conditions." *J. Comput. Inf. Sci. Eng.*, vol. 20, no. 4, 2020.

[52] Y. Song, Y. Li, L. Jia and M. Qiu, "Retraining Strategy-Based Domain Adaption Network for Intelligent Fault Diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6163-6171, Sept. 2020.

[53] Y. Pan, F. Mei, H. Miao, J. Zheng, K. Zhu and H. Sha, "An approach for HVCB mechanical fault diagnosis based on a deep belief network and a transfer learning strategy," *J. Elect. Eng. Technol.*, vol. 14, no. 1, pp. 407-419, Jan. 2019.

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

[55] J. Zhou, X. Yang, L. Zhang, S. Shao, G. Bian, "Multisignal VGG19 Network with Transposed Convolution for Rotating Machinery Fault Diagnosis Based on Deep Transfer Learning," *Shock and Vibration*, vol. 2020.

[56] T. Lu, F. Yu, B. Han and J. Wang, "A Generic Intelligent Bearing Fault Diagnosis System Using Convolutional Neural Networks With Transfer Learning," *IEEE Access*, vol. 8, pp. 164807-164814, 2020.

[57] R. Zhang, H. Tao, L. Wu and Y. Guan, "Transfer Learning With Neural Networks for Bearing Fault Diagnosis in Changing Working Conditions," *IEEE Access*, vol. 5, pp. 14347-14357, 2017。

[58] M. J. Hasan and J.-M. Kim, "Bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning," *Appl. Sci.-Basel*, vol. 8, no. 12, Dec. 2018.

[59] M. J. Hasan, M. M. M. Islam and J.-M. Kim, "Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions," *Measurement*, vol. 138, pp. 620-631, May 2019.

[60] T. Han, C. Liu, W. Yang and D. Jiang, "Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions," *ISA Trans.*, vol. 93, pp. 341-353, Oct. 2019.

[61] Z. He, H. Shao, X. Zhang, J. Cheng and Y. Yang, "Improved Deep Transfer Auto-Encoder for Fault Diagnosis of Gearbox Under Variable Working Conditions With Small Training Samples," *IEEE Access*, vol. 7, pp. 115368-115377, 2019.

[62] Z. He, H. Shao, X. Zhong and X. Zhao, "Ensemble transfer CNNs driven by multi-channel signals for fault diagnosis of rotating machinery cross working conditions", *Knowl.-Based Syst.*, vol. 207, Nov. 2020.

[63] B. Zhao, X. Zhang, Z. Zhan, and S. Pang, "Deep multi-scale convolutional transfer learning network: A novel method for intelligent fault diagnosis of rolling bearings under variable working conditions and domains," *Neurocomputing*, vol. 407, pp. 24–38, Sep. 2020.

[64] Y. Du, A. Wang, S. Wang, B. He, G. Meng, "Fault Diagnosis under Variable Working Conditions Based on STFT and Transfer Deep Residual Network", *Shock and Vibration*, vol. 2020.

[65] Z. Chen, J. Cen and J. Xiong, "Rolling Bearing Fault Diagnosis Using Time-Frequency Analysis and Deep Transfer Convolutional Neural Network," *IEEE Access*, vol. 8, pp. 150248-150261, 2020.

[66] F. Li, J. Chen, J. Pan and T. Pan, "Cross-domain learning in rotating machinery fault diagnosis under various operating conditions based on parameter transfer," *Meas. Sci. Technol.*, vol. 31, no. 8, 085104, 2020.

[67] Z. Wang, Q. Liu, H. Chen and X. Chu, "A deformable CNN-DLSTM based transfer learning method for fault diagnosis of rolling bearing under multiple working conditions," I*nternational Journal of Production Research*, 2020. DOI: 10.1080/00207543.2020.1808261

[68] H. Shao, M. Xia, G. Han, Y. Zhang and J. Wan, "Intelligent Fault Diagnosis of Rotor-Bearing System Under Varying Working Conditions With Modified Transfer Convolutional Neural Network and Thermal Images," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3488-3496, May 2021.

[69] X. Cao, B. Chen and N. Zeng, "A deep domain adaption model with multi-task networks for planetary gearbox fault diagnosis," *Neurocomputing*, vol. 409, pp. 173-190, Oct. 2020.

[70] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang and T. Zhang, "Deep Model Based Domain Adaptation for Fault Diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296-2305, March 2017.

[71] X. Li, W. Zhang and Q. Ding, "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning," *Neurocomputing*, vol. 310, pp. 77-95, 2018.

[72] Z. Tong, W. Li, B. Zhang, F. Jiang and G. Zhou, "Bearing Fault Diagnosis Under Variable Working Conditions Based on Domain Adaptation Using Feature Transfer Learning," *IEEE Access*, vol. 6, pp. 76187-76197, 2018.

[73] Z. Tong, W. Li, B. Zhang and M. Zhang, "Bearing fault diagnosis based on domain adaptation using transferable features under different working conditions," *Shock Vib.*, vol. 2018, Jun. 2018.

[74] B. Zhang, W. Li, X. Li and S. Ng, "Intelligent Fault Diagnosis Under Varying Working Conditions Based on Domain Adaptive Convolutional Neural Networks," *IEEE Access*, vol. 6, pp. 66367-66384, 2018.

[75] D. Xiao, Y. Huang, L. Zhao, C. Qin, H. Shi and C. Liu, "Domain Adaptive Motor Fault Diagnosis Using Deep Transfer Learning," *IEEE Access*, vol. 7, pp. 80937-80949, 2019.

[76] Z. An, S. Li, Y. Xin, K. Xu and H. Ma, "An intelligent fault diagnosis framework dealing with arbitrary length inputs under different working conditions," *Meas. Sci. Technol.*, vol. 30, no. 12, Dec. 2019.

[77] Q. Li, B. Tang, L. Deng, Y. Wu and Y. Wang, "Deep balanced domain adaptation neural networks for fault diagnosis of planetary gearboxes with limited labeled data," *Measurement*, vol. 156, May 2020.

[78] M. Azamfar, X. Li and J. Lee, "Intelligent ball screw fault diagnosis using a deep domain adaptation methodology," *Mechanism Mach. Theory*, vol. 151, Sep. 2020.

[79] M. Azamfar, X. Li and J. Lee, "Deep Learning-Based Domain Adaptation Method for Fault Diagnosis in Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 3, pp. 445-453, Aug. 2020.

[80] Y. Zhang, Z. Ren, S. Zhou, "A New Deep Convolutional Domain Adaptation Network for Bearing Fault Diagnosis under Different Working Conditions," *Shock and Vibration*, vol. 2020, Article ID 8850976.

[81] J. Zhu, N. Chen and C. Shen, "A New Deep Transfer Learning Method for Bearing Fault Diagnosis Under Different Working Conditions," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8394-8402, 1 Aug.1, 2020.

[82] J. Singh, M. Azamfar, A. Ainapure and J. Lee, "Deep learning-based cross-domain adaptation for gearbox fault diagnosis under variable speed conditions," *Meas. Sci. Technol.*, vol. 31, no. 5, Feb. 2020.

[83] Y. Li, Y. Song, L. Jia, S. Gao, Q. Li and M. Qiu, "Intelligent Fault Diagnosis by Fusing Domain Adversarial Training and Maximum Mean Discrepancy via Ensemble Learning," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2833-2841, April 2021.

[84] T. Han, C. Liu, W. Yang and D. Jiang, "Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application," *ISA Trans.*, vol. 97, pp. 269-281, Feb. 2020.

[85] Han T, Li Y F, Qian M. A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions[J]. IEEE Transactions on Instrumentation and Measurement, 2021.

[86] Wei D, Han T, Chu F, et al. Weighted domain adaptation networks for machinery fault diagnosis[J]. Mechanical Systems and Signal Processing, 2021, 158: 107744.

[87] X. Wang, C. Shen, M. Xia, D. Wang, J. Zhu and Z. Zhu, "Multi-scale deep intra-class transfer learning for bearing fault diagnosis," *Rel. Eng. Syst. Saf.*, vol. 202, Oct. 2020.

[88] C. Shen, X. Wang, D. Wang, Y. Li, J. Zhu and M. Gong, "Dynamic Joint Distribution Alignment Network for Bearing Fault Diagnosis Under Variable Working Conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-13, 2021, Art no. 3510813.

[89] X. Li, W. Zhang, Q. Ding and J.-Q. Sun, "Multi-layer domain adaptation method for rolling bearing fault diagnosis," *Signal Process.*, vol. 157, pp. 180-197, 2019.

[90] J. Wu, T. Tang, M. Chen, Y. Wang, and K. Wang, “A study on adaptation lightweight architecture based deep learning models for bearing fault diagnosis under varying working conditions,” *Expert Syst. Appl.*, vol. 160, Dec. 2020, Art. no. 113710.

[91] B. Yang, Q. Li, L. Chen, C. Shen, "Bearing Fault Diagnosis Based on Multilayer Domain Adaptation," *Shock and Vibration*, vol. 2020, Article ID 8873960.

[92] W. Qian, S. Li and J. Wang, "A New Transfer Learning Method and its Application on Rotating Machine Fault Diagnosis Under Variant Working Conditions," *IEEE Access*, vol. 6, pp. 69907-69917, 2018.

[93] W. Qian, S. Li and X. Jiang, "Deep transfer network for rotating machine fault analysis," *Pattern Recognit.*, vol. 96, Dec. 2019.

[94] X. Wang, H. He and L. Li, "A Hierarchical Deep Domain Adaptation Approach for Fault Diagnosis of Power Plant Thermal System," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5139-5148, Sept. 2019.

[95] J. An, P. Ai, D. Liu, "Deep Domain Adaptation Model for Bearing Fault Diagnosis with Domain Alignment and Discriminative Feature Learning," *Shock and Vibration*, vol. 2020, Article ID 4676701.

[96] P. Xiong, B. Tang, L. Deng, M. Zhao, X. Yu, "Multi-block domain adaptation with central moment discrepancy for fault diagnosis," *Measurement*, vol. 169, Feb. 2021.

[97] X. Li, Y. Hu, J. Zheng, M. Li, W. Ma, "Central moment discrepancy based domain adaptation for intelligent bearing fault diagnosis," *Neurocomputing*, vol. 429, pp. 12-24, Mar. 2020.

[98] H. Bao, Z. Yan, S. Ji, J. Wang, S. Jia, G. Zhang and B. Han, "An Enhanced Sparse Filtering Method for Transfer Fault Diagnosis Using Maximum Classifier Discrepancy", *Meas. Sci. Technol.*, 2021, DOI: https://doi.org/10.1088/1361-6501/abe56f.

[99] K. Xu, S. Li, X. Jiang, J. Lu, T. Yu, R. Li, "A novel transfer diagnosis method under unbalanced sample based on discrete-peak joint attention enhancement mechanism," *Knowl.-Based Syst.*, vol. 212, Jan. 2021.

[100] Z. Huang et al., "A multi-source dense adaptation adversarial network for fault diagnosis of machinery," in IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2021.3086707.

[101] X. Li, W. Zhang and Q. Ding, "Cross-Domain Fault Diagnosis of Rolling Element Bearings Using Deep Generative Neural Networks," *IEEE Trans. Ind. Electron.*, vol. 66, no. 7, pp. 5525-5534, July 2019.

[102] T. Zheng, L. Song, J. Wang, W. Teng, X. Xu, C. Ma, " Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings," *Measurement*, vol. 158, July. 2020.

[103] P. Liang, C. Deng, J. Wu and Z. Yang, "Intelligent fault diagnosis of rotating machinery via wavelet transform generative adversarial nets and convolutional neural network," *Measurement*, vol. 159, Jul. 2020.

[104] P. Liang, C. Deng, J. Wu, G. Li, Z. Yang, and Y. Wang, “Intelligent fault diagnosis via semisupervised generative adversarial nets and wavelet transform,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4659–4671, Jul. 2020.

[105] H. Tao, P. Wang, Y. Chen, V. Stojanovic, and H. Yang, “An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks,” *J. Franklin Inst.*, vol. 357, no. 11, pp. 7286–7307, Jul. 2020.

[106] S. Shao, P. Wang, and R. Yan, “Generative adversarial networks for data augmentation in machine fault diagnosis,” *Comput. Ind.*, vol. 106, pp. 85–93, Apr. 2019.

[107] Q. Guo, Y. Li, Y. Song, D. Wang, and W. Chen, “Intelligent fault diagnosis method based on full 1-D convolutional generative adversarial network,” *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2044–2053, Mar. 2020.

[108] Z. Shi, J. Chen, Y. Zi and Z. Zhou, "A Novel Multitask Adversarial Network via Redundant Lifting for Multicomponent Intelligent Fault Detection Under Sharp Speed Variation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-10, 2021, Art no. 3511010.

[109] T. Han, C. Liu, W. Yang and D. Jiang, "A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults,"*Knowl.-Based Syst.*, vol. 165, pp. 474-487, Feb. 2019.

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

[111] J. Shao, Z. Huang and J. Zhu, "Transfer Learning Method Based on Adversarial Domain Adaption for Bearing Fault Diagnosis," *IEEE Access*, vol. 8, pp. 119421-119430, 2020.

[112] Z. Chai and C. Zhao, "A Fine-Grained Adversarial Network Method for Cross-Domain Industrial Fault Diagnosis," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 3, pp. 1432-1442, July 2020.

[113] Z. Chen, G. He, J. Li, Y. Liao, K. Gryllias and W. Li, "Domain Adversarial Transfer Network for Cross-Domain Fault Diagnosis of Rotary Machinery," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 11, pp. 8702-8712, Nov. 2020.

[114] F. Li, T. Tang, B. Tang and Q. He, "Deep convolution domain-adversarial transfer learning for fault diagnosis of rolling bearings," *Measurement*, vol. 169, Feb. 2021.

[115] W. Mao, Y. Liu, L. Ding, A. Safian and X. Liang, "A New Structured Domain Adversarial Neural Network for Transfer Fault Diagnosis of Rolling Bearings Under Different Working Conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-13, 2021, Art no. 3509013

[116] Z. Liu, B. Lu, H. Wei, L. Chen, X. Li and M. Rätsch, "Deep Adversarial Domain Adaptation Model for Bearing Fault Diagnosis," *IEEE Trans. Syst. Man Cybern. Syst.*, doi: 10.1109/TSMC.2019.2932000.

[117] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Deep learning-based adversarial multi-classifier optimization for cross-domain machinery fault diagnostics," *Journal of Manufacturing Systems*, vol. 55, pp. 334– 347, 2020.

[118] X. Li, W. Zhang, H. Ma, Z. Luo and X. Li, "Domain generalization in rotating machinery fault diagnostics using deep neural networks," *Neurocomputing*, vol. 403, pp. 409-420, 2020.

[119] Y. Zhang, Z. Ren, S. Zhou and T. Yu, "Adversarial domain adaptation with classifier alignment for cross-domain intelligent fault diagnosis of multiple source domains," *Meas. Sci. Technol.*, vol. 32, no. 3, 2021.

[120] Q. Li, C. Shen, L. Chen and Z. Zhu, "Knowledge mapping-based adversarial domain adaptation: A novel fault diagnosis method with high generalizability under variable working conditions," *Mech. Syst. Signal Process.*, vol. 147, Jan. 2021.

[121] M. Zhang, D. Wang, W. Lu, J. Yang, Z. Li and B. Liang, "A Deep Transfer Model With Wasserstein Distance Guided Multi-Adversarial Networks for Bearing Fault Diagnosis Under Different Working Conditions," *IEEE Access*, vol. 7, pp. 65303-65318, 2019.

[122] J. Jiao, J. Lin, M. Zhao and K. Liang, "Double-level adversarial domain adaptation network for intelligent fault diagnosis," *Knowl.-Based Syst.*, vol. 205, Oct. 2020.

[123] J. Jiao, M. Zhao and J. Lin, "Unsupervised Adversarial Adaptation Network for Intelligent Fault Diagnosis," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9904-9913, Nov. 2020.

[124] C. Cheng, B. Zhou, G. Ma, D. Wu and Y. Yuan, "Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data," *Neurocomputing*, vol. 409, pp. 35-45, Oct. 2020.

[125] X. Wang and F. Liu, "Triplet loss guided adversarial domain adaptation for bearing fault diagnosis," *Sensors*, vol. 20, no. 1, pp. 1-11, Apr. 2020.

[126] Y Zou, Y Liu, J Deng, Y Jiang, W Zhang, "A novel transfer learning method for bearing fault diagnosis under different working conditions," *Measurement*, vol. 171, Feb. 2021.

[127] B Han, X Zhang, J Wang, Z An, S Jia, G Zhang, "Hybrid distance-guided adversarial network for intelligent fault diagnosis under different working conditions," *Measurement*, vol. 176, May 2021.

[128] Y. Wang, X. Sun, J. Li and Y. Yang, "Intelligent Fault Diagnosis With Deep Adversarial Domain Adaptation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-9, 2021, Art no. 2503509.

[129] X. Yu et al., "Conditional Adversarial Domain Adaptation With Discrimination Embedding for Locomotive Fault Diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-12, 2021, Art no. 3503812.

[130] D. She, M. Jia and M. Pecht, "Weighted Entropy Minimization Based Deep Conditional Adversarial Diagnosis Approach under Variable Working Conditions," *IEEE/ASME Trans. Mechatronics*, doi: 10.1109/TMECH.2020.3040175.

[131] M. Ragab et al., "Adversarial Multiple-Target Domain Adaptation for Fault Classification," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021, Art no. 3500211.

[132] Y. Liao, R. Huang, J. Li, Z. Chen and W. Li, "Deep Semisupervised Domain Generalization Network for Rotary Machinery Fault Diagnosis Under Variable Speed," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 8064-8075, Oct. 2020.

[133] S. Pang and X. Yang, "A Cross-Domain Stacked Denoising Autoencoders for Rotating Machinery Fault Diagnosis Under Different Working Conditions," *IEEE Access*, vol. 7, pp. 77277-77292, 2019.

[134] Z. Liu, B. Lu, H. Wei, X. Li and L. Chen, "Fault Diagnosis for Electromechanical Drivetrains Using a Joint Distribution Optimal Deep Domain Adaptation Approach," *IEEE Sensors Journal*, vol. 19, no. 24, pp. 12261-12270, 15 Dec.15, 2019.

[135] Z. Liu, L. Jiang, H. Wei, L. Chen and X. Li, "Optimal Transport-Based Deep Domain Adaptation Approach for Fault Diagnosis of Rotating Machine," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-12, 2021, Art no. 3508912.

[136] L. Wen, L. Gao and X. Li, "A New Deep Transfer Learning Based on Sparse Auto-Encoder for Fault Diagnosis," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 49, no. 1, pp. 136-144, Jan. 2019.

[137] Z. Wan, R. Yang, M. Huang, "Deep Transfer Learning-Based Fault Diagnosis for Gearbox under Complex Working Conditions," *Shock and Vibration*, vol. 2020, Article ID 8884179.

[138] Z. Tang, L. Bo, X. Liu, D. Wei, "An autoencoder with adaptive transfer learning for intelligent fault diagnosis of rotating machinery," *Meas. Sci. Technol.*, 2021, DOI: https://doi.org/10.1088/1361-6501/abd650.

[139] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, May 2019.

[140] Y. Gao, X. Liu, H. Huang, J. Xiang. "A hybrid of FEM simulations and generative adversarial networks to classify faults in rotor-bearing systems," *ISA Trans.*, vol. 108, pp. 356–366, Feb. 2021.

[141] K. Yu, Q. Fu, H. Ma, TR. Lin, X. Li, "Simulation data driven weakly supervised adversarial domain adaptation approach for intelligent cross-machine fault diagnosis," *Structural Health Monitoring*, Jan. 2021. doi:10.1177/1475921720980718

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

[143] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, “Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing,” *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2416–2425, Apr. 2019.

[144] H. Zheng, Y. Yang, J. Yin, Y. Li, R. Wang and M. Xu, "Deep Domain Generalization Combining A Priori Diagnosis Knowledge Toward Cross-Domain Fault Diagnosis of Rolling Bearing," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021, Art no. 3501311.

[145] B. Yang, Y. Lei, S. Xu and C. Lee, "An Optimal Transport-embedded Similarity Measure for Diagnostic Knowledge Transferability Analytics across Machines," IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2021.3095804.

[146] Z. Wu, H. Jiang, K. Zhao and X. Li, "An adaptive deep transfer learning method for bearing fault diagnosis," *Measurement*, vol. 151, Feb. 2020.

[147] P. Wang and R. X. Gao, "Transfer learning for enhanced machine fault diagnosis in manufacturing," *CIRP Ann.*, vol. 69, no. 1, pp. 413-416, 2020.

[148] H. Shao, Z. Ding, J. Cheng, and H. Jiang, “Intelligent fault diagnosis among different rotating machines using novel stacked transfer auto-encoder optimized by PSO,” *ISA Trans.*, vol. 105, pp. 308–319, Oct. 2020.

[149] Xudong Li et al., "Fault diagnostics between different type of components: A transfer learning approach," *Applied Soft Computing*, vol. 86, pp. 105950, 2020.

[150] H. Zhiyi, S. Haidong, J. Lin, C. Junsheng and Y. Yu, "Transfer fault diagnosis of bearing installed in different machines using enhanced deep auto-encoder," *Measurement*, vol. 152, Feb. 2020.

[151] Z. Chen, K. Gryllias and W. Li, "Intelligent Fault Diagnosis for Rotary Machinery Using Transferable Convolutional Neural Network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 339-349, Jan. 2020.

[152] L. Guo, Y. Lei, S. Xing, T. Yan and N. Li, "Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines With Unlabeled Data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316-7325, Sept. 2019.

[153] B. Yang, Y. Lei, F. Jia, N. Li and Z. Du, "A Polynomial Kernel Induced Distance Metric to Improve Deep Transfer Learning for Fault Diagnosis of Machines," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9747-9757, Nov. 2020.

[154] X. Li, X.-D. Jia, W. Zhang, H. Ma, Z. Luo and X. Li, "Intelligent cross-machine fault diagnosis approach with deep auto-encoder and domain adaptation," *Neurocomputing*, vol. 383, pp. 235-247, Mar. 2020.

[155] Z. Wu, H. Jiang, T. Lu, K. Zhao, "A deep transfer maximum classifier discrepancy method for rolling bearing fault diagnosis under few labeled data," *Knowl.-Based Syst.*, vol. 196, May 2020.

[156] C. Zheng, X. Wang, Y. Hao, K. Wang and X. Xiong, "Normalized Recurrent Dynamic Adaption Network: A New Framework With Dynamic Alignment for Intelligent Fault Diagnosis," *IEEE Access*, vol. 8, pp. 80243-80255, 2020.

[157] M. Lv, S. Liu, X. Su and C. Chen, "Deep Transfer Network With Multi-Kernel Dynamic Distribution Adaptation for Cross-Machine Fault Diagnosis," *IEEE Access*, vol. 9, pp. 16392-16409, 2021.

[158] K. Zhao, H. Jiang, Z. Wu, T. Lu, "A novel transfer learning fault diagnosis method based on Manifold Embedded Distribution Alignment with a little labeled data," *Journal of Intelligent Manufacturing*, Sept. 2020, DOI: https://doi.org/10.1007/s10845-020-01657-z

[159] X. Li, H. Jiang, R. Wang, M. Niu, "Rolling bearing fault diagnosis using optimal ensemble deep transfer network," *Knowl.-Based Syst.*, vol. 213, Feb. 2021.

[160] Y. Tan, L. Guo, H. Gao and L. Zhang, "Deep Coupled Joint Distribution Adaptation Network: A Method for Intelligent Fault Diagnosis Between Artificial and Real Damages," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-12, 2021, Art no. 3507212.

[161] J. Chen, J. Wang, J. Zhu, T. H. Lee and C. De Silva, "Unsupervised Cross-domain Fault Diagnosis Using Feature Representation Alignment Networks for Rotating Machinery," *IEEE/ASME Trans. Mechatronics*, doi: 10.1109/TMECH.2020.3046277.

[162] Zhang B, Hong X, Liu Y. Distribution adaptation deep transfer learning method for cross-structure health monitoring using guided waves[J]. Structural Health Monitoring, 2021, DOI: https://doi.org/10.1177/14759217211010709.

[163] Zhang B, Hong X, Liu Y. Multi-task deep transfer learning method for guided wave-based integrated health monitoring using piezoelectric transducers[J]. IEEE Sensors Journal, 2020, 20(23): 14391-14400.

[164] Zhang B, Hong X, Liu Y, Deep Convolutional Neural Network Probability Imaging for Plate Structural Health Monitoring Using Guided Waves[J]. IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-10, 2021, Art no. 2510610.

[165] X. Wang, F. Liu, D. Zhao, "Cross-Machine Fault Diagnosis with Semi-Supervised Discriminative Adversarial Domain Adaptation," *Sensors*, vol. 20, no. 3, pp. 1-16, July 2020.

[166] Y. Feng, J. Chen, Z. Yang, X. Song, Y. Chang, S. He, E. Xu, Z. Zhou, "Similarity-based meta-learning network with adversarial domain adaptation for cross-domain fault identification," *Knowl.-Based Syst.*, vol. 217, Apr. 2021.

[167] J. Zhu, N. Chen and C. Shen, "A New Multiple Source Domain Adaptation Fault Diagnosis Method Between Different Rotating Machines," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2020.3021406.

[168] Liao Y, Huang R, Li J, et al. Dynamic Distribution Adaptation Based Transfer Network for Cross Domain Bearing Fault Diagnosis[J]. Chinese Journal of Mechanical Engineering, 2021, 34(1): 1-10.

[169] N. Lu, T. Yin, "Transferable common feature space mining for fault diagnosis with imbalanced data," *Mech. Syst. Signal Process.*, Vol. 156, July 2021.

[170] D. Xiao, Y. Huang, C. Qin, Z. Liu, Y. Li and C. Liu, "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.*, vol. 233, no. 14, pp. 5131-5143, Jul. 2019.

[171] X. Li, H. Jiang, K. Zhao and R. Wang, "A Deep Transfer Nonnegativity-Constraint Sparse Autoencoder for Rolling Bearing Fault Diagnosis With Few Labeled Data," *IEEE Access*, vol. 7, pp. 91216-91224, 2019.

[172] H. Kim and B. D. Youn, "A New Parameter Repurposing Method for Parameter Transfer With Small Dataset and Its Application in Fault Diagnosis of Rolling Element Bearings," *IEEE Access*, vol. 7, pp. 46917-46930, 2019.

[173] Z. He, H. Shao, P. Wang, J. Lin, J. Cheng and Y. Yang, "Deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearbox with few target training samples," *Knowl.-Based Syst.*, vol. 191, Mar. 2020.

[174] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," *Measurement*, vol. 152, Feb. 2020.

