

REVIEW

Knowledge transfer in fault diagnosis of rotary machines

Guokai Liu¹  | Weiming Shen¹  | Liang Gao¹  | Andrew Kusiak² 

¹State Key Laboratory of Digital Manufacturing Equipment and Technology, School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China

²Department of Industrial and System Engineering, Seamans Center, The University of Iowa, Iowa City, Iowa, USA

Correspondence

Weiming Shen, State Key Laboratory of Digital Manufacturing Equipment and Technology, School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China.
Email: wshen@iecc.org

Funding information

China Scholarship Council, Grant/Award Number: 201906160078; the Fundamental Research Funds for the Central Universities of China, Grant/Award Number: 2021GCRC058

Abstract

Data-driven fault diagnosis has prevailed in machine condition monitoring in the past decades. However, traditional machine- and deep-learning-based fault diagnosis methods assumed that the source and target data share the same distribution and ignored knowledge transfer in dynamic working environments. In recent years, knowledge transfer approaches have been developed and have shown promising results in intelligent fault diagnosis and health management of rotary machines. This paper presents a comprehensive review of knowledge transfer approaches and their applications in fault diagnosis of rotary machines. A problem-oriented taxonomy of knowledge transfer in fault diagnosis is proposed. The knowledge transfer paradigms, approaches, and applications are categorised and analysed. Future research challenges and directions are explored from data, modelling, and application perspectives.

1 | INTRODUCTION

The concept of knowledge transfer, including transfer learning and domain adaptation, is of great interest to fault diagnosis of rotary machinery. Knowledge learnt in one application is transferred to another. Knowledge transfer is a key to prognostics and health management of rotary machines. Monitoring the health conditions of machines reduces the risk of downtime, expensive repairs, and catastrophic failures.

In the past three decades, research on rotary machine fault diagnosis (RMFD) has evolved over four stages (See Figure 1): knowledge-based [1–4], machine learning [5–7], deep learning [8–10], and knowledge transfer [11–13].

At the first stage, the expert experience was the basis of RMFD. Physics-derived parameters and the record of past events were used to conduct knowledge-based systems (KBSs). The symptom of the monitored machines was fed to a KBS making maintenance decisions. The growing complexity of rotation machinery has made it more difficult to build dependable KBSs.

Machine learning emerged as a development of KBSs. Based on the historical maintenance data and expert knowledge, machine-learning algorithms such as Bayesian network

(BN) [14], random forest [15], and support vector machine (SVM) [16] were developed with feature engineering [17–20] to build models for fault diagnosis. The growing number of parameters and data, quite often from heterogeneous sources, has led to the big data challenge [21].

In 2012, AlexNet [22] won the championship of the ImageNet Large Scale Visual Recognition Challenge. The amazing performance of the AlexNet has generated interest in deep learning-based algorithms such as convolutional neural networks (CNNs) [23, 24], recurrent neural networks [25, 26], long short-term memory networks [27, 28], and generative adversarial networks (GANs) [29–31]. Compared with conventional machine learning, deep learning algorithms are equipped with feature extraction abilities. The end-to-end paradigm has been widely used and proven to be effective in intelligent fault diagnosis. However, the assumption that training and future data must be in the same feature space and have the same distribution does not hold in many industrial applications.

Knowledge transfer has attracted attention in recent years. Deep learning algorithms suffer from time-consuming training, especially from large-scale unlabelled data. Difficulties in knowledge reuse arise in the presence of different

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *IET Collaborative Intelligent Manufacturing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