[175] L. Zou, Y. Li and F. Xu, "An adversarial denoising convolutional neural network for fault diagnosis of rotating machinery under noisy environment and limited sample size case," *Neurocomputing*, vol. 407, pp. 105-120, Sep. 2020.

[176] Y. Zhang, Z. Ren and S. Zhou, "An Intelligent Fault Diagnosis for Rolling Bearing Based on Adversarial Semi-Supervised Method," *IEEE Access*, vol. 8, pp. 149868-149877, 2020.

[177] X. Li, W. Zhang, Q. Ding and X. Li, "Diagnosing Rotating Machines With Weakly Supervised Data Using Deep Transfer Learning," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1688-1697, March 2020.

[178] M Zareapoor, P Shamsolmoali, J Yang, "Oversampling adversarial network for class-imbalanced fault diagnosis," *Mech. Syst. Signal Process.*, Vol. 149, Feb. 2021.

[179] T. Zhang, J. Chen, F. Li, T. Pan and S. He, "A Small Sample Focused Intelligent Fault Diagnosis Scheme of Machines via Multi-modules Learning with Gradient Penalized Generative Adversarial Networks," *IEEE Trans. Ind. Electron.*, doi: 10.1109/TIE.2020.3028821.

[180] C. Li, D. Cabrera, F. Sancho, R. -V. Sanchez, M. Cerrada and J. Valente de Oliveira, "One-shot fault diagnosis of 3D printers through improved feature space learning," *IEEE Trans. Ind. Electron.*, doi: 10.1109/TIE.2020.3013546.

[181] X. Li, W. Zhang, N. Xu and Q. Ding, "Deep Learning-Based Machinery Fault Diagnostics With Domain Adaptation Across Sensors at Different Places," *IEEE Trans. Ind. Electron.*, vol. 67, no. 8, pp. 6785-6794, Aug. 2020.

[182] S Siahpour, X Li, J Lee, "Deep learning-based cross-sensor domain adaptation for fault diagnosis of electro-mechanical actuators," Int. J. Dynam. Control, vol. 8, pp. 1054–1062, July 2020.

[183] V. Pandhare, X. Li, M. Miller, X. Jia and J. Lee, "Intelligent Diagnostics for Ball Screw Fault Through Indirect Sensing Using Deep Domain Adaptation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021, Art no. 2504211.

[184] Y. Qin, X. Wang, Q. Qian, H. Pu and J. Luo, "Multi-scale transfer voting mechanism: a new strategy for domain adaption," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2020.3045392.

[185] S. Wu, X. Jing, Q. Zhang, F. Wu, H. Zhao and Y. Dong, "Prediction Consistency Guided Convolutional Neural Networks for Cross-Domain Bearing Fault Diagnosis," *IEEE Access*, vol. 8, pp. 120089-120103, 2020

[186] Z. Cao, M. Long, J. Wang, and M. I. Jordan, "Partial transfer learning with selective adversarial networks," *in Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2724–2732.

[187] J. Jiao, M. Zhao, J. Lin, and C. Ding, "Classifier inconsistency-based domain adaptation network for partial transfer intelligent diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5965–5974, Sep. 2020.

[188] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Partial transfer learning in machinery cross-domain fault diagnostics using class-weighted adversarial networks," *Neural Netw.*, vol. 129, pp. 313–322, Sep. 2020.

[189] W. Li, Z. Chen and G. He, "A Novel Weighted Adversarial Transfer Network for Partial Domain Fault Diagnosis of Machinery," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1753-1762, March 2021.

[190] Z. Liu, B. Lu, H. Wei, L. Chen, X. Li and C. Wang, "A Stacked Auto-Encoder Based Partial Adversarial Domain Adaptation Model for Intelligent Fault Diagnosis of Rotating Machines," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2020.3045002.

[191] X. Li and W. Zhang, "Deep Learning-Based Partial Domain Adaptation Method on Intelligent Machinery Fault Diagnostics," *IEEE Trans. Ind. Electron.*, vol. 68, no. 5, pp. 4351-4361, May 2021.

[192] T. Han, C. Liu, R. Wu, D. Jiang, "Deep transfer learning with limited data for machinery fault diagnosis," *Applied Soft Computing*, vol. 103, pp. 107150, 2020.

[193] Y. Deng, D. Huang, S. Du, G. Li, C. Zhao, J. Lv, "A double-layer attention based adversarial network for partial transfer learning in machinery fault diagnosis," *Comput. Ind.*, Vol. 127, May 2021, 103399.

[194] B. Yang, CG. Lee, Y. Lei, N. Li, N. Lu, "Deep partial transfer learning network: A method to selectively transfer diagnostic knowledge across related machines," *Mech. Syst. Signal Process.*, vol. 116, July 2021.

[195] Q. Wang, G. Michau and O. Fink, "Missing-Class-Robust Domain Adaptation by Unilateral Alignment," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 663-671, Jan. 2021.

[196] S. Zhang, M. Wang, W. Li, J. Luo, and Z. Lin, "Deep learning with emerging new labels for fault diagnosis," *IEEE Access*, vol. 7, pp. 6279–6287, 2019.

[197] C. Wang, C. Xin, Z. Xu, "A novel deep metric learning model for imbalanced fault diagnosis and toward open-set classification," *Knowl.-Based Syst.*, vol. 220, May 2021.

[198] L. Feng and C. Zhao, "Fault Description Based Attribute Transfer for Zero-Sample Industrial Fault Diagnosis," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1852-1862, March 2021.

[199] J. Li, R. Huang, W. Li, "Intelligent Fault Diagnosis for Bearing Dataset Using Adversarial Transfer Learning based on Stacked Auto-Encoder," *Procedia Manufacturing*, Vol. 49, pp. 75-80, 2020.

[200] J. Li, R. Huang, G. He, S. Wang, G. Li and W. Li, "A Deep Adversarial Transfer Learning Network for Machinery Emerging Fault Detection," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8413-8422, 1 Aug.1, 2020.

[201] H. Wang, X. Bai, J. Tan, J. Yang, "Deep prototypical networks based domain adaptation for fault diagnosis," *Journal of Intelligent Manufacturing*, 2020, https://doi.org/10.1007/s10845-020-01709-4

[202] W. Zhang, X. Li, H. Ma, Z. Luo and X. Li, "Open Set Domain Adaptation In Machinery Fault Diagnostics Using Instance-Level Weighted Adversarial Learning," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2021.3054651.

[203] W. Zhang, X. Li, H. Ma, Z. Luo and X. Li, "Universal Domain Adaptation in Fault Diagnostics with Hybrid Weighted Deep Adversarial Learning," IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3064377.

[204] Yang B, Xu S, Lei Y, et al. Multi-source transfer learning network to complement knowledge for intelligent diagnosis of machines with unseen faults[J]. Mechanical Systems and Signal Processing, 2022, 162: 108095.

[205] J. Li, R. Huang, G. He, Y. Liao, Z. Wang, and W. Li, "A two-stage transfer adversarial network for intelligent fault diagnosis of rotating machinery with multiple new faults," *IEEE/ASME Trans. Mechatronics*, early access, Sep. 22, 2020, doi: 10.1109/TMECH.2020.3025615.

[206] X. Yu et al., "Deep Learning-Based Open Set Fault Diagnosis by Extreme Value Theory," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2021.3070324.

[207] P. P. Busto and J. Gall, “Open set domain adaptation,” *in Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 754–763.

[208] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, "Open set domain adaptation by backpropagation," *in Proc. ECCV*, 2018, pp. 153–168.

[209] D. Zhao, W. Cheng, R. X. Gao, R. Yan, and P. Wang, “Generalized Vold– Kalman filtering for nonstationary compound faults feature extraction of bearing and gear,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 401–410, Feb. 2020.

[210] G. Tang, Y. Wang, Y. Huang, N. Liu, and J. He, “Compound bearing fault detection under varying speed conditions with virtual multichannel signals in angle domain,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 8, pp. 5535–5545, Aug. 2020.

[211] N. Li, W. Huang, W. Guo, G. Gao, and Z. Zhu, “Multiple enhanced sparse decomposition for gearbox compound fault diagnosis,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 770–781, Mar. 2020.

[212] L. Cui, Y. Sun, J. Zhang, H. Wang, “Adapted dictionary-free orthogonal matching pursuit and 0-1 programming to solve the isolation and diagnosis of bearing and gear compound faults,” *Measurement*, Mar. 2021. Art. no. 109331.

[213] Shao H, Jiang H, Zhang H, et al. Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing[J]. Mechanical Systems and Signal Processing, 2018, 100: 743-765.

[214] Wang S, Xiang J, Zhong Y, et al. A data indicator-based deep belief networks to detect multiple faults in axial piston pumps[J]. Mechanical Systems and Signal Processing, 2018, 112: 154-170.

[215] Zhao X, Jia M, Liu Z. Semisupervised Graph Convolution Deep Belief Network for Fault Diagnosis of Electormechanical System With Limited Labeled Data[J]. IEEE Transactions on Industrial Informatics, 2020, 17(8): 5450-5460.

[216] Shao H, Lin J, Zhang L, et al. A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance[J]. Information Fusion, 2021, 74: 65-76.

[217] Sun G, Wang Y, Sun C, et al. Intelligent detection of a planetary gearbox composite fault based on adaptive separation and deep learning[J]. Sensors, 2019, 19(23): 5222.

[218] Lin M C, Han P Y, Fan Y H, et al. Development of Compound Fault Diagnosis System for Gearbox Based on Convolutional Neural Network[J]. Sensors, 2020, 20(21): 6169.

[219] Zhang D, Ren X, Zuo H. Compound Fault Diagnosis for Gearbox Based Using of Euclidean Matrix Sample Entropy and One-Dimensional Convolutional Neural Network[J]. Shock and Vibration, 2021, 2021.

[220] Xue Y, Dou D, Yang J. Multi-fault diagnosis of rotating machinery based on deep convolution neural network and support vector machine[J]. Measurement, 2020, 156: 107571.

[221] P. Liang, C. Deng, J. Wu, Z. Yang, J. Zhu, and Z. Zhang, "Single and simultaneous fault diagnosis of gearbox via a semi-supervised and high-accuracy adversarial learning framework," *Knowl.-Based Syst.*, vol. 198, Jun. 2020, Art. no. 105895.

[222] R. Huang, W. Li, and L. Cui, "An intelligent compound fault diagnosis method using one-dimensional deep convolutional neural network with multi-label classifier," in Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC), Auckland, New Zealand, May 2019, pp. 1–6.

[223] P. Liang, C. Deng, J. Wu, Z. Yang, J. Zhu, and Z. Zhang, ‘‘Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform,’’ *Comput. Ind.*, vol. 113, Dec. 2019, Art. no. 103132.

[224] Y. Jin, C. Qin, Y. Huang, and C. Liu, “Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network,” *Measurement*, vol. 173, Mar. 2021. Art. no. 108500.

[225] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” 2017, arXiv:1710.09829. [Online]. Available: http://arxiv.org/abs/1710.09829

[226] R. Huang, Y. Liao, S. Zhang, and W. Li, "Deep decoupling convolutional neural network for intelligent compound fault diagnosis," *IEEE Access*, vol. 7, pp. 1848–1858, 2019.

[227] R. Huang, J. Li, W. Li, and L. Cui, "Deep ensemble capsule network for intelligent compound fault diagnosis using multisensory data," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2304–2314, May 2020.

[228] A. Dibaj, MM. Ettefagh, R. Hassannejad, M. Ehghaghi, "A hybrid fine-tuned VMD and CNN scheme for untrained compound fault diagnosis of rotating machinery with unequal-severity faults," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114094.

[229] Xing S, Lei Y, Wang S, et al. A label description space embedded model for zero-shot intelligent diagnosis of mechanical compound faults[J]. Mechanical Systems and Signal Processing, 2022, 162: 108036.

[230] R. Huang, Z. Wang, J. Li, J. Chen, and W. Li, "A transferable capsule network for decoupling compound fault of machinery," *in Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, Dubrovnik, Croatia, May 2020, pp. 1–6.

[231] R. Huang, J. Li, Y. Liao, J. Chen, Z. Wang and W. Li, "Deep Adversarial Capsule Network for Compound Fault Diagnosis of Machinery Toward Multidomain Generalization Task," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021, Art no. 3506311.

[232] Z. Chen, A. Mauricio, W. Li, and K. Gryllias, ‘‘A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks,’’ *Mech. Syst. Signal Process.*, vol. 140, Jun. 2020, Art. no. 106683.

[233] D. Zhu, X. Cheng, L. Yang, Y. Chen and S. X. Yang, "Information Fusion Fault Diagnosis Method for Deep-Sea Human Occupied Vehicle Thruster Based on Deep Belief Network," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2021.3055770.

[234] L. Bottou, "Large-scale machine learning with stochastic gradient descent," *in Proc. 19th Int. Conf. Comput. Statist.*, 2010, pp. 177–186.

[235] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.

[236] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *in Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–41.

[237] X. Yuan, P. He, Q. Zhu and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805-2824, Sept. 2019.

[238] M. Hutson, “Artificial intelligence faces reproducibility crisis,” *Science*, vol. 359, no. 6377, pp. 725–726, 2018.

[239] Z. C. Lipton, “The mythos of model interpretability,” *in Proc. Int. Conf. Mach. Learn. (ICML) Workshop*, 2016, pp. 1–9.

[240] A. Rasheed, O. San and T. Kvamsdal, "Digital Twin: Values, Challenges and Enablers From a Modeling Perspective," *IEEE Access*, vol. 8, pp. 21980-22012, 2020.

[241] C. Molnar. (2019). Interpretable Machine Learning. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[242] Zhou Z, Li T, Zhang Z, et al. Bayesian Differentiable Architecture Search for Efficient Domain Matching Fault Diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-11.

[243] Qiang Yang; Yang Liu; Yong Cheng; Yan Kang; Tianjian Chen; Han Yu, Federated Learning, Morgan & Claypool, 2019, doi: 10.2200/S00960ED2V01Y201910AIM043.

[244] Z. Mao, M. Xia, B. Jiang, D. Xu and P. Shi, "Incipient Fault Diagnosis for High-Speed Train Traction Systems via Stacked Generalization," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2020.3034929.

[245] T. T. Nguyen, N. D. Nguyen and S. Nahavandi, "Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3826-3839, Sept. 2020.

# A Perspective Survey on Deep Transfer Learning for Fault Diagnosis in Industrial Scenarios: Theories, Applications and Challenges

Weihua Li [a, b], Ruyi Huang [a, c], Jipu Li [c], Yixiao Liao [c], Zhuyun Chen [c], Guolin He [c], Ruqiang Yan [d], Konstantinos Gryllias [e]

[a] Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 510640, China

[b] PazhouLab, Guangdong Artificial Intelligence and Digital Economy Laboratory, Guangzhou 510335, China

[c] School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510640, China

[d] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China

[e] Department of Mechanical Engineering, KU Leuven, Celestijnenlaan 300, Leuven 3001, Belgium

***Abstract***—Deep Transfer Learning (DTL) is a new paradigm of machine learning, which can not only leverage the advantages of Deep Learning (DL) in feature representation, but also benefit from the superiority of Transfer Learning (TL) in knowledge transfer. As a result, DTL techniques can make DL-based fault diagnosis methods more reliable, robust and applicable, and they have been widely developed and investigated in the field of Intelligent Fault Diagnosis (IFD). Although several systematic and valuable review articles have been published on the topic of IFD, they summarized relevant research only from an algorithm perspective and overlooked practical applications in industry scenarios. Furthermore, comprehensive review on DTL-based IFD methods is still lacking. From this insight, it is particularly important and more necessary to comprehensively survey the relevant publications of DTL-based IFD with the goal of helping readers to conveniently understand the current state-of-the-art techniques and to quickly design an effective solution for solving IFD problems in practice. First, theoretical backgrounds of DTL are briefly introduced to present how the transfer learning techniques can be integrated with deep learning models. Then, major applications of DTL and their recent developments in the field of IFD are detailed and discussed. More importantly, suggestions on how to select DTL algorithms for IFD in practical applications, and some future challenges and research trends are shared. Finally, conclusions of this survey are given. As last, we have reason to believe that the works done in this article can provide convenience and inspiration for the researchers who want to devote his/her efforts in the progress and advance of IFD.

***Keywords***—Fault Diagnosis, Deep Learning, Transfer Learning, Domain Adaptation, Deep Transfer Learning

### Abbreviations (Abbr.)

| Abbr. | Terminology | Abbr. | Terminology |
|---|---|---|---|
| **Adagrad** | Adaptive Gradient | **IMS** | Intelligent Maintenance Systems |
| **Adam** | Adaptive Moment Estimation | **JDA** | Joint Distribution Adaptation |
| **ADDA** | Adversarial Discriminative Domain Adaptation | ***k*-NN** | k-Nearest Neighbors |
| **AI** | Artificial Intelligence | **KL** | Kullback-Leibler |
| **ANN** | Artificial Neural Network | **LSTM** | Long short-term memory |
| **CNs** | Capsule Networks | **MMD** | Maximum Mean Discrepancy |
| **CWRU** | Case Western Reserve University | **MVD** | Maximum Variance Discrepancy |
| **CMD** | Central Moment Discrepancy | **MAE** | Mean Absolute Error |
| **CMS** | Condition Monitoring System | **OSDA** | Open Set Domain Adaptation |
| **CMMD** | Conditional Maximum Mean Discrepancy | **PSO** | Particle Swarm Optimization |
| **CNNs** | Convolutional Neural Networks | **PK-MMD** | polynomial kernel induced MMD |
| **CORAL** | Correlation Alignment | **PHM** | Prognostics and Health Management |
| **DACN** | Deep Adversarial Capsule Network | **RNNs** | Recurrent Neural Networks |
| **DADAN** | Deep Adversarial Domain Adaptation Network | **RKHS** | Reproducing Kernel Hilbert Space |
| **DBNs** | Deep Belief Networks | **RMSE** | Root Mean Square Error |
| **DBM** | Deep Boltzmann Machines | **SAN** | Selective Adversarial Network |
| **DDCNN** | Deep Decoupling Convolutional Neural Network | **SAE** | Sparse Auto-Encoder |
| **DL** | Deep Learning | **SGD** | Stochastic Gradient Descent |
| **DNNs** | Deep Neural Networks | **SVM** | Support Vector Machine |
| **DTL** | Deep Transfer Learning | **TrAdaBoost** | Transfer Adaptive Boosting |
| **DANN** | Domain Adversarial Neural Network | **TCA** | Transfer Component Analysis |
| **GK-MMD** | Gaussian kernel induced MMD | **TL** | Transfer Learning |
| **GANs** | Generative Adversarial Networks | **TCNN** | Transferable Convolutional Neural Network |
| **GRL** | Gradient Reversal Layer | **WATN** | Weighted Adversarial Transfer Network |
| **GNNs** | Graph Neural Networks | **WCs** | Working Conditions |
| **IFD** | Intelligent Fault Diagnosis | **WD** | Wasserstein Distances |

## 1. Introduction

Powerfully driven by advanced computing, sensing, measuring and communicating technologies, the manufacturing industry is characterized by an irresistible trend from automatic to digital and to intelligent, and it has embraced the new era of the fourth industrial revolution (Industry 4.0), whose ultimate goal is to make precise self-perception, to enable autonomous decision-making, and to realize intelligent networking for machines during the process of manufacturing [1]-[3]. Industrial equipment (IE), one of the most crucial carriers for manufacturing industry in such trend and revolution, has been devoting itself to generating economic benefits such as quality improvement, efficiency enhancement, energy conservation and cost reduction. Meanwhile, the IE is typically asked to accomplish the herculean tasks that often have harsh operating environment and need providing long-term services [4]-[6]. To ensure the safety and reliability of the industrial environment, the health status of IE has to be monitored and diagnosed in time, which can reduce equipment downtime, formulate scheduled maintenance, increase economic benefits, and avoid tragic catastrophes [7], [8]. Because of the complexity and dynamicity associated with the manufacturing processes, which inevitably leads to degradation, failure and damage, how to precisely make fault diagnosis for IE in time was and remains a great challenge.

In past decades, more and more attention has been paid to timely and precise IFD from academics and industry researchers since it has been listed as a key concern by many governments and organizations. Fortunately, owing to the rapid development of Artificial Intelligence (AI) technologies, especially in deep learning and transfer learning, abundant intelligent algorithms have been developed by researchers and engineers to address various practical problems in industrial scenarios, and have also brought successful breakthroughs for intelligent fault diagnosis (IFD) of IE.

Deep learning, a branch of machine learning in AI, broadly refers to methods that utilize hierarchical architectures, such as Deep Neural Networks (DNNs), Deep Belief Networks (DBNs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) [9]-[13], to learn higher-level representations from raw inputs which are images of pixel data, files of audio data, documents of text data, etc., [14]. On one hand, deep learning technology has been proven to be a promising tool in many applications of manufacturing industry due to its phenomenal advantages in massive data processing, discriminative feature learning and effective pattern recognition, constructing intelligent models by mapping relationships between health conditions of IE and industrial data in an end-to-end way [15]-[18]. On the other hand, deep learning technology has limitations which inhibit its further progress, advance and application in complex real-world scenarios. The ideal and hypothetical application scenarios of deep learning present the following characteristics:

(1) Deep learning requires abundant labeled samples in advance for model training. One of the limitations of deep learning methods is that they learn how to perform tasks through observations. That is to say, deep learning methods heavily rely on large amounts of labeled training data, without which these methods are prone to overfitting and will lack a robust generalization performance.

(2) Deep learning has strict requirements for the distributions between training and testing data. If a deep learning model is trained on data that present distributions discrepancy with the target data, the performance of the model will decrease dramatically and even will not work.

Considering the practical applications in many industrial scenarios, however, it is time-consuming, labor-intensive and even unrealistic to collect sufficient labeled data, especially labeled fault data, because the IE is always kept in a good status with time- or condition-based schedule maintenance. More importantly, it is often the case that the IE operates at harsh, varying and complex environments, which makes the distributions of the data in future testing situations different from that of the data of the pretrained model.

Transfer learning, another branch of machine learning that focuses on learning common knowledge from one or more related but different application scenarios to help AI algorithms to obtain more powerful performance in an application scenario of interest, has been demonstrated as a promising methodology for helping deep learning to overcome the limitations mentioned above [19]. In analogy with the ability of human beings that can leverage only a few examples or previous experience to help tackle unforeseeable problems, transfer learning can endow an AI model with better learning performance even when training data is sparse and limited, and with robust generalization performance from the related but different application scenarios to a new one [20]. However, the traditional machine learning approaches might not be able to learn the discriminative representations in an effective way, which is a major roadblock for fulfilling the potential of transfer learning.

Combining the advantages of deep learning in feature representation and the benefits of transfer learning in knowledge transfer, Deep Transfer Learning (DTL), a new paradigm of machine learning developed in recent years, leverages deep learning technology for transfer learning, which can learn hidden transferable knowledge and capture complex patterns more effectively [21]. DTL would be better preferred in practical application scenarios for manufacturing industry because it can be easier integrated with deep learning models that are widely developed for IFD of IE and can make the deep-learning-based methods more reliable, robust and accessible [22], [23].

The goal of this survey is to offer an in-depth overview of DTL for fault diagnosis in industrial scenarios, which can provide a comprehensive guidance for the readers who want to devote his/her efforts in the progress and advance of IFD. Historically, several systematic review articles have been published on the topic of fault diagnosis. For instance, Jay Lee, the founding director of the National Science Foundation Industry/University Cooperative Research Center (NSF I/UCRC) for Intelligent Maintenance Systems (IMS), conducted a comprehensive overview for Prognostics and Health Management (PHM) of rotary machinery systems from designing PHM methodology to selecting appropriate algorithms and to making accurate diagnosis decision, in 2014 [24]. That literature review placed much emphasis on the traditional fault diagnosis and prognosis algorithms, which cannot reflect the state-of-the-art techniques at present. Chen *et al.* gave a broad comprehensive literature survey of AI algorithms in the fault diagnosis of rotating machinery from the aspect of theory and application, in 2018 [25], which mainly focuses in the following algorithms: k-Nearest Neighbors (k-NN), Naive Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN) and Deep Neural Network (DNN). Yan and Gao summarized the deep learning-based research work published before 2019 for machine health monitoring [26], in which the popular deep learning models, such as Sparse Auto-Encoder (SAE), Deep Belief Network (DBN), Deep Boltzmann Machines (DBM), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have been systematically reviewed and the corresponding data and codes have been opened for replicating the reported results. Lei and Nandi presented a review and roadmap for IFD methods based on machine learning [27], in which the development of IFD methods was divided into three periods: the past (traditional machine learning), present (deep learning) and future (transfer learning). Besides the survey articles mentioned above, there are some other related articles which focus on the deep learning-based [28], transfer learning-based [29], convolutional neural network-based [30], [31], AI-Enabled-based [32] or special components-based [33]-[35] methods for machine fault diagnosis, which will not be enumerated here in detail. Admittedly, as the valuable and systematic scholarly sources on the IFD, these literature reviews have contributed to the development of fault diagnosis from many aspects and guided the researchers towards a clearer future direction [36], [37]. Nevertheless, there are still some aspects that have not yet been comprehensively summarized by the previous literature articles:

(1) The historical reviews mainly concerned IFD on either traditional machine learning, deep learning or transfer learning. As a new promising tool to solve the problems faced by the engineer and researcher in the real industrial scenarios, there is still a lack of systematic review on DTL-based IFD methods.

(2) Throughout the above discussion, it is clear that almost all the reviews categorized the relevant research from the perspective of algorithm and analytic technology, resulting in the difficulty to select appropriate algorithms for engineers in specific industrial applications.

Therefore, it is particularly important and more necessary to overview the relevant publications of DTL with the goal of helping readers to conveniently understand the current state-of-the-art techniques related to IFD and to quickly design an effective solution for some challenges in practice.

To overcome the limitation forementioned, this review article attempts to provide a comprehensive survey on DTL for fault diagnosis in industrial scenarios. First, different from the existing review articles which mainly focused on the IFD methods using either traditional machine learning, deep learning or transfer learning, this review article aims at focusing on the IFD methods using the new paradigm of machine learning, i.e., DTL. Second, in contrast to the existing review articles summarized the related publications from the algorithm perspective, this review article categorizes the DTL-based IFD methods from the perspective of practical industrial scenarios, which could provide suggestions to select appropriate algorithms for engineers in specific industrial applications. Last, the existing review articles mainly covered the related publication before 2020. It is the fact that there are many new articles have been published in recent years since the IFD have attracted lots of attentions from both academics and industry researchers, but, by contrast, this review article has included most of the state-of-the-art DTL techniques before it summitted to the Journal.

The main contributions of this article are outlined as the following three aspects:

(1) Basic concepts and theories of DTL are introduced, including instance-based DTL, model-based DTL and feature-based DTL, which can present a comprehensive overview about the DTL from the algorithm perspective.