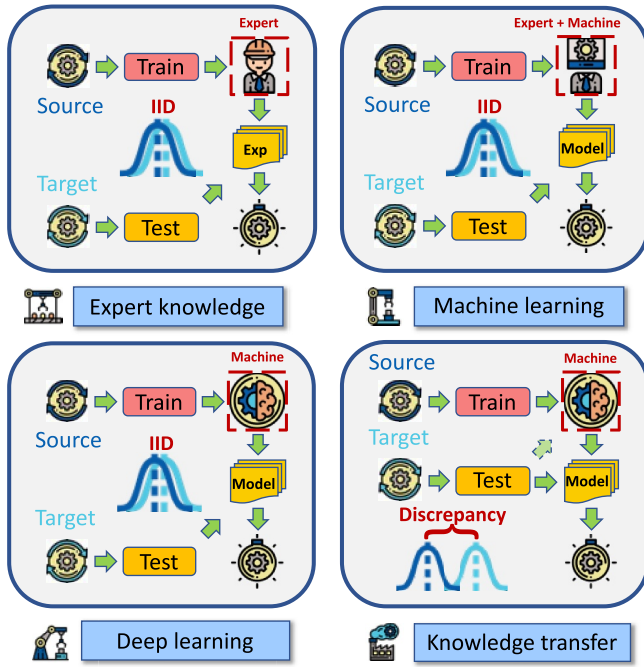


FIGURE 1 The development of fault diagnosis of rotary machines

distributions between the training and test data sets. Knowledge transfer can reduce reliance on labelled data and improve the generalisation ability of a model built from data following different distributions. These benefits have been demonstrated in applications such as computer vision, text classification, and time-series analysis.

A few survey papers on knowledge transfer have been published. Pan et al. [32] discussed the relationship between transfer learning and other approaches such as domain adaptation, multi-task learning, and sample selection bias as well as covariate shift. The definitions and applications of transfer learning were reviewed in Ref. [33]. Zhuang et al. [34] systematised the existing transfer learning research, summarised and interpreted the mechanisms and strategies applied. However, no survey paper related to knowledge transfer research in fault diagnosis was published prior to 2019. Zheng et al. [12] firstly reviewed advanced cross-domain research in fault diagnosis, mainly from an approach perspective. Meanwhile, Yan et al. [11] provided an overview of the latest developments in knowledge transfer of rotary machinery. A more recent survey on deep transfer learning in intelligent fault diagnosis focussed on industrial applications was published by Li et al. [13].

This paper tries to present a comprehensive review for knowledge transfer in fault diagnosis of rotary machines (KT-FDRM). The Scopus query string for this study follows ‘(TITLE-ABS-KEY [{knowledge transfer} OR {adaptation} OR {transfer learning} OR {Cross-domain}] AND TITLE-ABS-KEY [fault diagnosis])’. The search has been limited to the potential areas (engineering, physics and astronomy, decision sciences, neuroscience, and environmental science) related to the topic of this paper. The pioneering and representative efforts are reviewed in this study. Specifically, the main contributions of this review are summarised as follows:

- 1) The KT-FDRM paradigms are summarised while over 20 representative approaches with their variants and applications are introduced.
- 2) A problem-oriented taxonomy for KT-FDRM is proposed to explore the current advances and research frontiers.
- 3) Research challenges and future directions in KT-FDRM are explored from data, modelling, and application perspectives.

The remainder of this review is organised as follows: Section 2 presents an overview and preliminaries. Section 3 reviews the advanced knowledge transfer approaches and applications in fault diagnosis. Section 4 discusses research challenges and future directions. Section 5 concludes the paper.

2 | OVERVIEW

The terminology and definitions used throughout this paper are provided and standard knowledge transfer approaches based on knowledge content are introduced. A data-oriented taxonomy applicable to fault diagnosis of rotary machines is discussed.

2.1 | Terminology and definitions

The key terminology and definitions used throughout this paper are provided next. The frequently used symbols and their definitions are summarised in Table 1.

Fault detection: Indicates a component or a machine state that does not behave as expected;

Fault diagnosis: Determines the root cause of a fault;

Domain: $\mathcal{D} = \{\mathcal{X}, P(\mathbf{X})\}$ consists of two elements: a feature space \mathbf{X} , and a marginal probability distribution $P(\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$.

Task: $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ consists of two elements: a label space \mathcal{Y} and a predictive function $f(\cdot)$, where $f(\mathbf{X})$ is a conditional distribution $P(\mathbf{Y}|\mathbf{X})$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n \in \mathcal{Y}$.

Transfer learning: Given a source domain \mathcal{D}_S and a learning task \mathcal{T}_S , a target domain \mathcal{D}_T and learning task \mathcal{T}_T , transfer learning aims at the improvement of learning the target predictive function $f_T(\cdot)$ in \mathcal{D}_T using the knowledge in \mathcal{D}_S and \mathcal{T}_S , where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$.