(2) Applications of DTL approaches are summarized into four categories from the perspective of practical industrial scenarios, and each category in IFD are detailed, which would be instructive for engineers in specific industrial applications to select the appropriate algorithms.

(3) Future challenges and potential directions of DTL for IFD are concluded, attempting to provide new insights on the future works for potential newcomers and seasoned researchers.

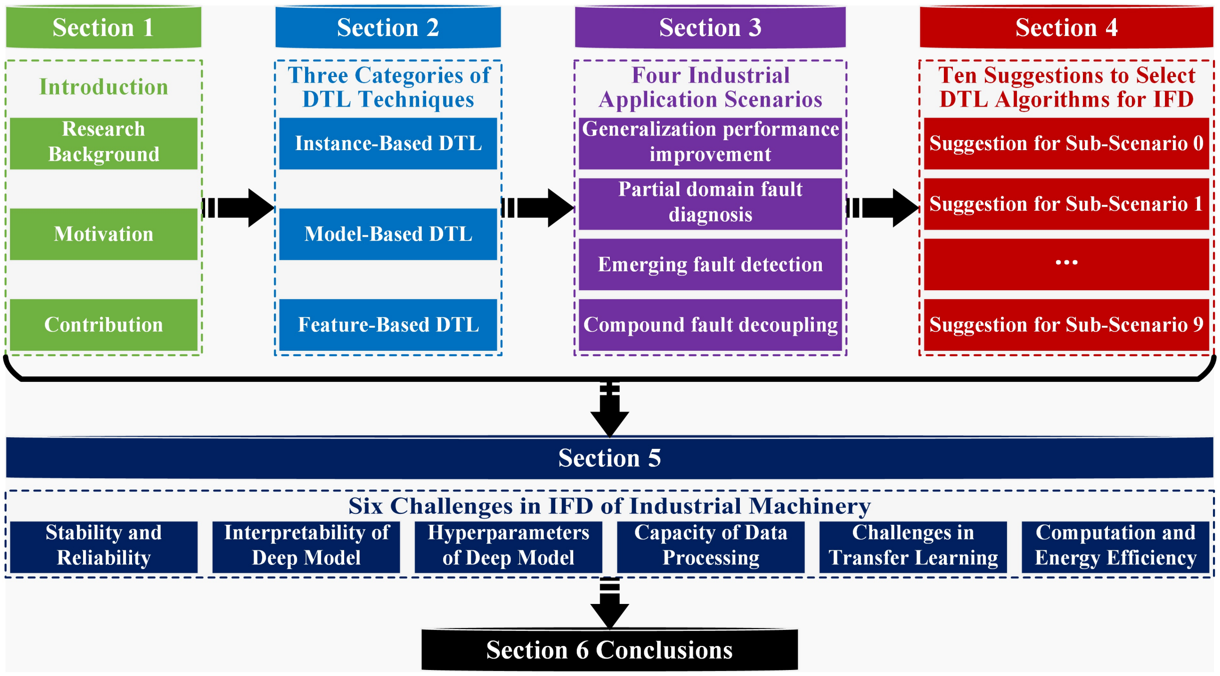

Fig. 1. Flow chart showing the overall logic of this literature review

As show in Fig. 1, the rest of this review article is organized as follows. In Section 2, the theoretical backgrounds of DTL, including basic definition of DTL and three categories of DTL, are briefly introduced to present how the transfer learning techniques can be integrated with deep learning models. Section 3 details the major applications of DTL and its recent developments in the field of IFD from the perspective of practical industrial scenarios, in which four application scenarios are formulated according to the task of fault classification. More importantly, suggestions that on how to select DTL algorithms for IFD in practical industrial applications and some future challenges are shared in Section 4 and 5, respectively. Finally, the conclusions of this survey are given in Section 6.

## 2. Theoretical Backgrounds of DTL

In this section, the basic definitions related to DTL are firstly introduced for convenience. The theories of DTL that explain how the transfer learning technologies can leverage the powerful representation ability of deep learning to extract and transfer knowledge, are summarized from the perspective of algorithm to allow readers understand the mechanisms and strategies of DTL approaches.

### *2.1 Basic Definitions of DTL*

According to the book of Transfer Learning [38], some basic definitions related to this survey, such as Domain, Task and Transfer Learning, are listed as follows:

***Domain***, denoted as $\mathcal{D}=\{\mathcal{X},P(X)\}$, consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X=\{x \mid x_i \in \mathcal{X}, i=1,\cdots,N\}$ is a dataset that contains $N$ instances. Generally, different domains are defined based on the fact that there are different feature spaces or different marginal probability distributions between these domains. In the scenarios of machinery fault diagnosis, different working conditions (WCs), locations and machines can be regarded as different domains.

***Task***, denoted as $\mathcal{T}=\{\mathcal{Y},f(\bullet)\}$ when giving a specific domain $\mathcal{D}$, consists of two components: a label space $\mathcal{Y}$ and a mapping function $f(\bullet)$, where $Y=\{y \mid y_i \in \mathcal{Y}, i=1,\cdots,N\}$ is a label set for the corresponding instances in $\mathcal{D}$. The mapping function $f(\bullet)$, also denoted as $f(\boldsymbol{x})=P(y \mid \boldsymbol{x})$, is a non-linear and implicit function that can bridge the relationship between the input instance and the predicted decision, which is expectedly learned from the given datasets. Similarly, different tasks are defined as there are different label spaces between these tasks. Different fault classes and types can be regarded as different tasks.

***Transfer Learning***, given a source domain $\mathcal{D}^S=\{\mathcal{X}^S,P^S(X^S)\}$ with the source task $\mathcal{T}^S=\{\mathcal{Y}^S,f^S(\bullet)\}$ and a target domain $\mathcal{D}^T=\{\mathcal{X}^T,P^T(X^T)\}$ with the target task $\mathcal{T}^T=\{\mathcal{Y}^T,f^T(\bullet)\}$, aims to learn a better mapping function $f^T(\bullet)$ for the target task $\mathcal{T}^T$ with the transferable knowledge gained from the source domain $\mathcal{D}^S$ and task $\mathcal{T}^S$. Contrary to the tradition machine learning and deep learning in which the domain and task between the source and target scenarios are identical (that is, $\mathcal{D}^S=\mathcal{D}^T$ and $\mathcal{T}^S=\mathcal{T}^T$), the transfer learning counters the problems where the domain and/or the task between the source and the target scenarios could be different (i.e., $\mathcal{D}^S \neq \mathcal{D}^T$ and/or $\mathcal{T}^S \neq \mathcal{T}^T$).

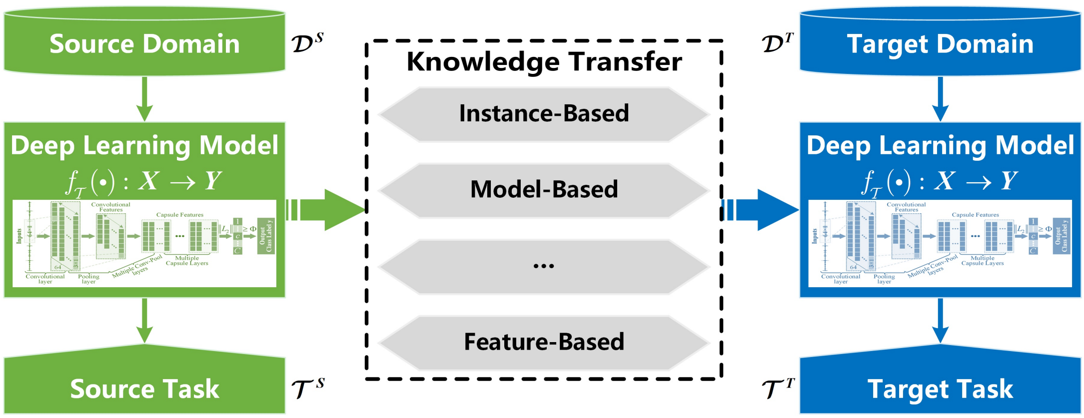


Fig. 2. An illustration of DTL

Based on the definition mentioned above, a definition of deep transfer learning can be formulated as: Given a transfer learning task $f^{S\rightarrow T}(\bullet): \boldsymbol{X}^T \rightarrow \boldsymbol{Y}^T$ based on $[\mathcal{D}^S, \mathcal{D}^T, \mathcal{T}^S, \mathcal{T}^T]$, deep transfer learning aims to learn the mapping function $f^{S\rightarrow T}(\bullet)$ by leveraging the powerful deep learning model, that is, deep neural networks, in which the transfer learning technique and the deep learning model can be integrated to a more robust AI method.

## *2.2 Categorization of DTL*

Fig. 2 shows a typical concept of DTL process that is capable of transferring the valuable knowledge by further exploiting the representation learning ability of deep neural networks. The literature on deep learning or transfer learning has gone through a considerable number of iterative updates. In contrast, few literatures focus on deep transfer learning as a new emerging technique. There is no mutual consensus on how to classify the categorization of DTL. According to the survey published by Tan et al. [21], the DTL approaches have been divided into four categories, that is, instances-based, mapping-based, network-based and adversarial-based DTL. However, these types of DTL approaches are associated and inter-related with each other, which makes it difficult to be well-categorized.

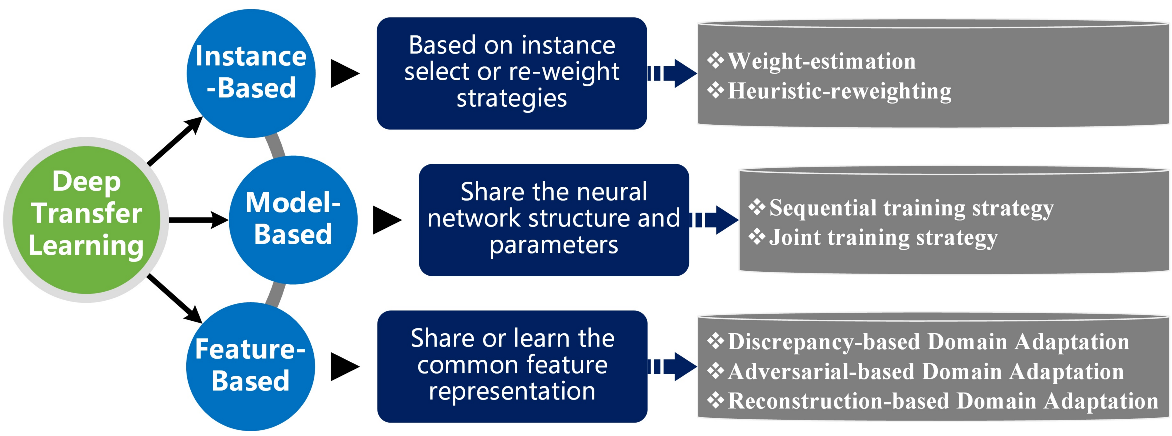


Fig. 3. Categorization of DTL

In this survey, from a viewpoint of the mechanism with which the deep learning model bridge the generalization errors between target and source domains by leveraging the transfer learning techniques, the DTL approaches are divided into three groups: instance-based, model-based and feature-based DTL. The categorization of DTL is illustrated in Fig. 3. Instance-based DTL approaches are typically based on instance select or re-weight strategies. Model-based DTL approaches mainly share the neural network structure and parameters between target and source domains. Feature-based DTL approaches share or learn the common feature representation between target and source domain. In the following parts of this section, the theorical backgrounds of each category of DTL will be introduced from the perspective of algorithm in detail.

### *2.2.1 Instance-based DTL*

Instance-Based DTL aims to train a more precise deep model under a transfer scenario where the difference between source and target domains/tasks only comes from either the different marginal probability distribution, i.e., $P^S(X^S) \neq P^T(X^T)$, or the conditional probability distribution, i.e., $P^S(Y^S \mid X^S) \neq P^T(Y^T \mid X^T)$, which also assumes that the labeled instances in the target domain are too limited to train a satisfied diagnosis model. An intuitive motivation behind instance-based DTL approaches is that directly merging the source data into the target data might deteriorate the performance of the target deep model and result in a negative transfer during the model training because some labeled instances in the source domain are significantly different from the target domain ones. Inspired by such motivation, the goal of the instance-based DTL approaches is to single out the instances in the source domain that are positive for target model training and to augment the target data by adapting the instance weighting strategies. A promising solution in terms of deep learning models is to automatically learn the instances weights of the source domain in the objective function. The general objective function of an instance-based DTL task can be formulated as

$$\mathcal{L} = \frac{1}{C^S}\sum_{i=1}^{N^S}\omega_i \mathcal{R}^S\left(f(\boldsymbol{x}_i^S), y_i^S\right) + \mathcal{R}^*\left(f(\boldsymbol{X}), \boldsymbol{Y}\right) \tag{1}$$

where $\omega_i$ is the weighting coefficient of the corresponding source instance, $C^S = \sum_{i=1}^{N^S}\omega_i$, $\mathcal{R}^S$ represents the risk function of selecting the source instance, and $\mathcal{R}^*$ denotes the second risk function related to the target task or the parameter regularization. The theoretical value of $\omega_i$ is defined as the ratio of the marginal probability distributions between the target domain and the source domain at the input instance $\boldsymbol{x}_i$

$$\omega_i = P^T(\boldsymbol{x}_i) / P^S(\boldsymbol{x}_i). \tag{2}$$

However, it is well known that such ratio is difficult to be directly computed with the marginal probability distribution. In this way, many effective methods have been developed to approximately estimate the aforementioned ratio bypassing the estimation step of the marginal probability distribution.

From the perspective of deep model training, the instance-based DTL can be further divided into the following two subcategories by considering whether the labeled instances are available in the target domain: the weight-estimation and the heuristic-reweighting method.

The weight-estimation method, which mainly focuses on the situation where there is a lack of labeled instances in target domain, convers the instance transfer problem into a weight estimation problem by leveraging the kernel embedding techniques. For instance, based on the theory of Maximum Mean Discrepancy (MMD) between distributions, the weights of source instances can be estimated by matching the means between the reweighted sources instances and the target instances in a Reproducing Kernel Hilbert Space (RKHS) [39], which can be obtained by optimizing the following objective

$$\begin{aligned} &\arg\min_{\boldsymbol{\omega}} \left\| \frac{1}{C^S}\sum_{i=1}^{N^S}\omega_i\Phi^S\left(\boldsymbol{x}_i^S\right) - \frac{1}{N^T}\sum_{j=1}^{N^T}\Phi^T\left(\boldsymbol{x}_j^T\right) \right\|_{\mathcal{H}}^2 \\ &\text{s.t. } \omega_i \geq 0,\ \left| \frac{1}{C^S}\sum_{i=1}^{N^S}\omega_i - 1 \right| \leq \varepsilon \end{aligned} \tag{3}$$

where $\varepsilon$ denotes a positive real number. There are some other tricks to estimate the weights by utilizing the Kullback-Leibler (KL) divergence [40]. With the weight of each source instance being estimated, Eq. 3 can be integrated into the objective function of the target task to learn a deep model. It is worth mentioning that the weight estimation of the source instances can be integrated into the training process of the deep model.

The heuristic-reweighting method, which is suitable for implementing the DTL task when some labeled instances are available in the target domain, aims to identify negative source instances by using instance reweighting strategies in a heuristic way. One of the most popular instance reweighting strategies is the Transfer Adaptive Boosting (TrAdaBoost) algorithm proposed by Dai *et al.* [41], in which the different weighting strategies are applied for the labeled instances in the source-domain and the target-domain to reduce the impact of negative source instances. Similar to the boosting-style algorithms, the weights of the source instances and the target instance can be updated through a lot of iterations, whose updating strategies are described as

$$\omega_i^S = \omega_i^S\left(1 + \sqrt{2\ln N^S / (N^S + N^T)}\right)^{-\mathcal{R}^S\left(f(\boldsymbol{x}_i^S),y_i^S\right)} \tag{4}$$

$$\begin{cases} \omega_j^T = \omega_j^T\left(\varepsilon / (1-\varepsilon)\right)^{-\mathcal{R}^T\left(f(\boldsymbol{x}_j^T),y_j^T\right)} \\ \varepsilon = \sum_{j=1}^{N^T}\omega_j^T\mathcal{R}^T\left(f(\boldsymbol{x}_j^T),y_j^T\right)\Big/\sum_{j=1}^{N^T}\omega_j^T \end{cases} \tag{5}$$

where $i = [1,\cdots,N^S]$, $j = [1,\cdots,N^T]$, and $\varepsilon$ denotes the mean loss of all target domain instances. It should be highlighted here that each iteration will learn a new weak deep model, and therefore ensemble techniques are used to form a final classifier by integrating all weak deep models. Besides the TrAdaBoost algorithm and its variations, the other heuristic-reweighting methods make full use of not only the labeled instances in the source and the target domain but also the unlabeled instances in the target domain. An intuitive solution of these methods is to decompose the objective function into three parts:

$$\begin{aligned}\mathcal{L} = \frac{1}{C^S}\sum_{i=1}^{N^S}\omega_i\mathcal{R}^S\left(f(\boldsymbol{x}_i^S), y_i^S\right) + \frac{1}{C^{T(L)}}\sum_{j=1}^{N^{T(L)}}\mathcal{R}^{T(L)}\left(f(\boldsymbol{x}_j^{T(L)}), y_j^{T(L)}\right) \\ + \frac{1}{C^{T(U)}}\sum_{k=1}^{N^{T(U)}}\gamma_k\mathcal{R}^{T(U)}\left(f(\boldsymbol{x}_k^{T(U)}), \hat{y}_k^{T(U)}\right)\end{aligned} \tag{6}$$

where the superscript of $S$, $T(L)$ and $T(U)$ denotes the labeled source, the labeled target and the unlabeled target, respectively. $C^S = \sum_{i=1}^{N^S}\omega_i$, $C^{T(L)} = N^{T(L)}$, $C^{T(U)} = \sum_{k=1}^{N^{T(U)}}\gamma_k$, $\gamma_k$ denotes the weight for the unlabeled target instance, and $\hat{y}_k^{T(U)} = P^T(y \mid \boldsymbol{x}_k^{T(U)})$ is the true conditional distributions of the unlabeled target instances. Generally, the optimal values of $\omega_i$, $\gamma_k$ and $\hat{y}_k^{T(U)}$ are unknown for computing these loss terms. Therefore, several techniques can be used during the deep model training to learn these parameters in a heuristic way. The typical procedure can be concluded as the following steps:

(1) An auxiliary classifier is firstly trained on the labeled target instances and then used to classify the labeled source and the unlabeled target instances to obtain the predicted probability of each instance.

(2) The labeled source and the unlabeled target instances are ranked based on its predicted probability, respectively.

(3) The $\omega_i$ of top n instances from the labeled source domain that are incorrectly predicted by the auxiliary classifier are set to zero, and the weights of others are set to one.

(4) The top n instances from the unlabeled target domain that have the highest prediction confidence are selected, for which the $\gamma_k$ is set to one and the $\hat{y}_k^{T(U)}$ is assigned to a pseudo label according to its predicted probability. Additionally, for all other instances from the unlabeled target domain, $\gamma_k$=0.

With the steps mentioned above, the whole loss can be calculated with the objective function presented in Eq. 6. Note that the selected labeled source and the unlabeled target instances can be used to train the auxiliary classifier again in the next iteration.

#### *2.2.2 Model-based DTL*

Model-Based DTL focuses on the transfer assumption that the tasks between the source and the target domains share some common knowledge in the model level, which means that the transferable knowledge is well embedded into a pretrained source deep model whose structure and parameters are general and helpful for learning a powerful target model. The goal of model-based DTL approaches is to exploit which part of the deep learning model pretrained in the source domain can help improving the model learning process for the target domain. Model-based DTL algorithms are based on the assumption that some labeled instances in the target domain should be available during the target model training. According to the way of training of the target deep model, the model-based DTL can be further divided into two subcategories: sequential training and joint training.

Sequential training establishes the target deep model by pretraining a deep learning model on auxiliary domains which have much richer and larger labeled instances and then fine-tuning the well-trained source model on the target domain which often lacks sufficient labeled instances. Specifically, sequential training-based DTL approaches typically contains two stages. In the first stage, i.e., the pretraining on auxiliary domains, a well-trained source model $\mathcal{F}^S(\bullet;\theta^S)$ has been learned from the source data, which can be defined as

$$\mathcal{F}^S(\bullet;\theta^S) = \arg\min \mathcal{R}^S\left(f^S(\boldsymbol{X}^S;\theta^S),\boldsymbol{Y}^S\right) \tag{7}$$

where $\theta^S = \{\theta_i^S\}_{i=1}^{L^S}$ is the model parameter set of the pretrained source model, $L^S$ denotes the layer number of the source model, $\mathcal{R}^S$ denotes the risk function for the source task. In the second stage, that is, the fine-tuning on the target domain, the target deep model $\mathcal{F}^T(\bullet;\theta^T)$ can be obtained by freezing some components of the well-trained source model and fine-tuning the rest components with the target domain data, or by reusing all the parameters of the well-trained source model to initialize the target deep model and retraining the whole target model with the target domain data. The processes of this stage can be formulated as

$$\begin{aligned}&\mathcal{F}^T(\bullet;\theta^T) = \arg\min \mathcal{R}^T\left(f^T(\boldsymbol{X}^T;\theta^T),\boldsymbol{Y}^T\right)\\ &\text{s.t. } \theta^T \text{ initialized/frozen with } \theta^*\end{aligned} \tag{8}$$

where $\theta^* = \{\theta_i^S, i \in [1,\cdots,L^S]\}$ is a subset of $\theta^S$ learned in the first stage, $\theta^T$ denotes the model parameter set expectedly learned in the second stage, $\mathcal{R}^T$ denotes the risk function for the target task. It is worth mentioning that the higher-level layers are prone to learn the task-specific representations and the lower-level layers are able to capture general representations in a deep learning model. Therefore, it is a classical fine-tuning strategy to freeze *n* lower-level layers learned from auxiliary domains and retrain the higher-level layer with limited target domain data.

Joint training tries to implement the source and the target tasks simultaneously. Different from the multi-task learning approaches which equally optimize the performance over all tasks, joint training-based DTL approaches focus on improving the performance of the target task by leveraging common knowledge from the source task. More specifically, there are two ways to joint training target task with source task. The first one is hard parameter sharing which shares the hidden layers directly while keeping the task-specific layers independently. The second one is soft parameter sharing which simply change the weight coefficient for the source and the target tasks or add regularization terms in the risk function. The processes of the soft parameter sharing can be defined as

$$\begin{aligned}\mathcal{F}^T(\bullet;\theta^T) = \arg\min[&\alpha\mathcal{R}^S\left(f^S(\boldsymbol{X}^S;\theta^S),\boldsymbol{Y}^S\right) + \beta\mathcal{R}^T\left(f^T(\boldsymbol{X}^T;\theta^T),\boldsymbol{Y}^T\right) \\ &+ \gamma\mathcal{R}^R\left(f(\boldsymbol{X}^S,\boldsymbol{X}^T;\theta^S,\theta^T)\right)]\end{aligned} \tag{9}$$

where $\mathcal{R}^S$, $\mathcal{R}^T$ and $\mathcal{R}^R$ are the risk functions of the source task, the target task and regularization terms, respectively; and $\alpha$, $\beta$ and $\gamma$ are the weight coefficients for the corresponding task.

### *2.2.3 Feature-based DTL*

Feature-Based DTL endows deep models with the ability to transfer knowledge by learning the common representations in the feature space level, rather than in the instances or the model level, which further relaxes the assumption in the instance-based DTL transfer learning scenario to allow the differences of feature spaces to exist in the source and target domains. An intuitive solution behind feature-based DTL approaches is to learn the mapping function as a bridge to convert the raw data in source and target domains from the different feature spaces to a common latent feature space, where the difference between domains can be reduced and the deep feature representations that are discriminative for the main learning task and indiscriminate with respect to the shift between different domains can be obtained. With these good representations, the performance of deep models can be significantly improved in accomplishing the target task.

From a broader perspective, feature-based DTL approaches intuitively covers two transfer styles without or with adaptation to target domain. The approaches without adaptation firstly extract the lower-level representations by using a pretrained source model, and then directly take the extracted representations as inputs for the target model, which are suitable and effective only when the target domain is closely related to the source domain. The approaches with adaptation adapt the feature representations across different domains through domain adaptation techniques, which can obtain a well performing model even if there is a shift or gap between source and target domains. Since the approaches without adaptation are easily implemented and their assumption may be too strong to be satisfied in most practical transfer scenarios, the following part mainly focus on the approaches with adaptation.

A crucial problem of feature-based DTL with domain adaptation in learning domain invariant features is how to estimate and learn representation invariance between source and target domains. The ways of constructing representation invariance measures generally include three strategies: leveraging criterions based on the discrepancy to the reduce difference of distribution, adding domain discriminative architectures to encourage the domain confusion through the adversarial mechanism, and combining the data reconstruction as an auxiliary task to help improving representations invariance. Therefore, the feature-based DTL with domain adaptation approaches can be further summarized into the following three subcategories.

The first subcategory is discrepancy-based domain adaptation, which aims to align the feature distribution shift and to improve the ability of learning transferable representations by reducing the discrepancy based on distance metrics or criterions defined between corresponding-level representations of the given source and target domains. The criterions that are proven to be successful for discrepancy-based domain adaptation include MMD [39], KL divergence [40], multiple kernels MMD (MK-MMD) [42], Correlation Alignment (CORAL) [43] and Wasserstein Distances (WD, also known as Earth-Mover distance) [44], among others. Taking the most commonly one, MMD, as an example, and given two domain representations $\boldsymbol{h}^S$ (source) and $\boldsymbol{h}^T$ (target), the criterion based on MMD can be empirically estimated as follows:

$$MMD(\boldsymbol{h}^S,\boldsymbol{h}^T)=\left\|\frac{1}{N_S}\sum_{i=1}^{N_S}\varphi(\boldsymbol{h}_i^S)-\frac{1}{N_T}\sum_{j=1}^{N_T}\varphi(\boldsymbol{h}_j^T)\right\|_{\mathcal{H}}^2 \tag{10}$$

Another common criterion is the Wasserstein Distances. The criterion based on the WD can be expressed as:

$$WD(\boldsymbol{h}^S,\boldsymbol{h}^T)=\inf_{\gamma\in\Gamma(\boldsymbol{h}^S,\boldsymbol{h}^T)}\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma}\left[\|\boldsymbol{x}-\boldsymbol{y}\|\right]=\sup_{\|f\|_L\le 1}\left(\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{h}^S}\left[f(\boldsymbol{x})\right]-\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{h}^T}\left[f(\boldsymbol{x})\right]\right) \tag{11}$$

The more details about WD can be found in [44].

In the process of model training, the deep neural network can be optimized by minimizing the classification loss on the labeled instance, $\mathcal{R}_C(\boldsymbol{X}_L,\boldsymbol{Y}_L)$, while the domain invariant representations are measured by one/multiple adaptation layer(s) with such criterion. The objective function of discrepancy-based domain adaptation is formulated as

$$\mathcal{L}=\mathcal{R}_C(\boldsymbol{X}_L,\boldsymbol{Y}_L)+\sum_{i=1}^{L^A}\lambda_i MMD(\boldsymbol{h}_i^S,\boldsymbol{h}_i^T) \tag{12}$$

where $L^A$ denotes the number of adaptation layers and the coefficient $\lambda_i$ is a penalty parameter for the $i$-th adaptation layer.

The second subcategory is the adversarial-based domain adaptation, which is inspired by the Generative Adversarial Networks (GANs) [45] and seeks to endow the deep neural network with the ability of learning domain-invariant representations. The GAN is typically composed of two components, that is, a generator (G) that generates fake data from noise and a discriminator (D) that distinguishes whether an instance is real or generated, which can be optimized by iteratively training D to maximize correct assignment of (real, fake) labels and training G to minimize the differences of real and generated data to confuse the discriminator:

$$\min_G\max_D\mathcal{L}(D,G)=\mathbb{E}_{\boldsymbol{x}\sim P_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{z}\sim P_z(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))] \tag{13}$$

In adversarial-based domain adaptation, such adversarial mechanism has been introduced into the deep neural network to ensure that the characteristics resulting from the difference of diverse domains cannot be distinguished. In light of whether to generate synthetic data, the historical adversarial-based domain adaptation approaches can be summarized as generative or non-generative adaptation model.

The generative adaptation model focuses on generating new data that are similar to the real data of the target domain by directly using GANs. More specifically, in the generative adaptation model, the generator $G(\boldsymbol{x}^S, \boldsymbol{z})$ generates an adapted instance $\boldsymbol{x}^G$ taking a source instance $\boldsymbol{x}^S$ and a noise vector $\boldsymbol{z}$ as inputs, and the discriminator tries to distinguish between the generated instances $\boldsymbol{x}^G$ (fake) and the target instances $\boldsymbol{x}^T$ (real). It is worth noting that, in contrast to the standard GANs in which the input of the generator is only a noise vector, the generative adaptation model's generator takes both a noise vector and a source instance as inputs. The generative adaptation model and its variants could be divided into two types from the perspective of neural network structure. The first one has two stages during model training: (1) generates the synthetic instances to augment the training dataset; (2) trains an extra classifier with both real and generated instances. The second one is usually augmented with a task-specific classifier (T) apart from the G and D, such that the goal of the generative adaptation model is to alternatively optimize the following minimax objective:

$$\min_{G,T} \max_{D} \mathcal{L}(G,D,T) = \alpha \mathcal{L}_{adv}(G,D) + \beta \mathcal{L}_{task}(G,T) \tag{14}$$

where $\alpha$ and $\beta$ are coefficients of the corresponding loss, $\mathcal{L}_{adv}$ and $\mathcal{L}_{task}$ denotes the adversarial loss and task loss, respectively.

The non-generative adaptation model pays more attention to learn the domain-invariant representations, rather than generating new data, by introducing the minimax loss or the domain-confusion loss into the deep model which typically consists of three parts: the feature extractor (instead of the generator), the domain discriminator and the task-specific classifier. One of the promising solutions in implementing the non-generative adaptation is to introduce a special Gradient Reversal Layer (GRL) between the feature extractor and the domain discriminator, which first was introduced in the Domain Adversarial Neural Network (DANN) [46]. DANN ensures that the representations learned from different domains are as closer as possible by maximizing the domain confusion loss through the GRL. The GRL function as an identity transformation during the forward propagation, while during the backward propagation it receives the gradient from the subsequent layer and reverses the sign of the gradient before delivering to the preceding layer:

$$\begin{cases} GRL(\boldsymbol{h}) = \boldsymbol{h}, & \text{forward propagation} \\ \dfrac{dGRL}{\boldsymbol{h}} = -\alpha \boldsymbol{I}, & \text{backward propagation} \end{cases} \tag{15}$$

With such GRL, the parameters of the feature extractor and the domain discriminator can be globally optimized and simultaneously updated. Another promising solution is to splits the optimization into two independent objectives: the parameter of the feature extractor $\theta_{FE}$ and the parameter of the domain discriminator $\theta_D$, and to perform iterative updates for the two objectives given the fixed parameters from the previous iteration:

$$\min_{\theta_D} \mathcal{L}_D(\boldsymbol{x}^S, \boldsymbol{x}^T, \theta_{FE}; \theta_D) = -\mathcal{R}^D\left(\boldsymbol{h}_D, y_D\right) \tag{16}$$

$$\min_{\theta_{FE}} \mathcal{L}_{conf}(\boldsymbol{x}^S, \boldsymbol{x}^T, \theta_D; \theta_{FE}) = -\mathcal{R}^{conf}\left(\boldsymbol{h}_D\right) \tag{17}$$

where $\boldsymbol{h}_D = D(FE(\boldsymbol{x}; \theta_{FE}); \theta_D)$ denotes the output of the domain discriminator, $\mathcal{R}^D$ is the risk function for the domain classification where the Cross-entropy loss function is commonly used, and $\mathcal{R}^{conf}$ is the risk function for the domain confusion where the probability density function of a uniform distribution is adapted based on the cross entropy between the predicted domain labels. Thus, the deep model can be optimized by adversarial training through minimizing the Eq. 15 only for updating $\theta_D$ and minimizing the Eq. 16 for updating $\theta_{FE}$, which can ensure that the learned representations is domain invariant.

The third subcategory is the reconstruction-based domain adaptation, which combines the auto-encoder neural networks with a task-specific classifier to jointly optimize a private encoder that captures domain-specific representations and a shared encoder that learns common representations between the domains. The reconstruction-based domain adaptation model integrates a shared decoder which learns to reconstruct the input instances with a reconstruction loss by taking both the private and the common representations as inputs. The reconstruction losses that have widely used in DTL are the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). The task-specific classifier is trained on the common representations learned by the shared encoder, which will be able to generalize across domains better since its inputs have been separated from the representations that are special to each domain.

Note that, besides the categories of DTL mentioned above, there exist many hybrid methods to build a DTL model using several of the aforementioned techniques simultaneously. The core idea of the hybrid methods is that the domain-invariant knowledge between source and target domain can be learned in any two or more of the levels, that is the instance-, model- and feature-level. Since the main definitions and theories of hybrid methods are almost the same with those mentioned above, this survey will not enumerate them in detail.

## 3 Formulations and Applications of DTL for Fault Diagnosis in Industrial Scenario

In the past few years, scientific researchers and engineers from both academic and industrial communities have already brought many impressive achievements and successful real-world application across a lot of DTL algorithms in implementing complex tasks. Examples include object recognition and detection based on image data collected in different conditions [47], speech recognition based on audio data sampled from different speakers [48], text classification and translation based on document data written in different languages [49], etc. Compared to the vast literature focused on the application in the field of computer vision and natural language processing, few surveys focus on the relative developments of DTL in industrial scenarios for the task of fault diagnosis. Therefore, in this section, the literature historically published in addressing the fault diagnosis problems with the DTL approaches is systematically reviewed, including the problem formulation of DTL for fault diagnosis in industrial scenarios and its main applications of each scenario.

### *3.1 Problem Formulation of DTL for Fault Diagnosis*

Within industrial scenarios, there exist many exact problems that have attracted considerable attentions and much emphasis has been placed on solving such problems. Understanding what type of problems have been faced with IFD and how to solve them is of great significance for researchers and engineers to correctly understand the reasons we survey this topic from the perspective of practical industrial scenarios, and to formulate the pattern of the DTL for fault diagnosis.

In the phase of current manufacturing industry, the major problems encountered in applying intelligent methods for machines are summarized as follows:

(1) The deep models learned from the given training data are not robust enough to be generalized from one application to a new or similar one, so it is difficult to deal with the uncertainty caused by the varying environment during machines working. For instance, the WCs of machines are various during long-term operation, and the health status is also declining with the degradation of crucial components. However, the generalization performance of deep models is insufficient in the face of changeable WCs and diversified data.

(2) Considering the fast upgrading and updating of the manufacturing products, the deep models also require periodic updates for the performance improvement. However, it is hard to collect and annotate the training data from scratch for the application of new products while reusing the labeled historical data collected and accumulated from the old products is relatively easy.

(3) The deep models learn how to make a fault diagnosis through the observations of given labeled data, so they encounter difficulties to recognize unknown patterns or faults. In order to step into the real industrial applications, it is a significant function that the fault diagnosis models can automatically detect a new anomaly since the unseen faults inevitably occur during the long-term services of the complex mechanical equipment.

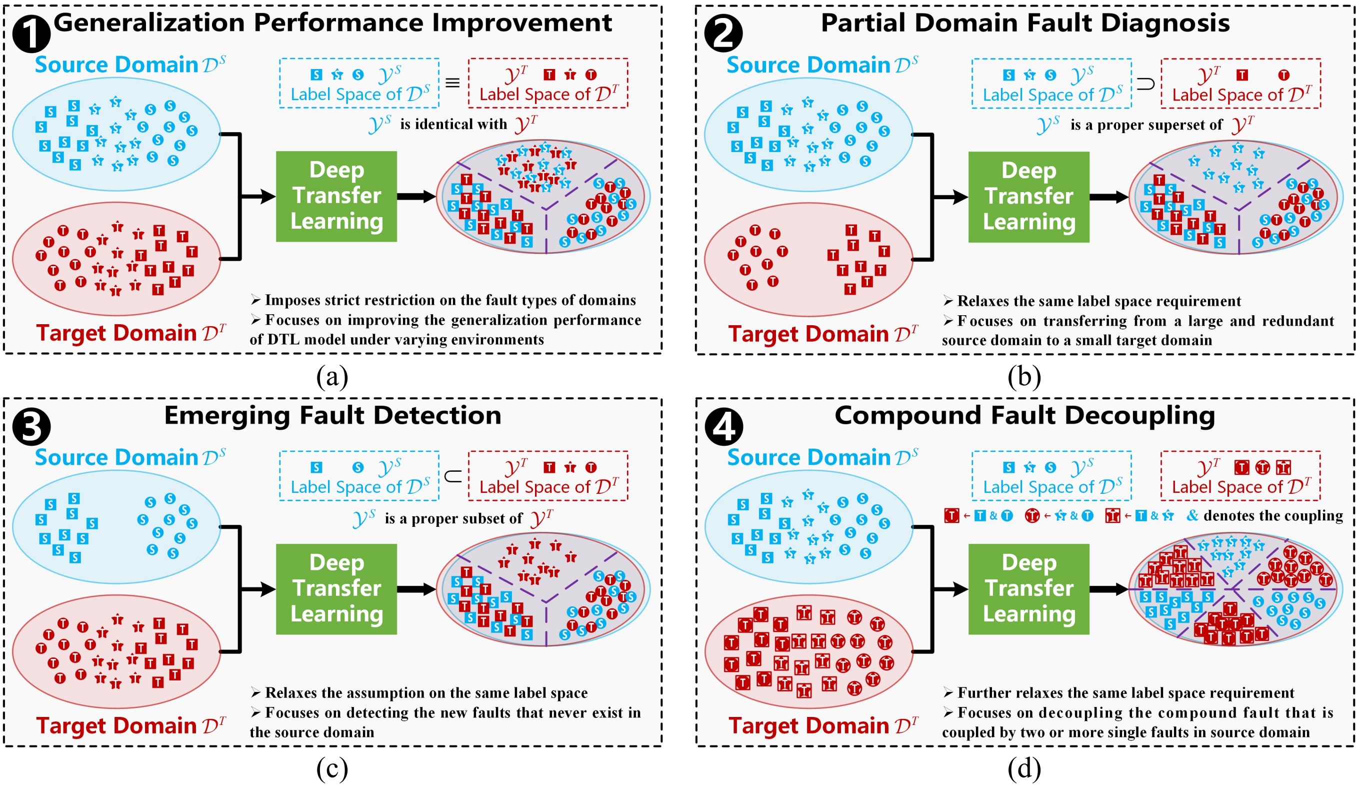


Fig. 4. Illustration of the four application scenarios for DTL (a) Generalization Performance Improvement, (b) Partial Domain Fault Diagnosis, (c) Emerging Fault Detection, (d) Compound Fault Decoupling

(4) The vast majority of researchers and engineers are concerned about the improvement of precision and accuracy for classifying the different faults. The compound fault, as a primary failure leading to expensive maintenance costs and tragic catastrophes in industrial scenarios, often emerges and evolves when multiple crucial components are simultaneously degraded or even broken. However, few time and effort have been paid to investigate the task of decoupling compound faults in an intelligent manner.

Aiming at solving the four problems mentioned above, the industrial application of DTL can be defined as four scenarios: generalization performance improvement, partial domain fault diagnosis, emerging fault detection, and compound fault decoupling, respectively. As descripted in the previous section and shown in Fig. 4, given a DTL task defined by $f^{S\to T}(\bullet): \boldsymbol{X}^T \to \boldsymbol{Y}^T$ based on [$\mathcal{D}^S$, $\mathcal{D}^T$, $\mathcal{T}^S$, $\mathcal{T}^T$], the four application scenarios can be formulated from the perspective of fault classification as follows:

***Generalization performance improvement***: In this scenario, the label space of target domain is identical with the label space of the source domain, that is, $\mathcal{Y}^T \equiv \mathcal{Y}^S$, which imposes strict restriction on the fault types of domains and mainly focuses on improving the generalization performance of DTL model under varying environments. Such scenario is called as generalization performance improvement.

***Partial domain fault diagnosis***: In this scenario, the label space of target domain is a proper subset of the label space of the source domain, that is, $\mathcal{Y}^T \subset \mathcal{Y}^S$, which relaxes the same label space requirement and mainly focuses on transferring knowledge from a large-scale but redundant source domain to an unknown small-scale target domain. Such scenario is referred to as partial domain fault diagnosis.

***Emerging fault detection***: In this scenario, the label space of the target domain is a proper superset of the label space of the source domain, that is, $\mathcal{Y}^T \supset \mathcal{Y}^S$, which also relaxes the assumption on the same label space and mainly focuses on detecting the new faults that never exist in the source domain. Such scenario is known as emerging fault detection.

***Compound fault decoupling***: In this scenario, the label space of the target domain is different from the label space of the source domain, but each fault in target domain is coupled by multiple single faults in the source domain. More specifically, a fault in the target domain $y_i^T$ is a compound fault which is coupled by two or more single faults in the source domain $y_j^S \,\&\cdots\&\, y_k^S$. Such scenario is defined as compound fault decoupling.

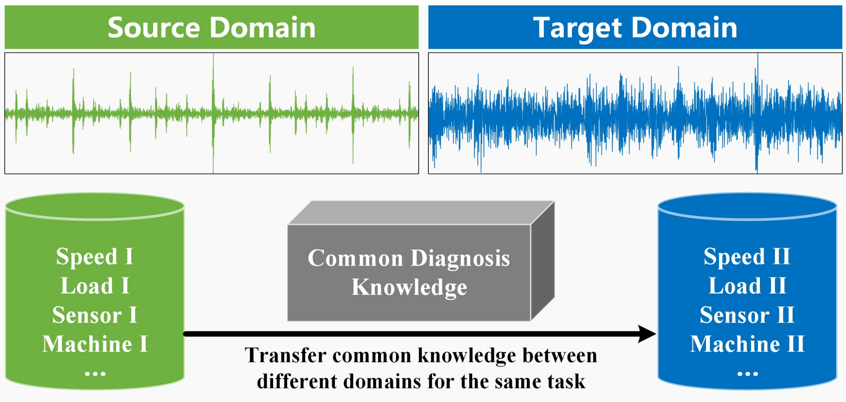


Fig. 5. Illustration of the motivation behind the scenario of generalization performance improvement

## *3.2 Generalization Performance Improvement*

### 3.2.1 Motivations and goals

As illustrated in Fig. 5, the motivation behind this scenario is that, in the real-world application, if the common knowledge, which does not contain the uncertainty information caused by varying environments, can be learned with limited source data for a specific task, a deep model with satisfactory generalization performance can be obtained for the same task even when it faces a new environment. Thus, as depicted in Fig. 4 (a), the ultimate goal in this scenario is to learn a robust deep model that should be able to implement the objective task under varying environments. Table I concludes the current solutions using the DTL-based fault diagnosis approaches for the generalization performance improvement from three application scenarios, that is varying WCs, across different machines, and other scenarios.

TABLE I Solutions for Generalization Performance Improvement

| Application Scenarios | Categorization of DTL | References | Common algorithms used |
|---|---|---|---|
| Varying Working Conditions | Instance-based | Zhang *et al.* [50] , Shen *et al.* [51], Song *et al.* [52], Pan *et al.* [53] | Weight-estimation with MVD/MMD, Heuristic-reweighting with pseudo-label, TrAdaBoost |
| | Model-based | Shao *et al.* [54], Zhou *et al.* [55], Lu *et al.* [56], Zhang *et al.* [57], Hasan *et al.* [58], [59], Han *et al.* [60], He *et al.* [61], [62], Zhao *et al.* [63], Du *et al.* [64], Chen *et al.* [65], Li *et al.* [66], Wang *et al.* [67], Shao *et al.* [68], Cao *et al.* [69] | Sequential training (DL+Fine-tune), Joint training (Y-Net) |
| | Feature-based | Lu *et al.* [70], Li *et al.* [71], Tong *et al.* [72], [73] Zhang *et al.* [74], Xiao *et al.* [75], An *et al.* [76], Li *et al.* [77], Azamfar *et al.* [78], [79], Zhang *et al.* [80], Zhu *et al.* [81], Singh *et al.* [82], Li *et al.* [83], Han *et al.* [84], [85], Wei *et al.* [86], Wang *et al.* [87], Shen *et al.* [88], Li *et al.* [89], Wu *et al.* [90], Yang *et al.* [91], Qian *et al.* [92], [93], Wang *et al.* [94], An *et al.* [95], Xiong *et al.* [96], Li *et al.* [97], Bao *et al.* [98], Xu *et al.* [99] Huang *et al.* [100] | Discrepancy-based (MMD, CMMD, MK-MMD, KL Divergence, CORAL, CMD, MCD) |
| | | Li *et al.* [101], Zheng *et al.* [102], Liang *et al.* [103], [104], Tao *et al.* [105], Shao *et al.* [106], Guo *et al.* [107], Shi *et al.* [108], Han *et al.* [109], Jiao *et al.* [110], Shao *et al.* [111], Chai *et al.* [112], Chen *et al.* [113], Li *et al.* [114], Mao *et al.* [115], Liu *et al.* [116], Li *et al.* [117], [118], Zhang *et al.* [119], Li *et al.* [120], Zhang *et al.* [121], Jiao *et al.* [122], [123], Cheng *et al.* [124], Wang *et al.* [125], Zou *et al.* [126], Han *et al.* [127], Wang *et al.* [128], Yu *et al.* [129], She *et al.* [130], Ragab *et al.* [131], Liao *et al.* [132] | Adversarial-based (GAN, DATN, DANN, GRL, ADDA, W-GAN) |
| | | Pang *et al.* [133], Liu *et al.* [134], [135], Wen *et al.* [136], Wan *et al.* [137], Tang *et al.* [138] | Reconstruction-based (SAE+TL) |
| Across Different Machines | Instance-based | Zheng *et al.* [144], Yang *et al.* [145], Wu *et al.* [146] | Instance-based discriminative loss |
| | Model-based | Wang *et al.* [147], Shao *et al.* [148], Li *et al.* [149], He *et al.* [150], Chen *et al.* [151] | VGG-19/SAE+Fine-tune |
| | Feature-based | Guo *et al.* [152], Yang *et al.* [139], [153], Li *et al.* [154], Wu *et al.* [155], Zheng *et al.* [156], Lv *et al.* [157], Zhao *et al.* [158], Li *et al.* [159], Tan *et al.* [160], Chen *et al.* [161], Zhang *et al.* [162], [163], [164] | Discrepancy-based (MMD, PK-MMD, MCD, Mutual information) |
| | | Wang, *et al.* [165], Feng *et al.* [166], Zhu *et al.* [167], Liao *et al.* [168] | Adversarial-based (GAN, DATN) |
| | | Lu. *et al.* [169] | Reconstruction-based (SAE+MMD) |
| Others (Imbalanced instances, Across sensors, etc.) | Instance-based | Xiao, *et al.* [170] | TrAdaBoost |
| | Model-based | Li *et al.* [171], Kim *et al.* [172], He, *et al.* [173] | CNN/SAE+Fine-tune |
| | Feature-based | Zhang *et al.* [174], Zou *et al.* [175], Zhang *et al.* [176], Li *et al.* [177], Zareapoor *et al.* [178], Zhang *et al.* [179], Li *et al.* [180], Li *et al.* [181], Siahpour *et al.* [182], Pandhare *et al.* [183], Qin *et al.* [184], Wu *et al.* [185] | GAN and its variants, DANN, One-shot learning, Unsupervised parallel data alignment |

### 3.2.2 Solutions for varying WCs

One of the main factors leading to distribution shift between training and testing data is that the WCs of IE are complex as the frequently changing of speeds, loads or operations. Therefore, in this case, a lot of solutions based on DTL have been investigated for enhancing the generalization performance of deep models which can effectively deal with the uncertainty caused by varying WCs during the long-term services of machines.

***Instance-based DTL solutions:*** Combining the maximum variance discrepancy (MVD) and the maximum mean discrepancy (MMD), Zhang *et al.* [50] proposed a weight-estimation method for bearing fault diagnosis to calculate the adaptation matrix between the source and target instances, which is used to reweight and down-weight the source instances that are negative for the target model training. As the most popular instance reweighting strategy, the fast TrAdaBoost algorithm was introduced by Shen *et al.* [51] as an instance reweighting strategy that can weaken the weight of the low-quality instances and enhance the weight of high-quality instances through iteratively update, which successfully employed to enhance the generalization performance for the fault diagnosis model of a gearbox operating under varying working conditions. Song *et al.* [52] proposed a retraining strategy-based domain adaption network (DAN-R) for IFD, which annotates the unlabeled instances in the target domain with pseudo-labels and then retrains the classification network using both training instances and pseudo-labeled testing instances. According to these instance-based DTL solutions [50]-[53], it can be found that instances-based approaches are effective and applicable for the application scenario of varying WCs. However, the performance of these methods discussed above are depended to some extent on the number or the quality of target instances, and they may have difficulty to tackle the problems in more challenging but complex scenarios which have significant discrepancy between source and target domain.

***Model-based DTL solutions:*** One of the major model-based DTL solutions for varying WCs is to pretrain a deep model (such as VGG-16 [54], VGG-19 [55] and AlexNet [56]) on source WCs and sequentially fine-tune it using the labeled instances in the target working condition [57]-[68]. For instance, Han *et al.* [60] finetuned a well-trained CNN with three transfer learning strategies at different levels of the CNN architecture: (1) just finetuning the classifier, (2) just finetuning the feature extractor, and (3) finetuning the whole CNN model, where the characteristics of each strategy are discussed and compared. The experimental validations show that the proposed transfer strategies can effectively transfer the useful features of the well-trained CNN for the target task and achieve the highest accuracy for the generalization problem of WCs. Another method is to jointly implement the source and target tasks in a deep model with multi-branches. Cao *et al.* [69] developed a multi-branch deep model, named Y-Net, to transfer knowledge for the fault diagnosis of planetary gearboxes, which consists of three components: two convolutional classification networks (one for the source task and another for the target task, and sharing weights with each other) and a reconstruction network. Compared with other solutions training a model from scratch, model-based DTL solutions benefit from the faster convergence rate and the reduction of the risk of overfitting. Additionally, these solutions have its inherent limitation that the fine-tune algorithm heavily relies on the dependence of labeled training data in the target scenario.

***Feature-based DTL solutions:*** For the application scenario of varying WCs, the feature learned by deep models is expected to be speed-insensitive and load-insensitive. Generally, the more insensitive to working conditions the features are, the better the generalization performance of deep model will be. Researchers, therefore, have been placed many efforts on how to learn universal features under varying WCs from the following three aspects.

From the aspect of discrepancy-based domain adaptation, a variety of criteria, such as MMD [70]-[83], Conditional Maximum Mean Discrepancy (CMMD) [84]-[88], MK-MMD [89]-[91], KL Divergence [92], [93], CORAL [94], [95], Central Moment Discrepancy (CMD) [96], [97], Maximum Classifier Discrepancy (MCD) [98] and others [99], [100] have been widely introduced into the objective function to measure the features discrepancy between the source and the target WCs. Specifically, Lu *et al.* [70] proposed a Deep neural network for domain Adaptation in Fault Diagnosis, named DAFD. The DAFD introduced the MMD term into the objective function of deep neural network for reducing the distribution discrepancy between different WCs in 2017, which was the first time that the transfer learning technique, i.e., the domain adaptation, was applied to train the deep model in the field of IFD. Aiming at minimizing the discrepancy of marginal and conditional distributions simultaneously, a deep transfer network (DTN) with joint distribution adaptation (JDA) was proposed by Han *et al.* [84] through the integration of marginal MMD and conditional MMD. Experiments carried on three practical industry datasets show that, comparing with the traditional deep learning- and transfer learning-based methods, the DTN with JDA achieves state-of-the-art diagnosis results regarding the application scenario of diverse operating WCs. In contrast to constructing a single layer with linear MMD in deep model, a multilayer domain adaptation (MLDA) method was proposed by Yang *et al.* [91]. The MLDA that matches the shift in both marginal and conditional distributions across WCs by adding MK-MMD and pseudo-label learning in multiple adaptation layers, can effectively extract working-condition-insensitive features for bearing fault diagnosis. Apart from using the MMD and its variations, Qian *et al.* [92], [93] utilized the KL Divergence to align the first and higher order moment discrepancies, Wang *et al.* [94] and An *et al.* [95] adopted CORAL to minimize the distribution gap between source and target WCs by aligning the second-order statistics, and Li *et al.* [96] employed CMD to reduce the discrepancy between different working conditions.