Domain adaptation: (a) A special subtopic of transfer learning, where the source domain differs from the target domain $\mathcal{D}_S \neq \mathcal{D}_T$ and performs the same task $\mathcal{T}_S = \mathcal{T}_T$. (b) The notion of domain adaptation is closely related to transfer learning. In some papers, transfer learning is used interchangeably with domain adaptation.

Knowledge transfer: Knowledge transfer refers to sharing or disseminating knowledge and providing inputs to problem-solving.

Without the loss of generality, the term knowledge transfer used in this paper accommodates the definitions of transfer learning and domain adaptation.

TABLE 1 Frequently used symbols and definitions

Symbol	Definition
\mathcal{D}_S	Source domain
\mathcal{D}_T	Target domain
\mathcal{X}	Feature space
\mathcal{Y}	Label space
\mathbf{X}	Data matrix
\mathbf{Y}	One-hot label matrix
\mathbf{x}	Data vector
\mathbf{y}	Label vector
\mathcal{T}_S	Source task
\mathcal{T}_T	Target task
$P(\mathbf{X})$	Marginal probability distribution
$f(\cdot)$	Prediction function
G	Generator
D	Discriminator
\mathbf{X}_S	Source data matrix
\mathbf{X}_T	Target data matrix
\mathbf{M}_S	Source subspace base
\mathbf{M}_T	Target subspace base
\mathbf{C}_S	Source covariance matrix
\mathbf{C}_T	Target covariance matrix
\mathbf{W}	Parameter weight matrix
\mathbf{B}	Parameter bias vector
\mathbf{H}	Extracted feature matrix
\mathbf{Q}	Extracted feature matrix
$\hat{\mathbf{X}}$	Reconstructed data matrix
$\mathcal{V}(\cdot)$	Data discrepancy function
\mathcal{L}	Loss function
J	Objection function
λ	Regularisation factor
β	Regularisation factor
μ	Balance factor
ξ	Parameter sensitivity

2.2 | Knowledge transfer approaches

To date, different taxonomies of knowledge transfer have been published (see Figure 2). Knowledge transfer can be homogeneous and heterogeneous. Each of the two categories can be supervised, semi-supervised, and unsupervised.

The knowledge content transfer perspective leads to the following four categories: instance-based, feature-based, relationship-based, and model-based knowledge transfer. The

first three categories are at the data level while the last one is at the model level. Based on the data and model level, the developments and challenges related to fault diagnosis of rotary machines are captured in four categories: statistics transfer, deep transfer, adversarial transfer, and adaptive transfer.

Statistical transfer. Statistical transfer (or traditional transfer) learning algorithms are usually versed in statistical analysis aimed at solving the probability discrepancy problem. Statistical transfer methods are usually used in homogeneous environments where domain data share the same feature space and similar distributions. Instance reuse, distribution adaptation, feature selection, and SA are the most widely used strategies.

Deep transfer. Deep transfer (or deep network transfer) algorithms take advantage of their feature learning abilities to improve the task performance using knowledge from a source and/or target domain. The selection of features in traditional machine learning and statistical transfer is time consuming while they are autonomously extracted in deep learning. Deep Inherited transfer utilises the shared parameters from the pre-trained models while knowledge is transferred from source task \mathcal{T}_S to target task \mathcal{T}_T . Deep synchronous transfer methods extract features, reduce the domain discrepancy, and solve the target task \mathcal{T}_T simultaneously while deep synchronous transfer methods do it sequentially.

Adversarial transfer. Adversarial transfer algorithms usually involve deep (or generative adversarial) structures composed of a generator G and a discriminator D . They are designed to confuse the source and target domain for D and learn domain invariant features. A generative network can serve as a generator of new target data, thus dealing with data imbalance. Adversarial transfer has shown promising results in many applications, including fault diagnosis.

Adaptive transfer. Adaptive transfer algorithms improve performance in target fault diagnosis where the independent identical distribution holds in the source and target domains. Special network structures or algorithms are carefully designed to improve the robustness of diagnosis models of rotary machines. Ensembles are applied when a single model is not able to handle data variability.