From the aspect of adversarial-based domain adaptation, the mechanism of GAN has been employed to help the deep model to learn task-sensitive but domain-insensitive features for the target tasks in the generative or non-generative adaptation ways. In the case of generative adaptation, deep GANs and its variants have been exploited to generate different types of data, such as frequency domain data [101], [102] and time-frequency domain data [103]-[105], with the help of available source data, and then these generated and real data are used to train an extra deep model, achieving reliable diagnosis results when testing data in target WCs are not available during model training. Shao *et al.* [106], Guo *et al.* [107] and Shi *et al.* [108] added an auxiliary classifier into the GAN, rather than training an extra classifier, to fully utilize the label information, hence, the enhanced models achieved higher diagnosis accuracy with few training data. However, one particular challenge for generative adaptation is the difficulty in evaluating the quality of the generated data with effective metrics quantitatively. In the case of non-generative adaptation, motivated by GAN, several deep DA frameworks such as the Domain Adversarial Transfer Network (DATN) [109]-[115], the DANN [116]-[119], the Adversarial Discriminative Domain Adaptation (ADDA) [120], the Wasserstein GAN (W-GAN) [121]-[128] and others [129], [130], have been developed and applied for fault diagnosis of machines under varying WCs. For instance, Chen *et al.* [113] exploited the discriminator of GANs as a domain classifier that performs binary domain classification and introduced a domain confusion loss, that is, the inverted label loss, to encourage the source and the target distributions to be a uniform distribution as close as possible. Liu *et al.* [116] utilized the DANN that integrates a GRL into the standard GAN to construct the deep model for bearing fault diagnosis, which largely enhances the generalization performance of diagnosis model under different speeds and loads. Following the principle of ADDA, a knowledge mapping-based adversarial domain adaptation (KMADA) method was proposed by Li and Shen *et al.* [120], which ensures that the feature space mapping from the target domain data can be updated until it is indistinguishable with the feature space mapping from the real source data. The KMADA achieved strong diagnosis results on an experimental bearing dataset and a locomotive bearing dataset. Wang *et al.* [128] proposed a Deep Adversarial Domain Adaptation Network (DADAN) for fault diagnosis of bearings and hard disk datasets collected from real-case data center, which employed a discriminator to measure the empirical Wasserstein distance between two domains instead of using a discriminator to classify the domain label. In addition to the methods mentioned above, some explorations have been proposed to deal with the problems of multiple target domains [131] and domain generalization [132]. Liao *et al.* [132] developed a deep semi-supervised domain generalization network to deal with a challenging diagnosis scenario where the well-trained model can generalize to an unseen working condition.

From the aspect of reconstruction-based domain adaptation, there are also some applications that utilized encoder–decoder reconstruction to enhance the generalization capability of diagnosis model under different WCs [133]-[138]. Wen *et al.* [136] constructed a three-layer SAE to learn the common representations from the raw data of different WCs in a reconstruction manner. An IFD method based on an autoencoder with adaptive transfer learning was proposed by Tang *et al.* [138], which use a shared encoder to learn transferable features using the reconstruction loss of RMSE and the adaptation loss of MMD.

### *3.2.3* Solutions for across different machines

Compared with the application scenario of varying WCs where the data used for model training and testing are both measured on the identical machine under different speeds, loads or operations, the main difference in the application scenario of across different machines is that those data are measured on related but different machines and suffer from more complicated factors, such as different mechanical structures, diverse material and various sizes. Such factors inevitably lead to more significant distribution shift between the training and testing data than the application scenario of varying WCs. Therefore, it is a more challenging task to transfer diagnosis knowledge across different machines. There are three typical applications of across different machines: transfer from laboratory to industry [139], transfer from simulation to reality [140], [141], and transfer from past to future [142], [143]. Once the bridge of transfer knowledge across different machines can be built, it will not only largely eliminate the dependency of the fault data collected from the target machine but also potentially reduce the economic cost spent for the maintenance of the target machine. With such demand increasing exponentially, fortunately, some solutions have been developed and investigated for further improving the practicability and generalization performance of diagnosis model.

***Instance-based DTL solutions:*** A Deep Domain Generalization Network for Fault Diagnosis (DDGFD), optimized with an instance-based discriminative loss, was proposed by Zheng *et al.* [144], aiming to explore the more challenging but practical across different machines scenarios where only normal samples are available in the dataset of the target machine. Yang *et al.* [145] developed a metric, named Optimal Transport-embedded Similarity Measure, for analyzing the transferability of diagnostic knowledge across machines, in which cluster-conditional distributions are explored to assign cluster labels for the target instances. Wu *et al.* [146] proposed a hybrid DTL method that combines the instance- and feature-transfer learning techniques to solve the diagnosis problem of bearings when sufficient labeled fault data in the practical engineering is lacking, which was validated in the application scenario of transfer from the Case Western Reserve University (CWRU) dataset to a locomotive bearing dataset collected in real industry.

***Model-based DTL solutions:*** Taking the rolling bearing fault diagnosis as a case study, Wang and Gao [147] adapted the VGG-19 network as the backbone model that was pretrained on non-manufacturing data, and then was finetuned on manufacturing machine for transferring common latent features among different machines. A Novel Stacked Transfer Auto-encoder (NSTAE), optimized using Particle Swarm Optimization (PSO), was proposed by Shao *et al.* [148] and was applied for IFD based on bearing and gear data collected from different rotating machines. Unlike the previous methods which focus on selecting the backbone model [149], [150], Chen *et al.* [151]

proposed a novel model-based DTL strategy for training a Transferable Convolutional Neural Network (TCNN), which exploits the knowledge learned from different source machines to improve the generalization performance of the target task. Its core idea is that the layers and the parameters of the pretrained TCNN are firstly subdivided into several blocks, and then each block is finetuned in reverse order. With respect to the model transfer, such strategy is suitable not only for the CNNs but also for other deep models such as DBN, SAE and Long short-term memory (LSTM). The model-based DTL solutions, especially the fine-tune algorithm, are comparatively easy to implement in the scenario of across different machines. But their performance would decrease dramatically if the labeled instances are insufficient or unavailable.

***Feature-based DTL solutions:*** For feature-based DTL, the discrepancy-based domain adaptation is still one of the most popular and promising solutions for fault diagnosis in the scenario across different machines and brings successful breakthroughs compared with traditional DL methods [152]-[164]. For example, Lei *et al.* have proposed several IFD methods based on discrepancy-based domain adaptation for transferring knowledge from laboratory to real industrial bearings [139], [152], [153]. A feature-based transfer neural network (FTNN) was proposed in [139] to learn transferable representation by combining multi-layer domain adaptation and pseudo label learning. In FTNN, a domain-shared CNN was trained by simultaneously minimizing three discrepancies: the classification discrepancy of the labeled instances in the source domain, the classification discrepancy of the unlabeled instances in the target domain with the help of the pseudo label learning, and the multilayer MMD discrepancy of the learned representations between across domains. In [153], a distance metric named polynomial kernel induced MMD (PK-MMD) was proposed to overcome the weakness of the Gaussian kernel induced MMD (GK-MMD). The experimental results showed that the PK-MMD based DTL method can not only improve the computation efficiency but also can achieve better performance for IFD in the across different machines scenario compared with other algorithms such as the Transfer Component Analysis (TCA), the DAFD, and the GK-MMD-based method. Meanwhile, Tan *et al.* [160] proposed a deep coupled joint distribution adaptation network (DCJDAN) to reduce the domain discrepancy between artificial and real damages, which has been validated on the dataset provided by Konstruktions- und Antriebstechnik (KAt) Bearing Data Center, Paderborn University. In addition, there are a few published methods which provide other solutions to solve the problems in across different machines scenarios by exploring the adversarial-based [165]-[168] and reconstruction-based domain adaptation techniques [169], which may not be as popular as the methods based on the discrepancy-based domain adaptation.

### 3.2.4 Solutions for other scenarios

Besides varying WCs and across different machines, there are other application scenarios as well, including across sensors and imbalanced instances. As for IFD methods, the locations, types and sampling frequency of sensors, as well as the number of training instances of each class, result in a huge distribution diversity between realistic industrial data. Many impressive studies have been applied to the application scenarios of imbalanced instances [170]-[180] and across different sensors [181]-[185], and paid much attentions to the investigation of how to improve the generalization performance of IFD models.

For the scenario of imbalanced instances, Zareapoor *et al.* [178] proposed a Minority oversampling Generative Adversarial Network (MoGAN) to deal with the problems where the number of each fault class are imbalanced during model training. The MoGAN converts the imbalanced problem into the balanced scenario by generating the minority instances through the GAN, which provides a potential solution for the scenarios where some labeled data are available in the target domain but they are not enough to train a satisfactory model. A one-shot learning method for fault diagnosis of 3D printers was proposed by Li *et al.* [180], which only requires one instance of each fault condition to accomplish the model training. Another scenario encountered in real industry is across different sensors. Prof. Jay Lee and his group have proposed several solutions for transferring diagnosis knowledge across sensors at different locations [181]-[185]. The proposed solutions are based on the unsupervised parallel data which are utilized to align the conditional distribution of the different health conditions. The experimental results showed that such solutions are promising to transfer common knowledge between the data from different locations of machines, and they can further improve the generalization performance of deep models in practical industry applications. Similarly, aiming at transferring the diagnosis model from one sensor to another, Qin *et al.* [184] designed a new transfer strategy for domain adaption, called Multi-Scale Transfer Voting Mechanism (MSTVM), which combines multi-scale feature learning and plurality voting operation techniques. The MSTVM can be used to the traditional domain adaption models, and the model's performance will be well improved.

## *3.3 Partial Domain Fault Diagnosis*

### 3.3.1 Motivations and goals

As illustrated in Fig. 6, the strong motivation behind this scenario is that, under the industrial big data environment, it is a promising solution to utilize the labeled historical data and the open-source industrial data which are collected from related scenarios, for training a diagnosis model that can transfer knowledge from large-scale but redundant source domain to unknown small-scale target domain. The challenges for partial domain fault diagnosis are due to the following two factors:

(1) Label space information of the target domain is lacking. In the trend of Industry 4.0, a large amount of monitoring data can be collected and stored for the target scenario. However, it is

expensive and unrealistic to annotate these large amounts of data, therefore the numbers and the types of faults are unknown.

(2) Outlier source faults may lead to negative transfer. From the viewpoint of big data, the large-scale but redundant source dataset is diverse enough to subsume all fault classes of the small-scale target dataset.

Thus, directly transferring between the entire source and target domains as the popular DTL methods is not an optimal and effective solution for the partial transfer scenario.

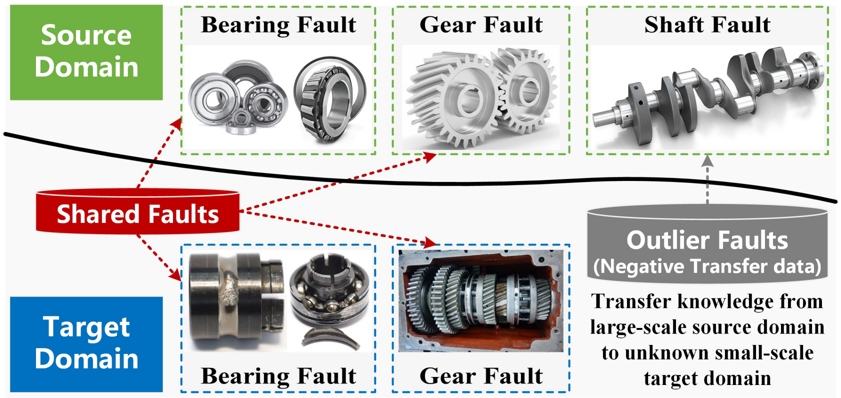


Fig. 6. Illustration of the motivation behind the scenario of partial domain fault diagnosis

As illustrated in Fig. 4 (b), the key goal in this scenario is to build an effective diagnosis model for partial domain fault diagnosis by aligning the distribution of source and target domains in the scope of the shared label space and singling out the outlier source data in the scope of the private label space. The DTL-based solutions developed for the partial domain fault diagnosis in recent years have been summarized in Table II. According to the experiments presented in the publications, these solutions not only have the capability to promote the positive transfer of the relevant data and to alleviate negative transfer of irrelevant data, but also can address the practical and challenging issues under the industrial big data environment.

TABLE II Solutions for Partial Domain Fault Diagnosis

| Application Scenarios | Categorization of DTL | References | Common algorithms used |
|---|---|---|---|
| Partial Domain Fault Diagnosis | Instance-based | Jiao *et al.* [187], Li *et al.* [188], Li *et al.* [189], Liu *et al.* [190] | Class weight-estimation strategy |
| | Feature-based | Li *et al.* [191], Han *et al.* [192], Deng *et al.* [193], Yang *et al.* [194], Wang *et al.* [195] | SAN, GAN+Attention/PK-MMD |

### 3.3.2 Solutions for partial domain fault diagnosis

As mentioned before, the assumption behind this scenario is that the label information of target data is unknown. Up to now, model-based DTL approaches are hardly applied to the problems in the scenario of partial domain fault diagnosis because they inherently rely on the label information of target instances. Therefore, in this case, most current DTL solutions have been developed on the basis of the instance-based and feature-based DTL methods.

***Instance-based DTL solutions:*** An intuitive solution to transfer knowledge from large-scale source dataset to small-scale target dataset is to select out the outlier instances in sources domain that are negative for building target model. Such an idea can be implemented by adapting the instance-level or class-level weighting strategies during the process of model training. Aiming at transferring knowledge from a large-scale dataset to a small-scale dataset (e.g., from ImageNet to Caltech-256), a Selective Adversarial Network (SAN) was firstly proposed by Cao *et al.* [186] in the Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR), 2018, for partial transfer learning. Inspired by the SAN, several class-level weighting methods have been proposed in the field of IFD [187]-[190]. For example, Jiao *et al.* [187] proposed a classifier inconsistency-based domain adaptation network (CIDA) for unsupervised partial domain fault diagnosis of planetary gearbox. The CIDA estimates the label space of target domain by calculating class weights through classifier inconsistency loss and selects out the source instances beyond the shared label space of source and target domains according to the class weights. The experimental results showed that the CIDA can implement the partial transfer diagnosis task from a working condition (containing all fault classes) to a target working condition (only containing a part of fault classes), and its performance is superior than that of the other popular DTL methods. Similarly, a Weighted Adversarial Transfer Network (WATN) was proposed by Li *et al.* [189] for partial domain fault diagnosis across different machines. In WATN, an auxiliary classifier is introduced to automatically learn the weight of each source instance, which can weight the contributions of each training instance to both feature learning and domain confusion. As a result, the role of irrelevant source instances can be effectively weakened during the knowledge transferring. However, these instance-based DTL methods are depended to some extent on the prediction distribution of the instances in the target domain.

***Feature-based DTL solutions:*** Besides the weighting mechanism described above, feature-based DTL solution have also been developed with promising results for partial domain fault diagnosis [191]-[195]. One example is that, inspired by GAN, Li and Zhang proposed an IFD method to address the partial domain adaptation problem by combining the techniques like conditional data alignment and unsupervised prediction consistency. Conditional data alignment is implemented by minimizing the distribution discrepancy between source and target domains through MMD. Unsupervised prediction consistency is achieved when the same prediction results of target domain data can be obtained after finishing the adversarial learning between multiple classification modules and the discriminator [191]. Similar application can be found in [192], which has been validated on a wind turbine fault dataset and achieved superior performance under different transfer scenarios than other traditional transfer learning methods. In addition, a double-layer attention based generative adversarial network (DA-GAN) was proposed by Deng *et al.* [193] for partial domain fault diagnosis of bearings, which aims to solve the problem, “where to transfer”, since the label space of target domain is unknown. In DA-GAN, the attention mechanism is introduced into two layers, one for domain attention and another for sample attention, which can provide guidance for the model to focus which fault classes should be shared or singled out. Yang *et al.* [194] further extended the partial domain fault diagnosis to a more practical and challenging setting where the instances imbalanced between fault classes, exist in the target domain, and proposed a deep partial transfer learning network (DPTL-Net) to selectively transfer diagnosis knowledge for planet gearbox. In DPTL-Net, a domain discriminator is employed to automatically learn domain asymmetry factors via adversarial learning, which can be utilized to weight the PK-MMD. The domain adaptation based on weighted PK-MMD can focus on the distribution discrepancy of source instances in the shared fault classes and filter out the instances in the outlier classes.

With the literatures surveyed above, the instance-based and feature-based DTL solutions have made significant breakthroughs to partial domain fault diagnosis, which can function as a bridge between the large-scale source domain to unknown small-scale target domain for the diagnosis knowledge transfer. However, it is obviously inappropriate to take all the labeled data as the source domain. Therefore, according to the characteristic of the target domain data, how to select the labeled source instances and determine the range of source domain from numerous low-quality industrial data is a challenging problem, which is ignored by the researcher as so far.

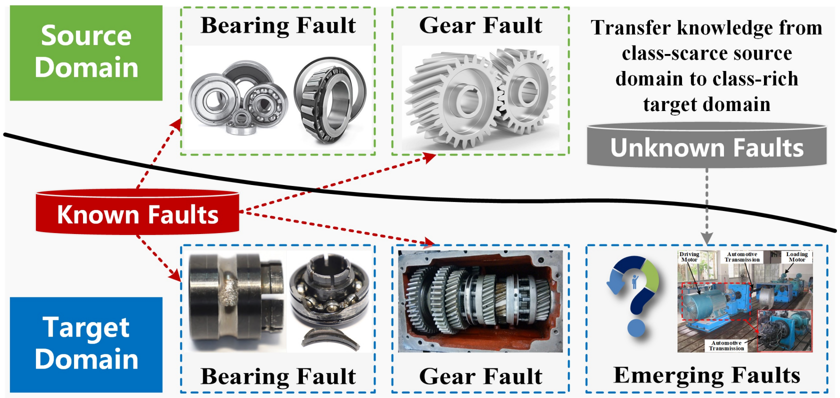


Fig. 7. An illustration of the motivation behind the scenario of emerging fault detection

### *3.4 Emerging Fault Detection*

#### 3.4.1 Motivations and goals

As illustrated in Fig. 7, the critical motivation behind this scenario is that, in practical industry applications, if IFD model can detect the unknown faults which are absent in the labeled source dataset and annotate these faults with correct labels, the IFD model will be able to precisely monitor the health conditions of machines and to continually expand its diagnosis knowledge. In the process of the emerging fault detection, the following two factors that should be taken fully into account are:

(1) Any knowledge about the faults is lacking. The unknown faults are emerging fault classes which newly occur in the target application scenario. More importantly, the unknown faults

never exist in the source domain. It is a challenging task to separate the known and unknown fault classes in an unsupervised manner.

(2) The emerging fault classes may also jeopardize the knowledge alignment between the source and target domains due to the absence of emerging faults in the source domain. In other words, negative transfer will happen if the distribution of the target domain is directly matched with that of the whole source domain.

Different from the partial domain fault diagnosis where knowledge is transferred from class-rich source domains to class-scarce target domain, the emerging fault detection aims at transferring diagnosis knowledge from class-scarce source domain to class-rich target domain. That is, as depicted in Fig. 4 (c), the main goal in this scenario is simultaneously to recognize the emerging faults as "unknown fault" classes and to classify the shared faults of two domains into the correct fault classes. Generally, unpredicted faults are prone to occur since the machines typically operate in complex and uncertain environments during long-term service. Such problem seriously restricts the practical application of the DTL-based methods. Consequently, it is an urgent demand for IFD methods to recognize emerging faults in practical engineering applications. However, there are only a few studies focusing on the emerging fault detection, for which Table III summarizes the current DL-based and DT-based solutions.

TABLE III Solutions for Emerging Fault Detection

| Application Scenarios | Categorization | References | Common algorithms used |
|---|---|---|---|
| Emerging Fault Detection | DL-based | Zhang *et al.* [196], Wang *et al.* [197], Feng *et al.* [198] | Similarity metric |
| | DTL-based | Li *et al.* [199], [200], Wang *et al.* [201], Zhang *et al.* [202], [203], Yang *et al.* [204], Li *et al.* [205], Yu *et al.* [206] | Open set domain adaptation |

### 3.4.2 Solutions for emerging fault detection

Detecting new faults during the testing scenario is one of the key steps for IFD methods when implementing the task of emerging fault detection. In terms of similarity metric learning, several DL-based solutions are established without transfer learning techniques to detecting the emerging faults [196]-[198]. For instance, Zhang *et al.* [196] proposed an emerging new labels method based on SAE (ENL-SAE) for detecting the emerging fault conditions of gearbox. The ENL-SAE forms a prior distribution of known faults with the Gaussian Distribution by utilizing the features extracted by SAE from the training samples, which can be employed to identify the unknown instances whose distribution deviate from the prior distribution of the known faults. These unknown instances are annotated with a new label as the emerging fault and used to retraining the diagnosis model. Simulation and realistic experimental results showed that the ENL-SAE can effectively recognize new faults and improve its practicality. Similarly, a deep metric learning (DML) model was proposed by Wang *et al.* [197], which has capability to classify the new fault by retrieving similarities. In DML, the raw data of each instance are firstly mapped into cosine space, and then the cosine similarity is used to retrieve the most similar fault. The methods mentioned above break through the limitations of the traditional intelligent algorithms owing to the capability of emerging fault detection [198]. However, an obvious bottleneck behind these methods is that they cannot deal with the diagnosis task under complex application scenarios where the distribution shift exists between the training and testing data. Furthermore, some of them rely on a few labeled instances of new faults.

Attempting to further break through the bottleneck mentioned above, the DTL-based solutions have been greatly developed for emerging fault detection in real industry application [199]-[206]. Inspired by the idea of Open Set Domain Adaptation (OSDA) [207], [208], Li *et al.* proposed an IFD method, called Deep Adversarial Transfer Learning Network (DATLN), for detecting the emerging faults of bearings and gearboxes [199], [200], which offered a highly successful attempts on this challenging diagnosis task. The DATLN consists of two components: a feature extractor and a classifier, which are trained by adversarial training. The feature extractor extracts features from input data, and the classifier outputs $K$+1 dimension probability, where $K$ represents the number of known faults in source domain and the $K$+1 th of the classifier output indicates the probability of the unknown fault. On the one hand, the feature extractor aligns the features extracted from the source and target domains, which can deceive the classifier. On the other hand, the classifier can build a decision boundary to recognize the unknown fault in the target domain. The experiments carried on bearing and gearbox datasets showed that the DATLN can not only align the distribution discrepancy between the different domain in the scope of the shared faults, but also can detect the emerging fault with high accuracy. Wang *et al.* [201] proposed a Deep Prototypical Networks based on DA (DPDAN), in which a prototypical layer was applied to learn the prototypes of each fault class and the classification is implemented by finding the nearest class prototype. The DPDAN is another attempt to address the problem where the fault classes of the target scenarios are partially overlapped with that of the source scenarios. Besides the aforementioned feature-based DTL methods, an OSDA method based on Instance-Level Weighted Adversarial Learning was proposed by Zhang *et al.* [202] and applied for IFD of machinery. The instance-level weighted mechanism is introduced to reflect the similarities of testing instances with known faults, therefore, the unknown faults, as well as the known faults, can be effectively identified. Admittedly, these methods are promising for the emerging fault detection and largely improve the applicability of IFD algorithms in the practical engineering. Nevertheless, a major limitation of them is that it can only detect all unknown faults as one category even if there exist multiple emerging faults.

To overcome such limitation, Li *et al.* [205] further extended the DATLN method to a Two-Stage Transfer Adversarial Network (TSTAN) for IFD of rotating machinery with multiple emerging faults. In the first stage, a DTL model is trained by the adversarial learning strategy and employed to single out the unknown fault instances as outliers from the known ones. In the second stage, an unsupervised convolutional SAE with silhouette coefficient is built to further recognize the number of the emerging faults. The TSTAN was validated on two OSDA scenarios: two and three new faults exist in the target domain respectively, and it achieved the highest diagnosis accuracy for the emerging fault detection compared with other state-of-the-art methods. To move one step forward, Yu *et al.* [206] proposed an open set fault diagnosis (OSFD) method with bilateral weighted adversarial networks (BWAN) and extreme value theory for the application scenario where the source and target domains share partially fault classes but hold its private fault classes at the same time. Such assumption is more in accordance with the case of practical engineering in industry. The experimental results on the CWRU and the Traction Motor Bearing (TMB) Dataset illustrate the superior performance of the proposed OSFD approach for emerging fault detection.

With the literature surveyed above, several excellent applications, have been witnessed the last years, addressing the challenging task of emerging fault detection for practical engineering. In terms of the more complex diagnosis task such as the across machines and sensors, however, there are still few or no solutions for the emerging fault detection. More efforts should be placed on these aspects.

## *3.5 Compound Fault Decoupling*

### 3.5.1 Motivations and goals

As illustrated in Fig. 8, the intuitive motivation behind this scenario is that, with the development of intelligent technology, the IFD model should certainly be endowed with the ability to decouple the compound fault in an intelligent manner by only leveraging upon the diagnosis knowledge learned from the data of the corresponding single faults. Such motivation, that is, intelligent compound fault decoupling, is inspired by the phenomenon that human beings are capable of separating the overlapping entities into multiple individual entities easily. As shown in the upper part of Fig. 8, taking the overlapping digits as a concrete example, humans can rapidly capture the key characteristics about each individual digit and can recognize multiple digits in the image even if digits overlap.

However, such an “easy” task is difficult for the majority of IFD algorithms. The challenges for intelligent compound fault decoupling mainly came from the following aspects:

(1) A compound fault occurs unpredictably when multiple key parts and components present defects or even damage at the same time. The monitoring signals become more complex since the fault characteristics of each component are coupled and exerted influence reciprocally, which dramatically increases the difficulties of IFD.

(2) The completeness of compound fault data within the training dataset is hard to be ensured. The practical challenge that is hardly avoidable is that it is difficult and unrealistic to

accumulate single-fault data in industrial applications, let alone to completely collect all types of compound fault data.

(3) The traditional classifier that utilizes the Softmax as the activation function of the last fully connected layer only outputs one label for a testing instance, which inherently determines that the compound fault is simplify regarded as an independent fault pattern for classification and the relationship between the compound fault and its corresponding single faults is ignored.

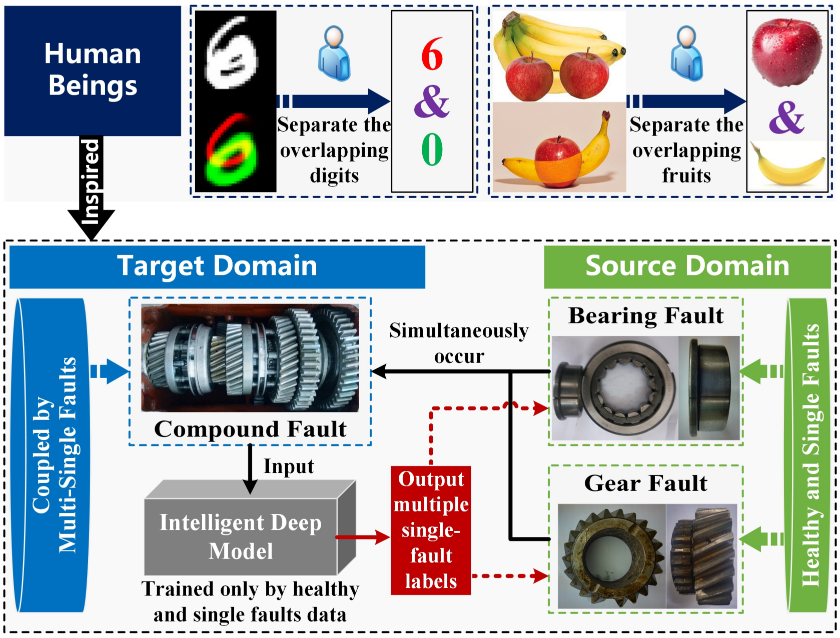


Fig. 8. Illustration of the motivation behind the scenario of compound fault decoupling

Based on the core idea of transfer learning, as illustrated in Fig. 4 (d), the goal here is to develop an IFD model for compound fault decoupling which can learn and capture useful fault characteristics from only the single-fault instances (source domain) and transfer the learned knowledge to help in making a right decoupling of compound fault instances (target domain). Following this insight, as summarized in Table IV, several successful attempts have been made for intelligent compound fault decoupling to imitate the learning ability of humans.