2.3 | Problem-oriented tasks

From a problem-oriented perspective [35], 4 scenarios and 11 tasks in Table 2 are considered in the knowledge transfer challenges in fault diagnosis of rotary machines. According to the attributes on data and label, five application scenarios and their transfer divergence can be presented as follows:

2.3.1 | Homogeneous data tasks

A homogeneous data scenario defines the transfer tasks where both source and target domains share the same feature and label spaces while feature distribution divergence exists, that is, $\mathcal{D}_S = \mathcal{D}_T$ and $\mathcal{Y}_S = \mathcal{Y}_T$ while $P(\mathbf{X}_S) \neq P(\mathbf{X}_T)$. This scenario widely appears in changing working conditions, deteriorating

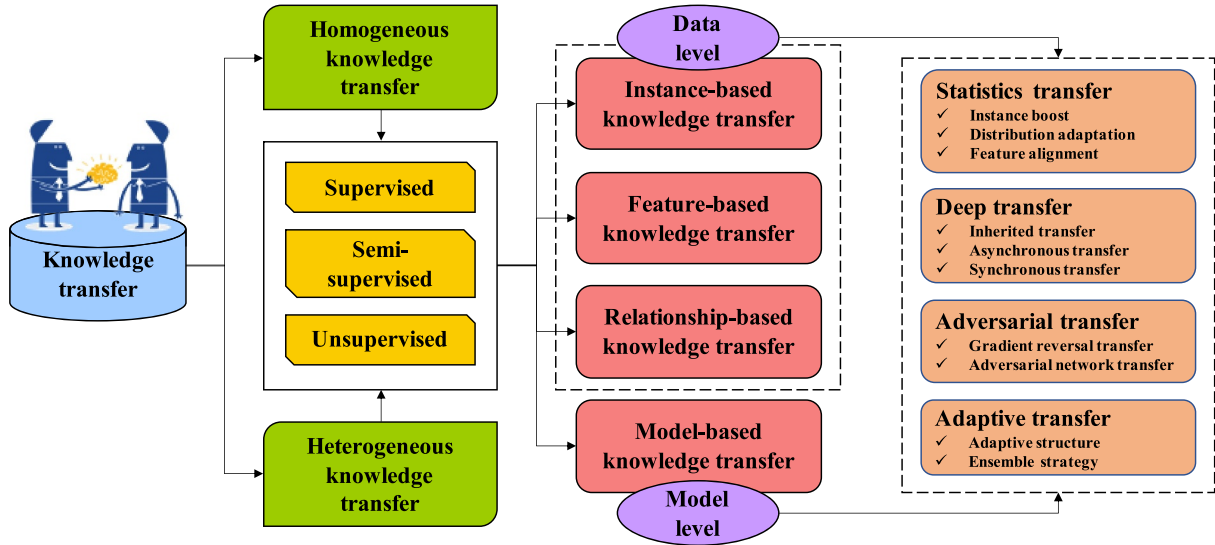


FIGURE 2 Taxonomy of knowledge transfer approaches in fault diagnosis

fault severities and relevant location between the sensor and faulty rotor part. According to the label accessibility in the task domain, homogeneous data transfer differs into four tasks: supervised, semi-supervised, unsupervised domain adaptation, and domain generalisation.

2.3.2 | Heterogeneous feature tasks

Heterogeneous feature scenario describes the transfer tasks, where source and target domains differ in feature spaces but share the same label spaces, that is, $\mathcal{D}_S \neq \mathcal{D}_T$ while $\mathcal{Y}_S = \mathcal{Y}_T$. This scenario occurs during transfer between sensors with different data type or sampling rate, and it suits a realistic collection divergence, where only massive labelled healthy condition data in the target domain are available while completely labelled faulty data in the source domain can be accessed. According to the label accessibility in the task domain, heterogeneous feature transfer differs into three tasks: supervised, semi-supervised, and unsupervised heterogeneous domain adaptation.

2.3.3 | Heterogeneous label tasks

Heterogeneous label scenario suits the transfer tasks, where both source and target share the same feature spaces while differs in label spaces, that is, $\mathcal{D}_S = \mathcal{D}_T$ while $\mathcal{Y}_S \neq \mathcal{Y}_T$. This scenario exists due to task divergence, such as transfer between different machines with different fault types, anomaly detection from normal conditions or multi-tasks problems.

2.3.4 | Heterogeneous data tasks

Heterogeneous data spaces scenario exists when source and target domains differ in both feature and label spaces, that is,

$\mathcal{D}_S \neq \mathcal{D}_T$ and $\mathcal{Y}_S \neq \mathcal{Y}_T$. This situation appears in many new applications, such as transfers from a digital twin system to a real physical system or transfer from a computation vision task to a fault diagnosis task, where domain divergence in both feature and label spaces widely exists.

Data diversity spawns varieties of transfer tasks and challenges in fault diagnosis as above. Different transfer learning algorithms have emerged in recent years, and we will introduce the representative knowledge transfer paradigms in the next part.

3 | APPROACHES AND APPLICATIONS

3.1 | Statistical transfer

Most of the traditional knowledge transfer algorithms before deep learning were versed in statistics. Limited fault diagnosis research based on statistical transfer left impressive achievements in the past decade. The surveyed statistical transfer in fault diagnosis can be grouped into three categories: (1) instance boost, (2) distribution adaptation, and (3) feature alignment.

3.1.1 | Instance boost

Instance transfer expands similar samples in the source domain to boost the performance of a classifier when limited labelled target data are available.

The importance and contribution of each instance, involved in KTFD problems, differ from each other. As a representative work of instance boost, TrAdaBoost proposed by Dai et al. [36] takes advantage of a similar iterative strategy like Adaboost [54] for transfer learning tasks. AdaBoost ensembles weak learners, iteratively trained with a unified weighting mechanism to all instances, to make a final decision. However, Adaboost only adapts to independent and identically distributed conditions.

TABLE 2 A problem-oriented taxonomy for knowledge transfer tasks in fault diagnosis (Abbreviations come from corresponding references)

Scenario	Task	X_S^L	X_S^U	X_T^L	X_T^U	Task name	Source label	Target label	Reference	The root of divergence
Homogeneous data	T1	1	0	1	0	Supervised domain adaptation	Fully labelled	Limited labelled	TrAdaBoost [36], MsTrAdaBoost [37], TaskTrAdaBoost [37]	Working conditions
	T2	1	0	1	1	Semi-supervised domain adaptation	Fully labelled	Limited labelled + massive unlabelled	FTNN [38]	Fault severities
	T3	1	0	0	1	Unsupervised domain adaptation	Fully labelled	Massive unlabelled	DTL [39], DCTLN [40]	Sensor locations
	T4	1	0	0	0	Domain generalisation	Fully labelled	Unavailable	AdaBN [41], MSDCNN [42], IDSCNN [43]	
Heterogeneous feature	T5	1	0	1	0	Supervised heterogeneous domain adaptation	Fully labelled	Limited labelled	MCDTN [44]	Sensor discrepancy
	T6	1	0	1	1	Semi-supervised heterogeneous domain adaptation	Fully labelled	Limited labelled + massive unlabelled	HT-SSAE [45]	Collection divergency
Heterogeneous label	T7	1	0	1	0	Supervised transfer learning	Fully labelled	Limited labelled	TLNN [46] (SNN [47])	Machines faulty types
	T8	1	0	0	1	Unsupervised transfer learning	Fully labelled	Massive unlabelled	DGNN [43] Cycle-GAN [48]	Anomaly detection multi-tasks
Heterogeneous data	T9	0	1	1	0	Self-learning	Unlabelled + (labelled)	Limited labelled	MMF [49]	Application objects
	T10	1	0	1	0	Supervised heterogeneous transfer learning	Fully labelled	Fully labelled	AlexNet [50], LeNet [40], VggNet [51]	
	T11	1	0	0	0	Semi-supervised heterogeneous transfer learning	Fully labelled	Limited labelled	DFDD [52], KTP [53]	

Conversely, TrAdaBoost designs two special weighting mechanisms to massive labelled training instances and limited labelled test instances, respectively. The key idea lies in TrAdaBoost is to assign high weights to the source data which are similar to the target ones while doing the opposite for the remainder. At the end of iteration training, half of the resultant weak classifiers are ensembled to a final classifier using a voting scheme. As a general training strategy, no specific classifier is assigned in this algorithm; thus, arbitrary machine learning- or deep learning-classifier can be used in TrAdaBoost.