TABLE IV Solutions for Compound Fault Decoupling

| Application Scenarios | Categorization | | References | Common algorithms used |
|---|---|---|---|---|
| Compound Fault Decoupling | Supervised | DL-based | Huang *et al.* [222], Liang *et al.* [223], Jin *et al.* [224] | DNN with multi-label classifier |
| | Unsupervised | DL-based | Huang *et al.* [226], [227], Dibaj *et al.* [228], Xing *et al.* [229] | CNs, Triple probabilistic terms, Zero-shot learning |
| | | DTL-based | Huang *et al.* [230], [231] | TCN, DACN |

### 3.5.2 Solutions for compound fault decoupling

Compound fault diagnosis was and remains a challenging but practical task in the field of fault diagnosis. Before the widely application of IFD, the traditional methods for compound fault diagnosis generally extract the fault characteristic frequencies of each single fault from the monitoring signals of compound fault to make an accurate diagnosis by utilizing the advanced signal processing algorithms [209]-[212]. For example, a compound fault diagnosis method based on multiple enhanced space decomposition was developed by Li *et al.* [211], which can extract the characteristic features of gear defect and bearing fault simultaneously. Cui *et al.* [212] proposed a method based on the Maximum Entropy Deconvolution Adjusted (MEDA) and Adapted Dictionary-free Orthogonal Matching Pursuit (ADOMP) to isolate the compound fault coupled by the gear and bearing faults. Although these solutions can be used to monitor the health states of IE, they heavily rely on the empirical knowledge and the engineering experience of experts, which is a major obstacle for its wide application in industry.

Benefitted from the advantages of DL in representation learning and pattern recognition, some phenomenal solutions have been proposed and applied for the compound fault diagnosis [213]-[221]. For example, in [213], [214] and [215], several DBN-based IFD methods were proposed and applied to diagnose the compound faults of machinery, which mainly focus on enhancing the structure of DBN to improve the performance of diagnosis model. Shao *et al.* [216] developed a multisensory fusion strategy using a stacked wavelet AE structure with a Morlet wavelet function and applied to the collaborative fault diagnosis of planetary gearbox with compound fault. Combining with other techniques, such as adaptive separation, Euclidean matrix sample entropy and adversarial learning, CNN were developed and enhanced for intelligent compound fault diagnosis in many fields [217]-[221].

It can be seen from the publications mentioned above that most of these solutions lose sight of an importance aspect that the compound fault is anything but an individual pattern when it comes to the corresponding single faults. It is inappropriate to simply regard the compound fault as an independent fault class for fault classification. To overcome the shortcoming mentioned above, an intelligent compound fault diagnosis framework based on Deep CNN with multiple-label classifier (DCNN-MLC) was proposed by Huang *et al.* [222] and validated on a gearbox dataset. The core idea of DCNN-MLC is that the sigmoid function, which can transform the output value of each neuron into [0, 1], is employed to substitute the Softmax as the activation function of the last fully connected layer. As a result, the MLC can output single or multiple labels for a testing instance by priorly setting a confidential threshold. The DCNN-MLC is trained with the single faults and compound faults instances, which can decouple the compound fault in a supervised manner by outputting multiple labels. Such an idea has been further investigated and applied for the compound fault diagnosis of gearboxes and bearings [223], [224]. The diagnosis model with MLC is effective for compound fault decoupling by having the ability of outputting multiple labels. However, these models heavily rely on the completeness of compound fault data, suffering setbacks when the labeled data of compound faults are incomplete or even unavailable.

Aiming at eliminating the dependance of completeness of compound fault data, scientific researchers proposed several DL-based solutions for compound fault decoupling in an unsupervised manner, in which the diagnosis model is only trained on the healthy and single faults instances and then can be used to diagnose the compound fault instances [226]-[229]. For instance, inspired by the Capsule Networks (CNs), a Deep Decoupling Convolutional Neural Network (DDCNN) was proposed by Huang *et al.* [226] and applied for intelligent compound fault decoupling of an automobile transmission. In DDCNN, a decoupling classifier is constructed with two capsule layers, rather than a fully connected layer, and is optimized by an agreement-based dynamic routing algorithm, which can decouple the compound fault via outputting multiple labels. The DDCNN is a first successful effort to realize the intelligent compound fault diagnosis by transferring the knowledge learned from the data of healthy and single faults in the scenario that the compound fault data are unavailable during model training. To achieve a common goal, a similar attempt has been investigated by Dibaj *et al.* [228]. The main idea of the method proposed in [228] is that the CNN is trained without the compound fault data, and triple probabilistic conditions are used to restrict the output label of the classifier by judging whether the acquired probabilities of each neuron satisfy these conditions. Thus, the untrained compound fault can be recognized in an intelligent manner. A label description space embedded model for intelligent fault diagnosis (LDS-IFD) was proposed by Xing *et al.* [229] to recognize the compound faults just using the single-faults data during the model training, which is validated by two datasets collected from bearing and planetary gearbox. Admittedly, these solutions have brought successful breakthroughs in intelligent compound fault diagnosis because they eliminate an important problem: the dependance of the completeness of the compound fault data. Nevertheless, the methods mentioned above still lack a robust generalization performance when they encounter a varying and harsh environment, restricting its further practical application in industry.

With the help of transfer learning techniques, the DTL-based solutions for intelligent compound fault decoupling have been attracted increasing attention and application in recent years. The compound fault diagnosis models are getting more generalizable and accessible under varying WCs [230], [231]. Huang *et al.* [230] further proposed a Transferable CN (TCN) for decoupling compound fault of rotating machinery under varying WCs. The TCN is a variant of DDCNN, which can reduce the distribution discrepancy between the source and the target domains by introducing the MMD into the last layer of the feature extractor and the decoupling classifier, respectively. The experimental results demonstrated that the TCN outperforms the DDCNN for the compound fault decoupling under varying WCs. To improve the practicality of diagnosis model, Huang *et al.* [231] further relaxed the assumption on training data by considering that the data cannot be obtained in advance for some special and extreme WCs, and proposed a Deep Adversarial Capsule Network (DACN) which embeds the domain generalization task into the intelligent compound fault diagnosis task. The DACN consists of three parts: the feature extractor (FE), the decoupling classifier (DC) and the multidomain classifier (MC), which is designed for representation learning, compound fault decoupling and multidomain adaptation, respectively. Using the single fault data collected under multiple WCs, the adversarial training strategy is employed to train the DACN. The comprehensive experiments carried on an automobile transmission demonstrates that the DACN is endowed with the ability to decouple the compound fault in an intelligent manner, as well as the ability of strong generalization performance across unseen working condition.

Through the literature surveyed above, the current solutions for intelligent compound fault decoupling have to some extent addressed the two problems: the dependence of data completeness and the lack of robust generalization performance. However, it seems that few studies focus on the more complex industrial scenarios, e.g., the compound fault coupled with three or more single faults, which might be more in accordance with the practical application in industry.

## 4 Suggestions to Select DTL Algorithms for IFD in Industry Applications

After the comprehensive literature survey in Section III, the recent development of DTL approaches in the field of IFD is systematically presented and discussed from the perspective of different industrial application scenarios. To provide a constructive guide for the readers who want to solve the practical industry problems via using DTL-based IFD methods, in this section, the general procedure of IFD based on DTL is concluded, as well as the suggestions to select DTL algorithms for IFD in industry applications.

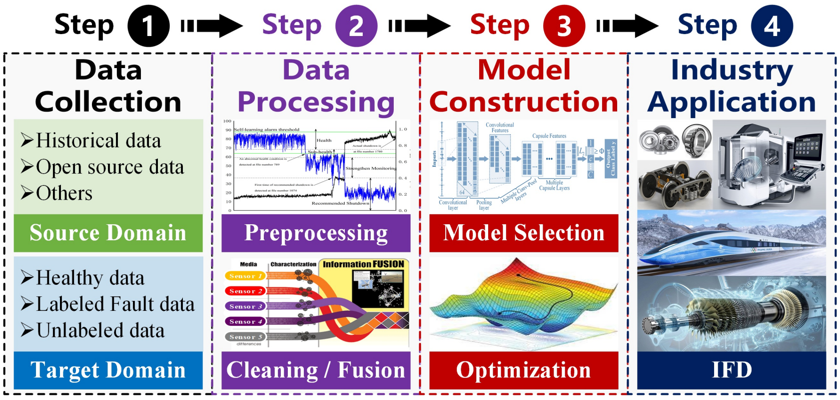


Fig. 9. General procedure of IFD based on DTL

### *4.1 General Procedure of IFD Based on DTL*

As demonstrated in Fig. 9, the general procedure of IFD based on DTL includes four crucial steps: data collection, data processing, model construction, and industry application. Following these steps, a practical IFD project can be implemented in industry applications.

***Step 1: Data Collection.*** In a systematic DTL approach to apply IFD methodologies for a specific task, the first step mostly focuses in collecting the available data from the source and the target domain. Before a DTL algorithm is utilized to accomplish the specific task, it is absolutely necessary and extremely beneficial to familiarize the characteristics of the collected monitoring data and the information of the interested equipment in terms of the key components, WCs, service intensity and all other important physical attributes. In other words, no matter what type of data, such as vibration, electrical and acoustic emission signals, can be collected, the quantity and the quality of data will be fundamental for the subsequent steps in developing an effective solution with dependable diagnosis accuracy via combining appropriate algorithms. As mentioned before, one of the advantages of DTL is that the labeled data in the related but different domains can be used to help training the target model. Therefore, the characteristics of data in the source domain largely affect the performance of the target model. Generally, there are mainly two ways to collect the source domain data. The first way is to use the labeled historical data, collected from similar machines, while the second one is to select similar data from open-source industrial big data. The public datasets, which have been provided by the PHM data challenges that have been held by the PHM Society since 2008, are real data collected from practical industry scenarios. All the datasets are fully opened to all researchers and covered the diagnostics and prognostics tasks in many industry fields, and can be downloaded by the website of PHM Society [1].

***Step 2: Data Processing.*** Contrary to the data collected from laboratory experiments, real industrial big data typically have four main characteristics: large volume, low value density, multi-source and heterogeneous data structure, and monitoring data stream [27]. Therefore, data processing is one of the key steps for improving the performance of the IFD model. Essentially, for an intelligent learning process, garbage data in inevitably leads to garbage results out. There is no one absolute way to prescribe the exact steps in data processing because the process would be better to combine some background information in the specifical scenario. Data cleaning, normalization and data fusion are popular and effective techniques for the processing of original industrial data [232], [233], which can remove errors and inconsistencies and improve the quality of the data that will be used to train the target model.

***Step 3: Model Construction.*** Along with the continuous progress in manufacturing industry, many advanced algorithms have been introduced, developed and benchmarked to implement the diagnostic and prognostic tasks in a supervised or unsupervised manner. Different algorithms are able to handle different problems depending on its adaptability. Therefore, a crucial step for developing an effective solution for IFD in industry scenarios is to select and adopt the most appropriate algorithms, based on available data and target tasks. The suggestions to select the DTL algorithms for IFD will be

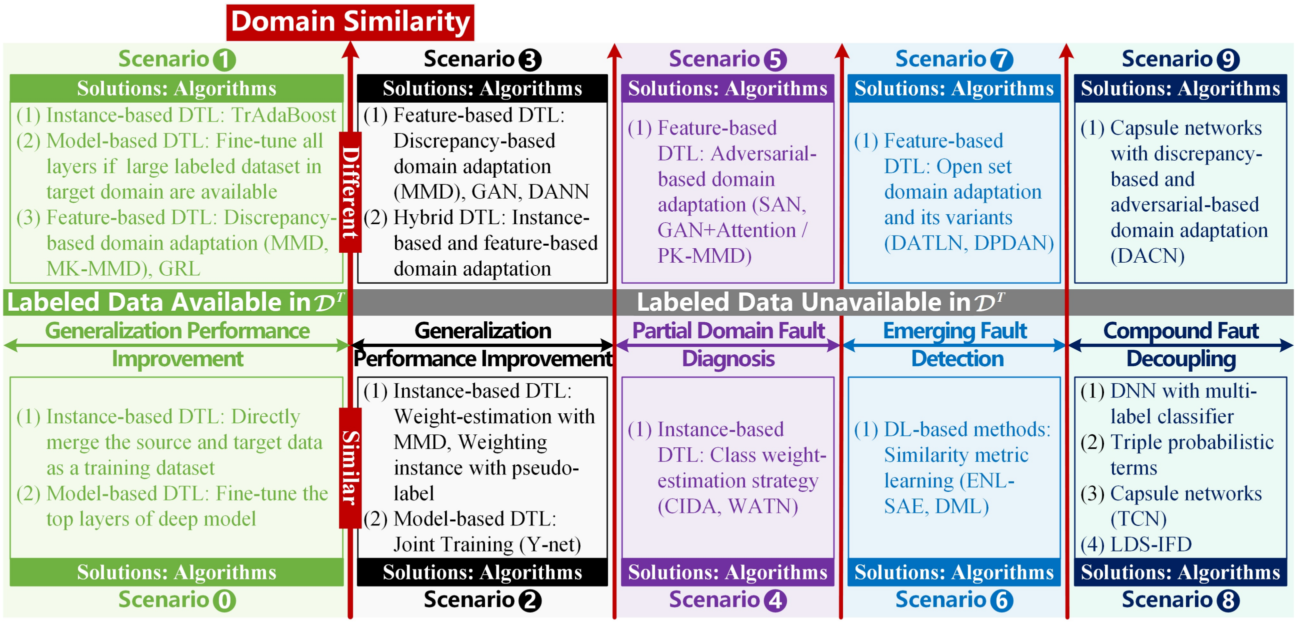


Fig. 10. Suggestions to select appropriate algorithms for practical industry applications

detailly introduced in Section IV, Part B. Once the DTL algorithm is determined, the target model can be optimized according to the source and target data via using gradient-based optimizers, such as Stochastic Gradient Descent (SGD) [234], Adaptive Gradient (Adagrad) [235] and Adaptive Moment Estimation (Adam) [236].

***Step 4: Industry Applications.*** After the diagnosis model has been constructed and optimized by feeding with the data, it can be ready for further application to monitor the health states of target equipment. In this step, it is important to use an Internet of Things (IoT) platform to support the IFD system to convey the useful information to the engineers through visualization tools.

### *4.2 Suggestions to Select DTL Algorithms for IFD*

After the general procedure for IFD methods has been systematically introduced, this subsection will offer some guidance and suggestions to select DTL algorithms according to specific scenarios of the industry applications.

It is an acknowledged truth that there is no general algorithm regarding to the IFD in industry application. At the beginning of selecting DTL algorithms, there are two factors which should be first considered. The first factor is to consider the circumstance of whether the labeled data are available in the target domain, while another one is to evaluate the similarity between the source and the target domain. As illustrated in Fig. 10, the corresponding algorithm selection strategies are provided according to the above two factors and the different industry application.

As for the generalization performance improvement, the application scenario can be further divided into four sub-scenarios: the scenario 0 to 3, in which the appropriate algorithms can be selected by considering the following suggestions.

***Scenario 0:*** Labeled data are available in the target domain, and the source and target domains are similar. Any DTL algorithms may work well in this situation. But, the most efficient and optimal option will be the model-based DTL algorithms, more specifically, the fine-tuning strategy. Since the labeled data are available, the target model can be trained in a supervised manner. Furthermore, since the gap between the source and target domain is small, the knowledge learned from the source domain will be also suitable to the target domain. therefore, it should be enough to directly merge the source and the target data as a training dataset or to fine-tune the top layers of the pre-trained deep model.

***Scenario 1:*** Labeled data are available in the target domain, and the source and target domains are different. In this situation, intuitively, instance-based DTL algorithms, e.g., TrAdaBoost, can be used to single out the similar instances in the source domain to augment the training dataset for the target task. However, such algorithms will be unsuccessful if the data are largely different between the source and the target domains. If the labeled data is sufficient for target model training, fortunately, another solution is to fine-tune all the layers of the pre-trained model. Further, if the labeled data is insufficient, it would be promising solutions to select the feature-based DTL algorithms, such as the discrepancy-based and non-generative domain adaptation.

***Scenario 2:*** Labeled data are unavailable in the target domain, and the source and target domains are similar. Implementing the target task in this situation will be a little bit more difficult than that of in the Scenario 0 due to the fact that the target instances are not annotated. However, since the instances in source and target are similar, the instance-based DTL algorithms, such as the weight-estimation based on kernel embedding techniques and the heuristic weighting strategy, would be a good choice to select out the positive instances in the source domain to help training the target model.

***Scenario 3:*** Labeled data are unavailable in the target domain, and the source and target domains are different. In this situation, the model-based and the instance-based DTL algorithms can hardly improve the generalization performance of the deep model because the label information of the target instances is lacking and the gap between the source and the target domain is large. Therefore, this situation will lead the engineer to the feature-based DTL algorithms (discrepancy-based and adversarial-based domain adaptation). The hybrid DTL algorithms which combine the instances-based and the feature-based domain adaptation will also be a promising tool in this scenario.

As for the other three application scenarios, their basic assumption is that the labeled data in the target domain are unavailable. Therefore, the main factor that should be considered for selecting algorithms is the domain similarity. Each application scenario can be further divided into two sub-scenarios, that is, scenario 4 & 5 (Partial Domain Fault Diagnosis), scenario 6 & 7 (Emerging Fault Detection), and scenario 8 & 9 (Compound Fault Decoupling).

***Scenario 4:*** Source and target domains are similar. It is important to use the similar data as the source dataset for partial domain fault diagnosis. For example, data collected from similar working conditions or same machines are perhaps the best option. As a result, the instance-based DTL algorithm, e.g., the class weight-estimation strategy, is recommended to single out the instances in the shared classes, and then used to train the target model.

***Scenario 5:*** Source and target domains are different. In this situation, since the similar source data are difficult to be collected, it is a potential solution to use the different but related data collected from related industry applications. Considering the demand for reducing the domain discrepancy and avoiding the negative transfer, the feature-based DTL algorithms, especially the adversarial-based domain adaptation (SAN or GAN+PK-MMD), should be given priority.

***Scenario 6:*** Source and target domains are similar. As for emerging fault detection, if the instances in the target domain are similar to those in the source domain, an effective method would be to apply the traditional DL-based methods that detect the new faults by calculating the similarity metric between the testing and the labeled instances.

***Scenario 7:*** Source and target domains are different. In the practical industry application, it is more common that the domain shift exists between the source and the target domains. Therefore, the OSDA algorithm and its variants would be more practical and effective to address the problems of emerging fault detection.

***Scenario 8:*** Source and target domains are similar. Even under an identical working condition and the same machine, it is a challenging task to intelligently decouple the compound fault via using a target model that just trained by single fault instances. If the labeled compound fault data is available, the DNN with MLC can be trained in a supervised manner and further applied for compound fault detection. Otherwise, the deep model can be trained only using the normal and single fault data, and then a rule (e.g., Triple probabilistic terms) can be used to restrict the outputting labels of classifier. From the results shown in literature, the capsule network is the best choice for compound fault decoupling.

***Scenario 9:*** Source and target domains are different. Since the domain shift is introduced with the varying environments, the DNN-based algorithms perform not well in this situation. Up to now, an effective and promising solution is to combine the capsule networks with the feature-based DTL algorithms, such as the discrepancy-based and the adversarial-based domain adaptation.

## 5 Future Challenges and Trends in IFD of Industrial Machinery

An obvious conclusion can be drawn from the comprehensive survey and discussion that, despite the fact that IFD algorithms based on DTL have made successful breakthroughs in many industry applications, there is still a long way to go until it is widely adopted in practical manufacturing industry systems. This is mainly because the performance of DTL algorithms lags far behind the requirements of manufacturing industry which places more emphasis on stability, standardization, accuracy, and repeatability. Before the IFD technologies can be fully embraced and applied in real world industry systems, the researchers in the related field should put significant effort into overcoming the following challenges.

### *5.1 Stability and Reliability*

Historically, the generalization performance of IFD model has been significant improved by leveraging upon transfer learning techniques. However, the current IFD methods based on DTL could only accomplish the well-defined transfer tasks that often have restrictions on WCs, machines, and other hypotheses, which lead to the fact that the IFD model is not yet robust enough in dealing with uncertain circumstances. For a trained IFD model, an uncertain change in the input could cause a large change in the output [237]. Furthermore, most IFD algorithms published in the papers had not been verified as reproducible [238] due to the complexity of model training process and the numerous hyperparameters. In fact, there are many uncertain deviations caused by human or non-human factors during the long-term service of IE, and such deviations will directly affect the robustness, generalization performance and reproducibility of IFD algorithms, resulting in their low stability and reliability in practical industry scenarios. Therefore, it is and remains a challenging task to improve the stability and reliability of IFD algorithms for the technology to truly be applied in practical manufacturing industry systems, which requires further breakthroughs in not only the improvement of generalization performance, but also the reproducibility of diagnosis results.

### *5.2 Interpretability of Deep Model*

Although DTL-based IFD methods have made phenomenal achievements in mechanical fault diagnostics and prognostics, an acknowledged limitation is that these methods have been perceived as black box techniques and are not interpretable, which does not provide a convincing insight into how and why they can make the final decision [239]. This issue may not only put in doubt the credibility of the decision itself, but also lacks compelling evidence to convince companies or industry that these techniques will work repeatedly. Applications in industry have strict requirements for safety and accuracy, and need to explain the reasonableness of the prediction decisions. As a result, the application of the DTL-based IFD methods in manufacturing industry are very limited. In recent years, fortunately, the theory of interpretable machine learning has captured increasing attention from the academic researchers. One way to make intelligent algorithms interpretable is to use only interpretable models, such as Naïve Bayes Classifier and K-Nearest Neighbors, which typically have the limitation that the performance of these model is inferior to other intelligent models [241]. Therefore, in-depth theoretical research should be placed putting more emphasis on opening the "black box" and increasing the transparency of IFD model. Besides the theoretical research, another potential research trend in recent years is to combine the IFD model with a physical/statistical model which is supported by rigorous theory. With the help of the domain knowledge in the physical/statistical model, the "black box" of the IFD model can be partly opened, and it would be easily understood how decisions are reached step-by-step.

### *5.3 Hyperparameters of Deep Model*

Generally, the architecture and hyperparameters of deep model significantly impact on the performance of DTL-based IFD methods. Therefore, it is a crucial step to select the hyperparameters during designing an effective solution with DTL-based IFD algorithms. However, there are no industry consensus on what the ways of selecting hyper-parameters works best. The hyper-parameters are typically selected in most publications via manual setting and experimental validation based on the grid search technique, which is a time-consuming way to ensure the model achieves the optimum performance. In the future, automatic machine learning might be an effective solution to solve such problem [242].

### *5.4 Capacity of Data Processing*

With regard to industry data, the challenges facing IFD right now mainly comes from the following aspects:

***(1) Data Quality.*** The performance of IFD models still depend heavily on the quantity and the quality of historical instances in the source domain, and annotating the industry data requires more engineering experience. In practical industry, it is often the case that, with more smart sensors embedded in machines and advances of measurement technology, large volume of monitoring data can be easily accumulated, but there have problems in data quality, such as lacking correct maintenance records, missing key parameters related to target components, existing misalignment of different variables, and coupling with strong background noise. Andrew Ng, a famous professor in Stanford, points out that the AI systems equals the integration of code (model/algorithm) and data, where the 80/20 rule for the data processing vs model training might be the right balance to achieve success. Therefore, it is necessary to monitor and improve the data quality before developing the IFD solution in practical application.

***(2) Imbalanced Data.*** It is a common case that, in the era of big industry data, the monitoring data of each health state are imbalanced. For ensuring the security and efficiency of production, the IE typically works under healthy conditions. As a result, the fault instances have a much lower chance of appearing than the healthy instances. This makes the data whether in source or target domain having an imbalanced distribution, which in turn makes the IFD model tending to learn biased decision boundaries that have a poorer diagnosis performance over the fault classes compared to the healthy class. Despite the fact that some publications have been focused on the problem of imbalanced data, it is difficult for the proposed solutions to deal with the imbalanced problems in more complex and uncertain industry environment. Therefore, to endow IFD algorithms with the ability to learn the discriminative representation from an extreme imbalanced dataset, more efforts would be necessary to simulate the knowledge transfer process in which humans can correctly guess that an object may belong to the class which share some physical characteristics, instead of brutally training the IFD model with "big data". Following this insight, the few-shot or zero-shot learning, which is inspired by the phenomenon that human beings can learn a new object with only a few instances or even without any instances, are the promising research trends for solving such issue in practical industry application of IFD.

***(3) Heterogeneous Data.*** The industry factory is a typical multiple source heterogeneous data environment. For instance, in wind farms, there are large amount of multiple source heterogeneous data, such as the high-frequency data (current, acoustic emission and vibration signals) and low-frequency data (environmental index, working condition information and control parameters), have been collected from the Supervisory Control and Data Acquisition (SCADA) system and the Condition Monitoring System (CMS). However, as surveyed in the previous sections, the majority of DTL-based IFD methods focus on cases where instances in source and target domains are homogeneous data (e.g., vibration data). The obvious limitation of existing IFD methods is that, if the target sensor malfunctions unexpectedly, the CMS will be out of operation, which in turn could lead to serious catastrophes. Since the multi-source heterogeneous data can provide different information for the same health states of machine, it is possible to transfer diagnosis knowledge from one sensor data to another ones, which may greatly improve the stability and reliability of IFD algorithms. Furthermore, up to now, few studies focus on the heterogeneous transfer learning in the field of IFD. Therefore, heterogeneous transfer learning between multiple sensors would also be one of the future research trends that more attention should be paid.