The available multi-source scenario provides more chance of discovering sources closely related to the target. Yao et al. [37] extended TrAdaBoost to multi-source scenes and proposed a novel algorithm named MsTrAdaBoost. In each iteration, parallel pair training using each source and the unique target

generates multiple candidate classifiers. The classifier with minimal errors is selected to update the instance weights in its source domain. At the end of the algorithm, all selected classifiers in the iterations are ensembled to form the final classifier.

New emerging data from the machine monitoring system change the distribution of target domain. The TaskTrAdaBoost [37] algorithm enables rapid retraining over new targets. In TaskTrAdaBoost, a group of candidate classifiers based on AdaBoost and threshold filtering are trained on each source data, respectively. Then, a modified AdaBoost is performed to select the optimal classifier which has the highest prediction accuracy on target instances. The weights of target instances are updated based on the performance of a selected classifier. Finally, all the selected classifiers are ensembled to predict the target task.

The representative fault diagnosis applications using instance boost can be found in [55, 56].

3.1.2 | Distribution adaptation

Data distribution discrepancy widely lies in cross-domain problems. Distribution adaptation aims to reduce the discrepancy by distribution alignment. We surveyed four typical distribution adaptation algorithms in knowledge transfer: marginal distribution adaptation (MDA), conditional distribution adaptation (CDA), joint distribution adaptation (JDA), and balanced distribution adaptation (BDA).

Marginal distribution adaptation

In general, if the generation mechanism of two failure datasets is different, their marginal probability distributions differ from each other in the feature spaces. For example, bearing and gear generate different marginal distributions due to the different failure generation mechanisms. MDA is suitable for heterogeneous feature tasks, where the feature spaces are different. Pan et al. [57] proposed a transfer component analysis (TCA) algorithm to deal with the marginal distribution mismatch problem. TCA learns a mapping function that projects all raw data into a subspace spanned by transfer components using maximum mean discrepancy. In the subspace, the data properties are preserved in a reduced dimension while the data distribution from both source and target domains are close to each other. The marginal distribution discrepancy decreases on the transfer components. Some TCA variants can be found in Ref. [58].

Conditional distribution adaptation

Conditional distribution discrepancy lies in the relationship between data and labels. For example, the failure categories of the two machines are different, which leads to the condition distribution discrepancy, and thus, CDA is suitable for heterogeneous label tasks. Wang et al. [59] proposed a stratified transfer learning (STL) algorithm to deal with the conditional distribution mismatch problem. STL first annotates pseudo labels to the target candidates via a majority voting strategy through basic learners trained on the source domain. Then, transfer components with their pseudo labels were iteratively updated in the projected subspace by solving an eigendecomposition problem based on a new kernel matrix and discrepancy criterion. The above iteration stops when the convergence condition is reached, and the last annotated pseudo labels are regarded as the prediction results. Similar CDA algorithms can be found in Ref. [60].

Joint distribution adaptation

Joint distribution discrepancy is realistic in machine monitoring. The varying working condition changes the marginal distribution while the new failure type with deteriorating severities changes the conditional distribution, and thus, both MDA and CDA are suggested to be considered in real complex fault diagnosis. Long et al. [61] proposed a JDA algorithm to reduce

marginal discrepancy and conditional discrepancy simultaneously. The JDA algorithm employs a kernel method and pseudo-label learning to reduce marginal- and conditional-discrepancy, respectively. The Lagrange method and singular value decomposition are used to optimize the objective function of JDA. An iterative strategy is taken to improve the distribution alignment with a basic classifier. Finally, the basic classifier outputs predictions when the convergence condition reaches. However, JDA assumes that the importance of MDA and CDA are the same, which may not hold in real applications.

Balanced distribution adaptation

In the knowledge transfer task for fault diagnosis, the importance of marginal- and the conditional-distribution discrepancy may be different. It's hard to judge whether varying working conditions or deteriorating severities dominate the influence. Wang et al. [62] proposed a BDA algorithm to minimise the weighted distribution divergence in both marginal discrepancy and conditional discrepancy. In BDA, the balance factor leverages the different importance in its objective function. The BDA algorithm can adaptively adjust the balance factor μ according to the specific data field. When $\mu \rightarrow 0$, it indicates that the significant discrepancy lies in both source and target data, and thus minimising the marginal discrepancy is more important. When $\mu \rightarrow 1$, it indicates that both domains share similarities, so trying to minimise the conditional distribution discrepancy is more appropriate. Both JDA and BDA are suitable for heterogeneous data tasks.