***(4) Data Privacy and Protection***. In the era of digital and intelligence, industry data is one of the most important assets a company has. For that reason alone, data privacy and protection should be a top priority for any company. It is difficult to reach an agreement and share labeled data among different companies and factories, which in turn results in data fragmentation and isolation. As a result, such restriction poses significant obstacles for the applications of the IFD algorithms in the practical industry. Therefore, how to solve the problem of data fragmentation and isolation while considering and complying with the restriction of data privacy and protection is one of the major challenges for the IFD algorithms to truly accommodate a wider range of application in practical industry. One potential research trend for addressing the above issues is to combine the federated learning and DTL to build and train an effective and accurate IFD model [243].

### *5.5 Challenges in Transfer Learning*

To design an effective DTL-based algorithms, there are still several key challenges should be placed more efforts on.

***(1) Identifying the Appropriate Source Domain***

The previous survey on DTL-based IFD algorithms elaborates several ways to transfer source domain knowledge for practical industrial applications, however, identifying an appropriate source domain is still a challenging problem due to the challenges caused by big industrial data. For example, for many industrial applications, it is difficult to find an appropriate source domain that includes sufficient training instances annotated with precise label information for implementing target tasks. Even worse, it may be unrealistic to find any failure data from similar or related industrial application. With the rapid development of digital technology, such as Digital Twins, one promising way is to utilize the simulation or generation techniques to generate training data as the source domain in such a scenario. In addition, transferring the knowledge from multiple source domains has been attracted more and more attentions recently.

***(2) Avoiding Negative Transfer***

Once the source domain is determined, avoiding negative transfer is also a challenging problem during building a DTL model. As illustrated in Section 3, although there are several tricks have been proposed for avoiding negative transfer, it should be highlighted that negative transfer still needs further systematic investigation. One of effective measures to improve the performance of the DTL-based IFD model in industrial scenarios is to transfer only the common knowledge that can contribute to the target learning task and to avoid negative transfer at the same time. For example, developing an accurate “distance” metrics between the domains might be a feasible solution for avoiding negative transfer since the existing metrics used in feature-based DTL are not powerful enough in developing a perfect transfer learning application.

***(3) Assessing Transferability***

Assessing the transferability across domains in quantitative is another challenging problem during developing a DTL-based IFD method in industrial scenario. However, as so far, there is still few publications focusing on assessing transferability between the source and target domains mathematically. We confident that assessing transferability across domain will be a significantly important research trend in the future, which will enhance the performance of the DTL-based methods and further boost the application of DTL in industrial scenarios.

### *5.6 Computation and Energy Efficiency*

According to the aforementioned literature survey, it is generally the case that the DTL-based IFD methods suffer from the high requirement of computational source and speed. The inefficient and large computation in deep model has hindered the successful application of IFD methods in real-time data analytics. However, the capability of real-time monitoring is fundamental to PHM systems, which can improve the security of machinal systems, identify potential faults as soon as they occur, allow for early maintenance, and avoid systems failures. Therefore, the real-time IFD algorithms should be encouraged to be investigated to ensure real-time decision-making for monitoring the incipient damages or unexpected faults [244]. Techniques, including efficient neural network compression, incremental learning and deep reinforcement learning [245], are potential research directions to facilitate the real-time ability of DTL based IFD algorithms.

## 6 Conclusions

In this survey article, the theory and strategies of DTL methods have been summarized from the algorithm perspective, which gives the basic definitions related to DTL and explain how the TL technologies can help improving the performance of DL model. The state-of-the-art applications of DTL-based IFD approaches have also been overviewed from the perspective of practical industrial applications, in which the four major application scenarios: generalization performance improvement, partial domain fault diagnosis, emerging fault detection, and compound fault decoupling, are formulated and fully discussed. Thereafter, the suggestions for the selection of DTL algorithms for a new IFD project have been detailed, as well as the future challenges and potential trends. This review article not only leads readers to easily understand the current state-of-the-art DTL techniques related to IFD and to quickly design an effective solution for solving IFD problems in practice, but also provides the main challenges facing IFD until it has wide adoption in practical manufacturing industry systems, as well as the future research trends, for researchers and scholars.

**Acknowledgment**

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1702400, in part by National Natural Science Foundation of China under Grant 51875208, in part by Key-Area Research and Development Program of Guangdong Province under Grant 2019B010154002. The last author acknowledges the support of Flanders Make and VLAIO in the frames of DGTwin Prediction project, the support of Research Foundation - Flanders (FWO) for its support through the ROBUSTIFY research grant no. S006119N and the Flemish Government under the “Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen” programme.

## References


[1] J. Lee, Industrial AI, Singapore:Springer, 2020.

[2] J. Wang, L. Ye, RX. Gao, C. Li, and L. Zhang, "Digital twin for rotating machinery fault diagnosis in smart manufacturing," *International Journal of Production Research,* vol.57, no.12, pp: 3920-3934, 2019.

[3] J. Wan, X. Li, H. -N. Dai, A. Kusiak, M. Martínez-García and D. Li, "Artificial-Intelligence-Driven Customized Manufacturing Factory: Key Technologies, Applications, and Challenges," *Proceedings of the IEEE*, doi: 10.1109/JPROC.2020.3034808.

[4] R. Yan, R. X. Gao and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," *Signal Process.*, vol. 96, pp. 1-15, Mar. 2014.

[5] X. Yang, K. Ding, and G. He, “Phenomenon-model-based AM-FM vibration mechanism of faulty spur gear,” *Mech. Syst. Signal Process.*, vol. 134, Dec. 2019, Art. no. 106366.

[6] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, “Machinery health prognostics: A systematic review from data acquisition to RUL prediction,” *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, 2018.

[7] R. Huang, J. Li, S. Wang, G. Li and W. Li, "A Robust Weight-Shared Capsule Network for Intelligent Machinery Fault Diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6466-6475, Oct. 2020.

[8] Datong Liu, Yuchen Song, Lyu Li, Haitao Liao, Yu Peng. On-line life cycle health assessment for lithium-ion battery in electric vehicles, *Journal of Cleaner Production (JCLP)*, 199: 1050-1065, 2018.

[9] B. Zhang, S. Zhang, and W. Li, “Bearing performance degradation assessment using long short-term memory recurrent network,” *Comput. Ind.*, vol. 106, pp. 14–29, Apr. 2019.

[10] Z. Chen, K. Gryllias, and W. Li, “Mechanical fault diagnosis using convolutional neural networks and extreme learning machine,” *Mech. Syst. Signal Process.*, vol. 133, Nov. 2019, Art. no. 106272.

[11] X. Hong, B. Zhang, Y. Liu, H. Qi and W. Li, “Deep-learning-based guided wave detection for liquid-level state in porcelain bushing type terminal,” *Structural Control and Health Monitoring*, vol. 28, no. 1, Jan. 2021, e2651, DOI: https://doi.org/10.1002/stc.2651.

[12] Li X, Yang Y, Shao H, et al. Symplectic weighted sparse support matrix machine for gear fault diagnosis[J]. Measurement, 2021, 168: 108392.

[13] Z. Chen, J. Xu, T. Peng and C. Yang, "Graph Convolutional Network-Based Method for Fault Diagnosis Using a Hybrid of Measurement and Prior Knowledge," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2021.3059002.

[14] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, May 2015.

[15] S. Khan, and T. Yairi, “A review on the application of deep learning in system health management,” *Mech. Syst. Signal Process.*, vol. 107, pp. 241–265, 2018.

[16] Li Q, Ding X, He Q, et al. Manifold Sensing-Based Convolution Sparse Self-Learning for Defective Bearing Morphological Feature Extraction[J]. IEEE Trans. Ind. Informatics, 2021, 17(5): 3069-3078.

[17] S. Zhang, Z. Sun, C. Li, D. Cabrera, J. Long and Y. Bai, "Deep Hybrid State Network With Feature Reinforcement for Intelligent Fault Diagnosis of Delta 3-D Printers," IEEE Transactions on Industrial Informatics, vol. 16, no. 2, pp. 779-789, Feb. 2020.

[18] S. Zhang, Z. Sun, M. Wang, J. Long, Y. Bai and C. Li, "Deep Fuzzy Echo State Networks for Machinery Fault Diagnosis," IEEE Transactions on Fuzzy Systems, vol. 28, no. 7, pp. 1205-1218, July 2020.

[19] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[20] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43-76, Jan. 2021.

[21] C. Tan, F. Sun and T. Kong, "A survey on deep transfer learning," *Proc. Int. Conf. Artif. Neural Netw.*, pp. 270-279, 2018.

[22] C. Li et al., "Fusing convolutional generative adversarial encoders for 3D printer fault detection with only normal condition signals," *Mech. Syst. Signal Process.*, vol. 147, Jan. 2021, Art. no. 107108

[23] S. Zhang, X. Duan, C. Li, M. Liang, "Pre-classified reservoir computing for the fault diagnosis of 3D printers," *Mech. Syst. Signal Process.*, vol. 146, Jan. 2021, Art. no. 106961

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

[25] R. Liu, B. Yang, E. Zio and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vol. 108, pp. 33-47, Aug. 2018.

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

[27] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020.

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

[29] R. Yan, F. Shen, C. Sun and X. Chen, "Knowledge Transfer for Rotary Machine Fault Diagnosis," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8374-8393, Aug. 2020.

[30] J. Y. Jiao, M. Zhao, J. Lin and K. X. Liang, "A comprehensive review on convolutional neural network in machine fault diagnosis," *Neurocomputing*, vol. 417, pp. 36-63, 2020.

[31] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, ''1D convolutional neural networks and applications: A survey,'' *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107398.

[32] Zhao Z, Wu J, Li T, et al. Challenges and Opportunities of AI-Enabled Monitoring, Diagnosis & Prognosis: A Review[J]. Chinese Journal of Mechanical Engineering, 2021, 34(1): 1-29.

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

[34] S. Lu, R. Yan, Y. Liu and Q. Wang, "Tacholess Speed Estimation in Order Tracking: A Review With Application to Rotating Machine Fault Diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 7, pp. 2315-2332, July 2019.

[35] A. Stetco, F. Dinmohammadi, X. Zhao, V. Robu, D. Flynn, M. Barnes, et al., "Machine learning methods for wind turbine condition monitoring: A review," *Renew. Energ.*, vol. 133, pp. 620-635, Apr. 2019.

[36] C. Li, S. Zhang, Y. Qin and E. Estupinan, "A systematic review of deep transfer learning for machinery fault diagnosis," *Neurocomputing*, vol. 407, pp. 121-135, Sep. 2020.

[37] S. Zhang, Z. Sun, J. Long, C. Li, and Y. Bai, "Dynamic condition monitoring for 3D printers by using error fusion of multiple sparse autoencoders," Comput. Ind., vol. 105, pp. 164–176, Feb. 2019.

[38] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, Transfer Learning. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[39] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf and G. R. Lanckriet, “Hilbert space Embeddings and metrics on probability measures,” *J. Mach. Learning Res.*, vol. 99, pp. 1517–1561, 2010.

[40] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, “Direct importance estimation for covariate shift adaptation,” *Ann. Inst. Stat. Math.*, vol. 60, no. 4, pp. 699–746, Dec. 2008.

[41] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” *in Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 193–200.

[42] M. Long, Y. Cao, J. Wang, and M. I. Jordan, “Learning transferable features with deep adaptation networks,” *in Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[43] B. Sun and K. Saenko, “Deep Coral: Correlation alignment for deep domain adaptation,” *in Proc. European Conf. Computer Vision*, 2016, pp. 443–450.

[44] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.

[45] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[46] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” *in Proc. ICML*, 2015, pp. 1180–1189.

[47] H. Chen, Y. Li and D. Su, "Discriminative Cross-Modal Transfer Learning and Densely Cross-Level Feedback Fusion for RGB-D Salient Object Detection," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4808-4820, Nov. 2020.

[48] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, vol. 7, pp. 19143-19165, 2019.

[49] D. Wang et al., "Softly Associative Transfer Learning for Cross-Domain Classification," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4709-4721, Nov. 2020.

[50] Z. Zhang, H. Chen, S. Li and Z. An, "Unsupervised domain adaptation via enhanced transfer joint matching for bearing fault diagnosis," *Measurement*, vol. 165, Dec. 2020.

[51] F. Shen, R. Langari, and R. Yan. "Exploring Sample/Feature Hybrid Transfer for Gear Fault Diagnosis Under Varying Working Conditions." *J. Comput. Inf. Sci. Eng.*, vol. 20, no. 4, 2020.

[52] Y. Song, Y. Li, L. Jia and M. Qiu, "Retraining Strategy-Based Domain Adaption Network for Intelligent Fault Diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6163-6171, Sept. 2020.

[53] Y. Pan, F. Mei, H. Miao, J. Zheng, K. Zhu and H. Sha, "An approach for HVCB mechanical fault diagnosis based on a deep belief network and a transfer learning strategy," *J. Elect. Eng. Technol.*, vol. 14, no. 1, pp. 407-419, Jan. 2019.

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

[55] J. Zhou, X. Yang, L. Zhang, S. Shao, G. Bian, "Multisignal VGG19 Network with Transposed Convolution for Rotating Machinery Fault Diagnosis Based on Deep Transfer Learning," *Shock and Vibration*, vol. 2020.

[56] T. Lu, F. Yu, B. Han and J. Wang, "A Generic Intelligent Bearing Fault Diagnosis System Using Convolutional Neural Networks With Transfer Learning," *IEEE Access*, vol. 8, pp. 164807-164814, 2020.

[57] R. Zhang, H. Tao, L. Wu and Y. Guan, "Transfer Learning With Neural Networks for Bearing Fault Diagnosis in Changing Working Conditions," *IEEE Access*, vol. 5, pp. 14347-14357, 2017。

[58] M. J. Hasan and J.-M. Kim, "Bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning," *Appl. Sci.-Basel*, vol. 8, no. 12, Dec. 2018.

[59] M. J. Hasan, M. M. M. Islam and J.-M. Kim, "Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions," *Measurement*, vol. 138, pp. 620-631, May 2019.

[60] T. Han, C. Liu, W. Yang and D. Jiang, "Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions," *ISA Trans.*, vol. 93, pp. 341-353, Oct. 2019.

[61] Z. He, H. Shao, X. Zhang, J. Cheng and Y. Yang, "Improved Deep Transfer Auto-Encoder for Fault Diagnosis of Gearbox Under Variable Working Conditions With Small Training Samples," *IEEE Access*, vol. 7, pp. 115368-115377, 2019.

[62] Z. He, H. Shao, X. Zhong and X. Zhao, "Ensemble transfer CNNs driven by multi-channel signals for fault diagnosis of rotating machinery cross working conditions", *Knowl.-Based Syst.*, vol. 207, Nov. 2020.

[63] B. Zhao, X. Zhang, Z. Zhan, and S. Pang, “Deep multi-scale convolutional transfer learning network: A novel method for intelligent fault diagnosis of rolling bearings under variable working conditions and domains,” *Neurocomputing*, vol. 407, pp. 24–38, Sep. 2020.

[64] Y. Du, A. Wang, S. Wang, B. He, G. Meng, "Fault Diagnosis under Variable Working Conditions Based on STFT and Transfer Deep Residual Network", *Shock and Vibration*, vol. 2020.

[65] Z. Chen, J. Cen and J. Xiong, "Rolling Bearing Fault Diagnosis Using Time-Frequency Analysis and Deep Transfer Convolutional Neural Network," *IEEE Access*, vol. 8, pp. 150248-150261, 2020.

[66] F. Li, J. Chen, J. Pan and T. Pan, "Cross-domain learning in rotating machinery fault diagnosis under various operating conditions based on parameter transfer," *Meas. Sci. Technol.*, vol. 31, no. 8, 085104, 2020.

[67] Z. Wang, Q. Liu, H. Chen and X. Chu, "A deformable CNN-DLSTM based transfer learning method for fault diagnosis of rolling bearing under multiple working conditions," I*nternational Journal of Production Research*, 2020. DOI: 10.1080/00207543.2020.1808261

[68] H. Shao, M. Xia, G. Han, Y. Zhang and J. Wan, "Intelligent Fault Diagnosis of Rotor-Bearing System Under Varying Working Conditions With Modified Transfer Convolutional Neural Network and Thermal Images," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3488-3496, May 2021.

[69] X. Cao, B. Chen and N. Zeng, "A deep domain adaption model with multi-task networks for planetary gearbox fault diagnosis," *Neurocomputing*, vol. 409, pp. 173-190, Oct. 2020.

[70] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang and T. Zhang, "Deep Model Based Domain Adaptation for Fault Diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296-2305, March 2017.

[71] X. Li, W. Zhang and Q. Ding, "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning," *Neurocomputing*, vol. 310, pp. 77-95, 2018.

[72] Z. Tong, W. Li, B. Zhang, F. Jiang and G. Zhou, "Bearing Fault Diagnosis Under Variable Working Conditions Based on Domain Adaptation Using Feature Transfer Learning," *IEEE Access*, vol. 6, pp. 76187-76197, 2018.

[73] Z. Tong, W. Li, B. Zhang and M. Zhang, "Bearing fault diagnosis based on domain adaptation using transferable features under different working conditions," *Shock Vib.*, vol. 2018, Jun. 2018.

[74] B. Zhang, W. Li, X. Li and S. Ng, "Intelligent Fault Diagnosis Under Varying Working Conditions Based on Domain Adaptive Convolutional Neural Networks," *IEEE Access*, vol. 6, pp. 66367-66384, 2018.

[75] D. Xiao, Y. Huang, L. Zhao, C. Qin, H. Shi and C. Liu, "Domain Adaptive Motor Fault Diagnosis Using Deep Transfer Learning," *IEEE Access*, vol. 7, pp. 80937-80949, 2019.

[76] Z. An, S. Li, Y. Xin, K. Xu and H. Ma, "An intelligent fault diagnosis framework dealing with arbitrary length inputs under different working conditions," *Meas. Sci. Technol.*, vol. 30, no. 12, Dec. 2019.

[77] Q. Li, B. Tang, L. Deng, Y. Wu and Y. Wang, "Deep balanced domain adaptation neural networks for fault diagnosis of planetary gearboxes with limited labeled data," *Measurement*, vol. 156, May 2020.

[78] M. Azamfar, X. Li and J. Lee, "Intelligent ball screw fault diagnosis using a deep domain adaptation methodology," *Mechanism Mach. Theory*, vol. 151, Sep. 2020.

[79] M. Azamfar, X. Li and J. Lee, "Deep Learning-Based Domain Adaptation Method for Fault Diagnosis in Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 3, pp. 445-453, Aug. 2020.

[80] Y. Zhang, Z. Ren, S. Zhou, "A New Deep Convolutional Domain Adaptation Network for Bearing Fault Diagnosis under Different Working Conditions," *Shock and Vibration*, vol. 2020, Article ID 8850976.

[81] J. Zhu, N. Chen and C. Shen, "A New Deep Transfer Learning Method for Bearing Fault Diagnosis Under Different Working Conditions," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8394-8402, 1 Aug.1, 2020.

[82] J. Singh, M. Azamfar, A. Ainapure and J. Lee, "Deep learning-based cross-domain adaptation for gearbox fault diagnosis under variable speed conditions," *Meas. Sci. Technol.*, vol. 31, no. 5, Feb. 2020.

[83] Y. Li, Y. Song, L. Jia, S. Gao, Q. Li and M. Qiu, "Intelligent Fault Diagnosis by Fusing Domain Adversarial Training and Maximum Mean Discrepancy via Ensemble Learning," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2833-2841, April 2021.

[84] T. Han, C. Liu, W. Yang and D. Jiang, "Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application," *ISA Trans.*, vol. 97, pp. 269-281, Feb. 2020.

[85] Han T, Li Y F, Qian M. A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions[J]. IEEE Transactions on Instrumentation and Measurement, 2021.

[86] Wei D, Han T, Chu F, et al. Weighted domain adaptation networks for machinery fault diagnosis[J]. Mechanical Systems and Signal Processing, 2021, 158: 107744.

[87] X. Wang, C. Shen, M. Xia, D. Wang, J. Zhu and Z. Zhu, "Multi-scale deep intra-class transfer learning for bearing fault diagnosis," *Rel. Eng. Syst. Saf.*, vol. 202, Oct. 2020.

[88] C. Shen, X. Wang, D. Wang, Y. Li, J. Zhu and M. Gong, "Dynamic Joint Distribution Alignment Network for Bearing Fault Diagnosis Under Variable Working Conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-13, 2021, Art no. 3510813.

[89] X. Li, W. Zhang, Q. Ding and J.-Q. Sun, "Multi-layer domain adaptation method for rolling bearing fault diagnosis," *Signal Process.*, vol. 157, pp. 180-197, 2019.

[90] J. Wu, T. Tang, M. Chen, Y. Wang, and K. Wang, “A study on adaptation lightweight architecture based deep learning models for bearing fault diagnosis under varying working conditions,” *Expert Syst. Appl.*, vol. 160, Dec. 2020, Art. no. 113710.

[91] B. Yang, Q. Li, L. Chen, C. Shen, "Bearing Fault Diagnosis Based on Multilayer Domain Adaptation," *Shock and Vibration*, vol. 2020, Article ID 8873960.

[92] W. Qian, S. Li and J. Wang, "A New Transfer Learning Method and its Application on Rotating Machine Fault Diagnosis Under Variant Working Conditions," *IEEE Access*, vol. 6, pp. 69907-69917, 2018.

[93] W. Qian, S. Li and X. Jiang, "Deep transfer network for rotating machine fault analysis," *Pattern Recognit.*, vol. 96, Dec. 2019.

[94] X. Wang, H. He and L. Li, "A Hierarchical Deep Domain Adaptation Approach for Fault Diagnosis of Power Plant Thermal System," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5139-5148, Sept. 2019.

[95] J. An, P. Ai, D. Liu, "Deep Domain Adaptation Model for Bearing Fault Diagnosis with Domain Alignment and Discriminative Feature Learning," *Shock and Vibration*, vol. 2020, Article ID 4676701.

[96] P. Xiong, B. Tang, L. Deng, M. Zhao, X. Yu, "Multi-block domain adaptation with central moment discrepancy for fault diagnosis," *Measurement*, vol. 169, Feb. 2021.

[97] X. Li, Y. Hu, J. Zheng, M. Li, W. Ma, "Central moment discrepancy based domain adaptation for intelligent bearing fault diagnosis," *Neurocomputing*, vol. 429, pp. 12-24, Mar. 2020.

[98] H. Bao, Z. Yan, S. Ji, J. Wang, S. Jia, G. Zhang and B. Han, "An Enhanced Sparse Filtering Method for Transfer Fault Diagnosis Using Maximum Classifier Discrepancy", *Meas. Sci. Technol.*, 2021, DOI: https://doi.org/10.1088/1361-6501/abe56f.

[99] K. Xu, S. Li, X. Jiang, J. Lu, T. Yu, R. Li, "A novel transfer diagnosis method under unbalanced sample based on discrete-peak joint attention enhancement mechanism," *Knowl.-Based Syst.*, vol. 212, Jan. 2021.

[100] Z. Huang et al., "A multi-source dense adaptation adversarial network for fault diagnosis of machinery," in IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2021.3086707.

[101] X. Li, W. Zhang and Q. Ding, "Cross-Domain Fault Diagnosis of Rolling Element Bearings Using Deep Generative Neural Networks," *IEEE Trans. Ind. Electron.*, vol. 66, no. 7, pp. 5525-5534, July 2019.

[102] T. Zheng, L. Song, J. Wang, W. Teng, X. Xu, C. Ma, " Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings," *Measurement*, vol. 158, July. 2020.

[103] P. Liang, C. Deng, J. Wu and Z. Yang, "Intelligent fault diagnosis of rotating machinery via wavelet transform generative adversarial nets and convolutional neural network," *Measurement*, vol. 159, Jul. 2020.

[104] P. Liang, C. Deng, J. Wu, G. Li, Z. Yang, and Y. Wang, "Intelligent fault diagnosis via semisupervised generative adversarial nets and wavelet transform," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4659–4671, Jul. 2020.

[105] H. Tao, P. Wang, Y. Chen, V. Stojanovic, and H. Yang, "An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks," *J. Franklin Inst.*, vol. 357, no. 11, pp. 7286–7307, Jul. 2020.

[106] S. Shao, P. Wang, and R. Yan, "Generative adversarial networks for data augmentation in machine fault diagnosis," *Comput. Ind.*, vol. 106, pp. 85–93, Apr. 2019.

[107] Q. Guo, Y. Li, Y. Song, D. Wang, and W. Chen, "Intelligent fault diagnosis method based on full 1-D convolutional generative adversarial network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2044–2053, Mar. 2020.

[108] Z. Shi, J. Chen, Y. Zi and Z. Zhou, "A Novel Multitask Adversarial Network via Redundant Lifting for Multicomponent Intelligent Fault Detection Under Sharp Speed Variation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-10, 2021, Art no. 3511010.

[109] T. Han, C. Liu, W. Yang and D. Jiang, "A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults,"*Knowl.-Based Syst.*, vol. 165, pp. 474-487, Feb. 2019.

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

[111] J. Shao, Z. Huang and J. Zhu, "Transfer Learning Method Based on Adversarial Domain Adaption for Bearing Fault Diagnosis," *IEEE Access*, vol. 8, pp. 119421-119430, 2020.

[112] Z. Chai and C. Zhao, "A Fine-Grained Adversarial Network Method for Cross-Domain Industrial Fault Diagnosis," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 3, pp. 1432-1442, July 2020.

[113] Z. Chen, G. He, J. Li, Y. Liao, K. Gryllias and W. Li, "Domain Adversarial Transfer Network for Cross-Domain Fault Diagnosis of Rotary Machinery," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 11, pp. 8702-8712, Nov. 2020.

[114] F. Li, T. Tang, B. Tang and Q. He, "Deep convolution domain-adversarial transfer learning for fault diagnosis of rolling bearings," *Measurement*, vol. 169, Feb. 2021.

[115] W. Mao, Y. Liu, L. Ding, A. Safian and X. Liang, "A New Structured Domain Adversarial Neural Network for Transfer Fault Diagnosis of Rolling Bearings Under Different Working Conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-13, 2021, Art no. 3509013

[116] Z. Liu, B. Lu, H. Wei, L. Chen, X. Li and M. Rätsch, "Deep Adversarial Domain Adaptation Model for Bearing Fault Diagnosis," *IEEE Trans. Syst. Man Cybern. Syst.*, doi: 10.1109/TSMC.2019.2932000.

[117] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Deep learning-based adversarial multi-classifier optimization for cross-domain machinery fault diagnostics," *Journal of Manufacturing Systems*, vol. 55, pp. 334– 347, 2020.

[118] X. Li, W. Zhang, H. Ma, Z. Luo and X. Li, "Domain generalization in rotating machinery fault diagnostics using deep neural networks," *Neurocomputing*, vol. 403, pp. 409-420, 2020.

[119] Y. Zhang, Z. Ren, S. Zhou and T. Yu, "Adversarial domain adaptation with classifier alignment for cross-domain intelligent fault diagnosis of multiple source domains," *Meas. Sci. Technol.*, vol. 32, no. 3, 2021.

[120] Q. Li, C. Shen, L. Chen and Z. Zhu, "Knowledge mapping-based adversarial domain adaptation: A novel fault diagnosis method with high generalizability under variable working conditions," *Mech. Syst. Signal Process.*, vol. 147, Jan. 2021.

[121] M. Zhang, D. Wang, W. Lu, J. Yang, Z. Li and B. Liang, "A Deep Transfer Model With Wasserstein Distance Guided Multi-Adversarial Networks for Bearing Fault Diagnosis Under Different Working Conditions," *IEEE Access*, vol. 7, pp. 65303-65318, 2019.

[122] J. Jiao, J. Lin, M. Zhao and K. Liang, "Double-level adversarial domain adaptation network for intelligent fault diagnosis," *Knowl.-Based Syst.*, vol. 205, Oct. 2020.

[123] J. Jiao, M. Zhao and J. Lin, "Unsupervised Adversarial Adaptation Network for Intelligent Fault Diagnosis," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9904-9913, Nov. 2020.

[124] C. Cheng, B. Zhou, G. Ma, D. Wu and Y. Yuan, "Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data," *Neurocomputing*, vol. 409, pp. 35-45, Oct. 2020.

[125] X. Wang and F. Liu, "Triplet loss guided adversarial domain adaptation for bearing fault diagnosis," *Sensors*, vol. 20, no. 1, pp. 1-11, Apr. 2020.

[126] Y Zou, Y Liu, J Deng, Y Jiang, W Zhang, "A novel transfer learning method for bearing fault diagnosis under different working conditions," *Measurement*, vol. 171, Feb. 2021.

[127] B Han, X Zhang, J Wang, Z An, S Jia, G Zhang, "Hybrid distance-guided adversarial network for intelligent fault diagnosis under different working conditions," *Measurement*, vol. 176, May 2021.

[128] Y. Wang, X. Sun, J. Li and Y. Yang, "Intelligent Fault Diagnosis With Deep Adversarial Domain Adaptation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-9, 2021, Art no. 2503509.

[129] X. Yu et al., "Conditional Adversarial Domain Adaptation With Discrimination Embedding for Locomotive Fault Diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-12, 2021, Art no. 3503812.

[130] D. She, M. Jia and M. Pecht, "Weighted Entropy Minimization Based Deep Conditional Adversarial Diagnosis Approach under Variable Working Conditions," *IEEE/ASME Trans. Mechatronics*, doi: 10.1109/TMECH.2020.3040175.

[131] M. Ragab et al., "Adversarial Multiple-Target Domain Adaptation for Fault Classification," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021, Art no. 3500211.

[132] Y. Liao, R. Huang, J. Li, Z. Chen and W. Li, "Deep Semisupervised Domain Generalization Network for Rotary Machinery Fault Diagnosis Under Variable Speed," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 8064-8075, Oct. 2020.

[133] S. Pang and X. Yang, "A Cross-Domain Stacked Denoising Autoencoders for Rotating Machinery Fault Diagnosis Under Different Working Conditions," *IEEE Access*, vol. 7, pp. 77277-77292, 2019.

[134] Z. Liu, B. Lu, H. Wei, X. Li and L. Chen, "Fault Diagnosis for Electromechanical Drivetrains Using a Joint Distribution Optimal Deep Domain Adaptation Approach," *IEEE Sensors Journal*, vol. 19, no. 24, pp. 12261-12270, 15 Dec.15, 2019.

[135] Z. Liu, L. Jiang, H. Wei, L. Chen and X. Li, "Optimal Transport-Based Deep Domain Adaptation Approach for Fault Diagnosis of Rotating Machine," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-12, 2021, Art no. 3508912.

[136] L. Wen, L. Gao and X. Li, "A New Deep Transfer Learning Based on Sparse Auto-Encoder for Fault Diagnosis," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 49, no. 1, pp. 136-144, Jan. 2019.

[137] Z. Wan, R. Yang, M. Huang, "Deep Transfer Learning-Based Fault Diagnosis for Gearbox under Complex Working Conditions," *Shock and Vibration*, vol. 2020, Article ID 8884179.

[138] Z. Tang, L. Bo, X. Liu, D. Wei, "An autoencoder with adaptive transfer learning for intelligent fault diagnosis of rotating machinery," *Meas. Sci. Technol.*, 2021, DOI: https://doi.org/10.1088/1361-6501/abd650.

[139] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, May 2019.

[140] Y. Gao, X. Liu, H. Huang, J. Xiang. "A hybrid of FEM simulations and generative adversarial networks to classify faults in rotor-bearing systems," *ISA Trans.*, vol. 108, pp. 356–366, Feb. 2021.

[141] K. Yu, Q. Fu, H. Ma, TR. Lin, X. Li, "Simulation data driven weakly supervised adversarial domain adaptation approach for intelligent cross-machine fault diagnosis," *Structural Health Monitoring*, Jan. 2021. doi:10.1177/1475921720980718

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

[143] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, "Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2416–2425, Apr. 2019.

[144] H. Zheng, Y. Yang, J. Yin, Y. Li, R. Wang and M. Xu, "Deep Domain Generalization Combining A Priori Diagnosis Knowledge Toward Cross-Domain Fault Diagnosis of Rolling Bearing," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021, Art no. 3501311.

[145] B. Yang, Y. Lei, S. Xu and C. Lee, "An Optimal Transport-embedded Similarity Measure for Diagnostic Knowledge Transferability Analytics across Machines," IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2021.3095804.

[146] Z. Wu, H. Jiang, K. Zhao and X. Li, "An adaptive deep transfer learning method for bearing fault diagnosis," *Measurement*, vol. 151, Feb. 2020.

[147] P. Wang and R. X. Gao, "Transfer learning for enhanced machine fault diagnosis in manufacturing," *CIRP Ann.*, vol. 69, no. 1, pp. 413-416, 2020.

[148] H. Shao, Z. Ding, J. Cheng, and H. Jiang, "Intelligent fault diagnosis among different rotating machines using novel stacked transfer auto-encoder optimized by PSO," *ISA Trans.*, vol. 105, pp. 308–319, Oct. 2020.

[149] Xudong Li et al., "Fault diagnostics between different type of components: A transfer learning approach," *Applied Soft Computing*, vol. 86, pp. 105950, 2020.

[150] H. Zhiyi, S. Haidong, J. Lin, C. Junsheng and Y. Yu, "Transfer fault diagnosis of bearing installed in different machines using enhanced deep auto-encoder," *Measurement*, vol. 152, Feb. 2020.

[151] Z. Chen, K. Gryllias and W. Li, "Intelligent Fault Diagnosis for Rotary Machinery Using Transferable Convolutional Neural Network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 339-349, Jan. 2020.

[152] L. Guo, Y. Lei, S. Xing, T. Yan and N. Li, "Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines With Unlabeled Data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316-7325, Sept. 2019.

[153] B. Yang, Y. Lei, F. Jia, N. Li and Z. Du, "A Polynomial Kernel Induced Distance Metric to Improve Deep Transfer Learning for Fault Diagnosis of Machines," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9747-9757, Nov. 2020.

[154] X. Li, X.-D. Jia, W. Zhang, H. Ma, Z. Luo and X. Li, "Intelligent cross-machine fault diagnosis approach with deep auto-encoder and domain adaptation," *Neurocomputing*, vol. 383, pp. 235-247, Mar. 2020.

[155] Z. Wu, H. Jiang, T. Lu, K. Zhao, "A deep transfer maximum classifier discrepancy method for rolling bearing fault diagnosis under few labeled data," *Knowl.-Based Syst.*, vol. 196, May 2020.

[156] C. Zheng, X. Wang, Y. Hao, K. Wang and X. Xiong, "Normalized Recurrent Dynamic Adaption Network: A New Framework With Dynamic Alignment for Intelligent Fault Diagnosis," *IEEE Access*, vol. 8, pp. 80243-80255, 2020.

[157] M. Lv, S. Liu, X. Su and C. Chen, "Deep Transfer Network With Multi-Kernel Dynamic Distribution Adaptation for Cross-Machine Fault Diagnosis," *IEEE Access*, vol. 9, pp. 16392-16409, 2021.

[158] K. Zhao, H. Jiang, Z. Wu, T. Lu, "A novel transfer learning fault diagnosis method based on Manifold Embedded Distribution Alignment with a little labeled data," *Journal of Intelligent Manufacturing*, Sept. 2020, DOI: https://doi.org/10.1007/s10845-020-01657-z

[159] X. Li, H. Jiang, R. Wang, M. Niu, "Rolling bearing fault diagnosis using optimal ensemble deep transfer network," *Knowl.-Based Syst.*, vol. 213, Feb. 2021.

[160] Y. Tan, L. Guo, H. Gao and L. Zhang, "Deep Coupled Joint Distribution Adaptation Network: A Method for Intelligent Fault Diagnosis Between Artificial and Real Damages," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-12, 2021, Art no. 3507212.

[161] J. Chen, J. Wang, J. Zhu, T. H. Lee and C. De Silva, "Unsupervised Cross-domain Fault Diagnosis Using Feature Representation Alignment Networks for Rotating Machinery," *IEEE/ASME Trans. Mechatronics*, doi: 10.1109/TMECH.2020.3046277.

[162] Zhang B, Hong X, Liu Y. Distribution adaptation deep transfer learning method for cross-structure health monitoring using guided waves[J]. Structural Health Monitoring, 2021, DOI: https://doi.org/10.1177/14759217211010709.

[163] Zhang B, Hong X, Liu Y. Multi-task deep transfer learning method for guided wave-based integrated health monitoring using piezoelectric transducers[J]. IEEE Sensors Journal, 2020, 20(23): 14391-14400.

[164] Zhang B, Hong X, Liu Y, Deep Convolutional Neural Network Probability Imaging for Plate Structural Health Monitoring Using Guided Waves[J]. IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-10, 2021, Art no. 2510610.

[165] X. Wang, F. Liu, D. Zhao, "Cross-Machine Fault Diagnosis with Semi-Supervised Discriminative Adversarial Domain Adaptation," *Sensors*, vol. 20, no. 3, pp. 1-16, July 2020.

[166] Y. Feng, J. Chen, Z. Yang, X. Song, Y. Chang, S. He, E. Xu, Z. Zhou, "Similarity-based meta-learning network with adversarial domain adaptation for cross-domain fault identification," *Knowl.-Based Syst.*, vol. 217, Apr. 2021.

[167] J. Zhu, N. Chen and C. Shen, "A New Multiple Source Domain Adaptation Fault Diagnosis Method Between Different Rotating Machines," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2020.3021406.

[168] Liao Y, Huang R, Li J, et al. Dynamic Distribution Adaptation Based Transfer Network for Cross Domain Bearing Fault Diagnosis[J]. Chinese Journal of Mechanical Engineering, 2021, 34(1): 1-10.

[169] N. Lu, T. Yin, "Transferable common feature space mining for fault diagnosis with imbalanced data," *Mech. Syst. Signal Process.*, Vol. 156, July 2021.

[170] D. Xiao, Y. Huang, C. Qin, Z. Liu, Y. Li and C. Liu, "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.*, vol. 233, no. 14, pp. 5131-5143, Jul. 2019.

[171] X. Li, H. Jiang, K. Zhao and R. Wang, "A Deep Transfer Nonnegativity-Constraint Sparse Autoencoder for Rolling Bearing Fault Diagnosis With Few Labeled Data," *IEEE Access*, vol. 7, pp. 91216-91224, 2019.

[172] H. Kim and B. D. Youn, "A New Parameter Repurposing Method for Parameter Transfer With Small Dataset and Its Application in Fault Diagnosis of Rolling Element Bearings," *IEEE Access*, vol. 7, pp. 46917-46930, 2019.

[173] Z. He, H. Shao, P. Wang, J. Lin, J. Cheng and Y. Yang, "Deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearbox with few target training samples," *Knowl.-Based Syst.*, vol. 191, Mar. 2020.

[174] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," *Measurement*, vol. 152, Feb. 2020.

[175] L. Zou, Y. Li and F. Xu, "An adversarial denoising convolutional neural network for fault diagnosis of rotating machinery under noisy environment and limited sample size case," *Neurocomputing*, vol. 407, pp. 105-120, Sep. 2020.

[176] Y. Zhang, Z. Ren and S. Zhou, "An Intelligent Fault Diagnosis for Rolling Bearing Based on Adversarial Semi-Supervised Method," *IEEE Access*, vol. 8, pp. 149868-149877, 2020.

[177] X. Li, W. Zhang, Q. Ding and X. Li, "Diagnosing Rotating Machines With Weakly Supervised Data Using Deep Transfer Learning," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1688-1697, March 2020.

[178] M Zareapoor, P Shamsolmoali, J Yang, "Oversampling adversarial network for class-imbalanced fault diagnosis," *Mech. Syst. Signal Process.*, Vol. 149, Feb. 2021.

[179] T. Zhang, J. Chen, F. Li, T. Pan and S. He, "A Small Sample Focused Intelligent Fault Diagnosis Scheme of Machines via Multi-modules Learning with Gradient Penalized Generative Adversarial Networks," *IEEE Trans. Ind. Electron.*, doi: 10.1109/TIE.2020.3028821.

[180] C. Li, D. Cabrera, F. Sancho, R. -V. Sanchez, M. Cerrada and J. Valente de Oliveira, "One-shot fault diagnosis of 3D printers through improved feature space learning," *IEEE Trans. Ind. Electron.*, doi: 10.1109/TIE.2020.3013546.

[181] X. Li, W. Zhang, N. Xu and Q. Ding, "Deep Learning-Based Machinery Fault Diagnostics With Domain Adaptation Across Sensors at Different Places," *IEEE Trans. Ind. Electron.*, vol. 67, no. 8, pp. 6785-6794, Aug. 2020.

[182] S Siahpour, X Li, J Lee, "Deep learning-based cross-sensor domain adaptation for fault diagnosis of electro-mechanical actuators," Int. J. Dynam. Control, vol. 8, pp. 1054–1062, July 2020.

[183] V. Pandhare, X. Li, M. Miller, X. Jia and J. Lee, "Intelligent Diagnostics for Ball Screw Fault Through Indirect Sensing Using Deep Domain Adaptation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021, Art no. 2504211.

[184] Y. Qin, X. Wang, Q. Qian, H. Pu and J. Luo, "Multi-scale transfer voting mechanism: a new strategy for domain adaption," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2020.3045392.

[185] S. Wu, X. Jing, Q. Zhang, F. Wu, H. Zhao and Y. Dong, "Prediction Consistency Guided Convolutional Neural Networks for Cross-Domain Bearing Fault Diagnosis," *IEEE Access*, vol. 8, pp. 120089-120103, 2020

[186] Z. Cao, M. Long, J. Wang, and M. I. Jordan, "Partial transfer learning with selective adversarial networks," *in Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2724–2732.

[187] J. Jiao, M. Zhao, J. Lin, and C. Ding, "Classifier inconsistency-based domain adaptation network for partial transfer intelligent diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5965–5974, Sep. 2020.

[188] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Partial transfer learning in machinery cross-domain fault diagnostics using class-weighted adversarial networks," *Neural Netw.*, vol. 129, pp. 313–322, Sep. 2020.

[189] W. Li, Z. Chen and G. He, "A Novel Weighted Adversarial Transfer Network for Partial Domain Fault Diagnosis of Machinery," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1753-1762, March 2021.

[190] Z. Liu, B. Lu, H. Wei, L. Chen, X. Li and C. Wang, "A Stacked Auto-Encoder Based Partial Adversarial Domain Adaptation Model for Intelligent Fault Diagnosis of Rotating Machines," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2020.3045002.

[191] X. Li and W. Zhang, "Deep Learning-Based Partial Domain Adaptation Method on Intelligent Machinery Fault Diagnostics," *IEEE Trans. Ind. Electron.*, vol. 68, no. 5, pp. 4351-4361, May 2021.

[192] T. Han, C. Liu, R. Wu, D. Jiang, "Deep transfer learning with limited data for machinery fault diagnosis," *Applied Soft Computing*, vol. 103, pp. 107150, 2020.

[193] Y. Deng, D. Huang, S. Du, G. Li, C. Zhao, J. Lv, "A double-layer attention based adversarial network for partial transfer learning in machinery fault diagnosis," *Comput. Ind.*, Vol. 127, May 2021, 103399.

[194] B. Yang, CG. Lee, Y. Lei, N. Li, N. Lu, "Deep partial transfer learning network: A method to selectively transfer diagnostic knowledge across related machines," *Mech. Syst. Signal Process.*, vol. 116, July 2021.

[195] Q. Wang, G. Michau and O. Fink, "Missing-Class-Robust Domain Adaptation by Unilateral Alignment," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 663-671, Jan. 2021.

[196] S. Zhang, M. Wang, W. Li, J. Luo, and Z. Lin, "Deep learning with emerging new labels for fault diagnosis," *IEEE Access*, vol. 7, pp. 6279–6287, 2019.

[197] C. Wang, C. Xin, Z. Xu, "A novel deep metric learning model for imbalanced fault diagnosis and toward open-set classification," *Knowl.-Based Syst.*, vol. 220, May 2021.

[198] L. Feng and C. Zhao, "Fault Description Based Attribute Transfer for Zero-Sample Industrial Fault Diagnosis," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1852-1862, March 2021.

[199] J. Li, R. Huang, W. Li, "Intelligent Fault Diagnosis for Bearing Dataset Using Adversarial Transfer Learning based on Stacked Auto-Encoder," *Procedia Manufacturing*, Vol. 49, pp. 75-80, 2020.

[200] J. Li, R. Huang, G. He, S. Wang, G. Li and W. Li, "A Deep Adversarial Transfer Learning Network for Machinery Emerging Fault Detection," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8413-8422, 1 Aug.1, 2020.

[201] H. Wang, X. Bai, J. Tan, J. Yang, "Deep prototypical networks based domain adaptation for fault diagnosis," *Journal of Intelligent Manufacturing*, 2020, https://doi.org/10.1007/s10845-020-01709-4

[202] W. Zhang, X. Li, H. Ma, Z. Luo and X. Li, "Open Set Domain Adaptation In Machinery Fault Diagnostics Using Instance-Level Weighted Adversarial Learning," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2021.3054651.

[203] W. Zhang, X. Li, H. Ma, Z. Luo and X. Li, "Universal Domain Adaptation in Fault Diagnostics with Hybrid Weighted Deep Adversarial Learning," IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3064377.

[204] Yang B, Xu S, Lei Y, et al. Multi-source transfer learning network to complement knowledge for intelligent diagnosis of machines with unseen faults[J]. Mechanical Systems and Signal Processing, 2022, 162: 108095.

[205] J. Li, R. Huang, G. He, Y. Liao, Z. Wang, and W. Li, "A two-stage transfer adversarial network for intelligent fault diagnosis of rotating machinery with multiple new faults," *IEEE/ASME Trans. Mechatronics*, early access, Sep. 22, 2020, doi: 10.1109/TMECH.2020.3025615.

[206] X. Yu et al., "Deep Learning-Based Open Set Fault Diagnosis by Extreme Value Theory," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2021.3070324.

[207] P. P. Busto and J. Gall, “Open set domain adaptation,” *in Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 754–763.

[208] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, "Open set domain adaptation by backpropagation," *in Proc. ECCV*, 2018, pp. 153–168.

[209] D. Zhao, W. Cheng, R. X. Gao, R. Yan, and P. Wang, “Generalized Vold– Kalman filtering for nonstationary compound faults feature extraction of bearing and gear,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 401–410, Feb. 2020.

[210] G. Tang, Y. Wang, Y. Huang, N. Liu, and J. He, "Compound bearing fault detection under varying speed conditions with virtual multichannel signals in angle domain," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 8, pp. 5535–5545, Aug. 2020.

[211] N. Li, W. Huang, W. Guo, G. Gao, and Z. Zhu, "Multiple enhanced sparse decomposition for gearbox compound fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 770–781, Mar. 2020.

[212] L. Cui, Y. Sun, J. Zhang, H. Wang, "Adapted dictionary-free orthogonal matching pursuit and 0-1 programming to solve the isolation and diagnosis of bearing and gear compound faults," *Measurement*, Mar. 2021. Art. no. 109331.

[213] Shao H, Jiang H, Zhang H, et al. Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing[J]. Mechanical Systems and Signal Processing, 2018, 100: 743-765.

[214] Wang S, Xiang J, Zhong Y, et al. A data indicator-based deep belief networks to detect multiple faults in axial piston pumps[J]. Mechanical Systems and Signal Processing, 2018, 112: 154-170.

[215] Zhao X, Jia M, Liu Z. Semisupervised Graph Convolution Deep Belief Network for Fault Diagnosis of Electormechanical System With Limited Labeled Data[J]. IEEE Transactions on Industrial Informatics, 2020, 17(8): 5450-5460.

[216] Shao H, Lin J, Zhang L, et al. A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance[J]. Information Fusion, 2021, 74: 65-76.

[217] Sun G, Wang Y, Sun C, et al. Intelligent detection of a planetary gearbox composite fault based on adaptive separation and deep learning[J]. Sensors, 2019, 19(23): 5222.

[218] Lin M C, Han P Y, Fan Y H, et al. Development of Compound Fault Diagnosis System for Gearbox Based on Convolutional Neural Network[J]. Sensors, 2020, 20(21): 6169.

[219] Zhang D, Ren X, Zuo H. Compound Fault Diagnosis for Gearbox Based Using of Euclidean Matrix Sample Entropy and One-Dimensional Convolutional Neural Network[J]. Shock and Vibration, 2021, 2021.

[220] Xue Y, Dou D, Yang J. Multi-fault diagnosis of rotating machinery based on deep convolution neural network and support vector machine[J]. Measurement, 2020, 156: 107571.

[221] P. Liang, C. Deng, J. Wu, Z. Yang, J. Zhu, and Z. Zhang, "Single and simultaneous fault diagnosis of gearbox via a semi-supervised and high-accuracy adversarial learning framework," *Knowl.-Based Syst.*, vol. 198, Jun. 2020, Art. no. 105895.

[222] R. Huang, W. Li, and L. Cui, "An intelligent compound fault diagnosis method using one-dimensional deep convolutional neural network with multi-label classifier," in Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC), Auckland, New Zealand, May 2019, pp. 1–6.

[223] P. Liang, C. Deng, J. Wu, Z. Yang, J. Zhu, and Z. Zhang, ''Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform,'' *Comput. Ind.*, vol. 113, Dec. 2019, Art. no. 103132.

[224] Y. Jin, C. Qin, Y. Huang, and C. Liu, "Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network," *Measurement*, vol. 173, Mar. 2021. Art. no. 108500.

[225] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," 2017, arXiv:1710.09829. [Online]. Available: http://arxiv.org/abs/1710.09829

[226] R. Huang, Y. Liao, S. Zhang, and W. Li, "Deep decoupling convolutional neural network for intelligent compound fault diagnosis," *IEEE Access*, vol. 7, pp. 1848–1858, 2019.

[227] R. Huang, J. Li, W. Li, and L. Cui, "Deep ensemble capsule network for intelligent compound fault diagnosis using multisensory data," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2304–2314, May 2020.

[228] A. Dibaj, MM. Ettefagh, R. Hassannejad, M. Ehghaghi, "A hybrid fine-tuned VMD and CNN scheme for untrained compound fault diagnosis of rotating machinery with unequal-severity faults," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114094.

[229] Xing S, Lei Y, Wang S, et al. A label description space embedded model for zero-shot intelligent diagnosis of mechanical compound faults[J]. Mechanical Systems and Signal Processing, 2022, 162: 108036.

[230] R. Huang, Z. Wang, J. Li, J. Chen, and W. Li, "A transferable capsule network for decoupling compound fault of machinery," *in Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, Dubrovnik, Croatia, May 2020, pp. 1–6.

[231] R. Huang, J. Li, Y. Liao, J. Chen, Z. Wang and W. Li, "Deep Adversarial Capsule Network for Compound Fault Diagnosis of Machinery Toward Multidomain Generalization Task," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021, Art no. 3506311.

[232] Z. Chen, A. Mauricio, W. Li, and K. Gryllias, ''A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks,'' *Mech. Syst. Signal Process.*, vol. 140, Jun. 2020, Art. no. 106683.

[233] D. Zhu, X. Cheng, L. Yang, Y. Chen and S. X. Yang, "Information Fusion Fault Diagnosis Method for Deep-Sea Human Occupied Vehicle Thruster Based on Deep Belief Network," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2021.3055770.

[234] L. Bottou, "Large-scale machine learning with stochastic gradient descent," *in Proc. 19th Int. Conf. Comput. Statist.*, 2010, pp. 177–186.

[235] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.

[236] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *in Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–41.

[237] X. Yuan, P. He, Q. Zhu and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805-2824, Sept. 2019.

[238] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725–726, 2018.

[239] Z. C. Lipton, "The mythos of model interpretability," *in Proc. Int. Conf. Mach. Learn. (ICML) Workshop*, 2016, pp. 1–9.

[240] A. Rasheed, O. San and T. Kvamsdal, "Digital Twin: Values, Challenges and Enablers From a Modeling Perspective," *IEEE Access*, vol. 8, pp. 21980-22012, 2020.

[241] C. Molnar. (2019). Interpretable Machine Learning. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[242] Zhou Z, Li T, Zhang Z, et al. Bayesian Differentiable Architecture Search for Efficient Domain Matching Fault Diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-11.

[243] Qiang Yang; Yang Liu; Yong Cheng; Yan Kang; Tianjian Chen; Han Yu, Federated Learning, Morgan & Claypool, 2019, doi: 10.2200/S00960ED2V01Y201910AIM043.

[244] Z. Mao, M. Xia, B. Jiang, D. Xu and P. Shi, "Incipient Fault Diagnosis for High-Speed Train Traction Systems via Stacked Generalization," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2020.3034929.

[245] T. T. Nguyen, N. D. Nguyen and S. Nahavandi, "Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3826-3839, Sept. 2020.

**Weihua Li:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.
**Ruyi Huang:** Conceptualization, Methodology, Software, Writing - Original Draft.
**Jipu Li:** Validation, Investigation, Resources.
**Yixiao Liao:** Validation, Investigation, Resources.
**Zhuyun Chen:** Validation, Investigation, Resources.
**Guolin He:** Visualization, Resources, Writing - Review & Editing.
**Ruqiang Yan:** Writing - Review & Editing, Supervision.
**Konstantinos Gryllias:** Writing - Review & Editing, Supervision.

**Declaration of interests**

☐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.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Weihua Li reports financial support was provided by National Key Research and Development Program of China. Weihua Li reports financial support was provided by National Natural Science Foundation of China. Weihua Li reports financial support was provided by Key-Area Research and Development Program of Guangdong Province.