The representative fault diagnosis applications using distribution adaptation transfer can be found in MDA [58, 63–65], JDA [66, 67], and BDA [68, 69].

3.1.3 | Feature alignment

In addition to explicit features alignment mentioned in instance boost and distribution adaptation, feature alignment tries to align implicit features, such as principal features, geometric features, and statistical features, in the subspace of original data. A typical feature alignment method usually contains three parts: subspace generation, SA, and classifier training [34].

Principal feature alignment

The raw time series, in machine condition monitoring, usually have high dimensions, which lead to dimension disaster. Feature extraction with reduced dimension is essential for KTFD problems. Fernando et al. [70] proposed an unsupervised domain adaptation algorithm named subspace alignment (SA). In SA, the source and target data are first projected into a subspace of principal components by applying principal component analysis. The leading eigenvectors are selected as the subspace bases, \mathbf{M}_S and \mathbf{M}_T . Then, a linear mapping problem $\arg \min_{\mathbf{W}} \|\mathbf{M}_S \mathbf{W} - \mathbf{M}_T\|_F^2$ is solved, and the optimal solution $\mathbf{W} = \mathbf{M}_S^T \mathbf{M}_T$ is produced. The SA algorithm transforms the source subspace coordinate system into the target subspace coordinate system. Finally, the aligned representation

in the source and target subspace is expressed as $\mathbf{X}_S \mathbf{M}_S \mathbf{W}$ and $\mathbf{X}^T \mathbf{W}$, respectively. Any classifier can be trained using the aligned features.

Geometric feature alignment

Gong et al. [71] proposed an unsupervised domain adaptation algorithm named geodesic flow kernel (GFK). The GFK projects the raw data into the geodesic flow subspace [72] with manifold space maintaining the geometric data attributes. The source and target data in the manifold space are considered as two points, and the transfer strategy is to find a sampling geodesic flow (SGF) and transfer the source features along SGF to the target features.

Statistical feature alignment

Sun et al. [73] proposed an unsupervised domain adaptation algorithm named correlation alignment (CORAL). CORAL aligns the second-order statistics of the source and target distributions, that is, the covariance matrices \mathbf{C}_S and \mathbf{C}_T from D_S and D_T , respectively. Given a linear transformation \mathbf{W} and using the Frobenius norm to measure the discrepancy, the optimisation problem is represented as $\arg \min \|\mathbf{W}^T \mathbf{C}_S \mathbf{W} - \mathbf{C}_T\|_F^2$.

The representative fault diagnosis applications using feature alignment can be found in SA [64], GFK [74], and CORAL [75].

3.2 | Deep transfer

Most recent studies on transfer learning in fault diagnosis of rotary machines have focussed on deep neural networks with

categories of transfer methods: (1) inherited transfer, (2) synchronous transfer, and (3) asynchronous transfer.

3.2.1 | Inherited transfer

Inherited transfer (or parameter-sharing) leverages knowledge learnt from a source domain to improve the performance of the target task (see Figure 3). The inherited transfer is widely used when the target data are limited or training from a scratch is time consuming. In the source task, a neural network is usually randomly initialised as shown in Figure 3. Then, parameters of the neural network are optimised for a source task. In the inherited transfer, the feature domains \mathcal{D}_S and \mathcal{D}_T are similar while the tasks can be different. For a classification task such as fault diagnosis, the classifier at the last layer is usually randomly initialised and then trained in the transfer process. Based on the layer or neurons sharing the weights (Figure 3), four options are followed: (1) fine-tune classifier, (2) fine-tune top layers, (3) fine-tune all layers, and (4) selective fine-tuning.

Fine-tune classifier

All parameters ahead of the classifier are inherited and frozen from the trained model, and only the parameters of a classifier are fine-tuned in the target training.

Fine-tune top layers

All parameters in the top layers including the classifier and top feature extractor layers are fine-tuned while the remaining parameters inherit the parameters from the trained model.

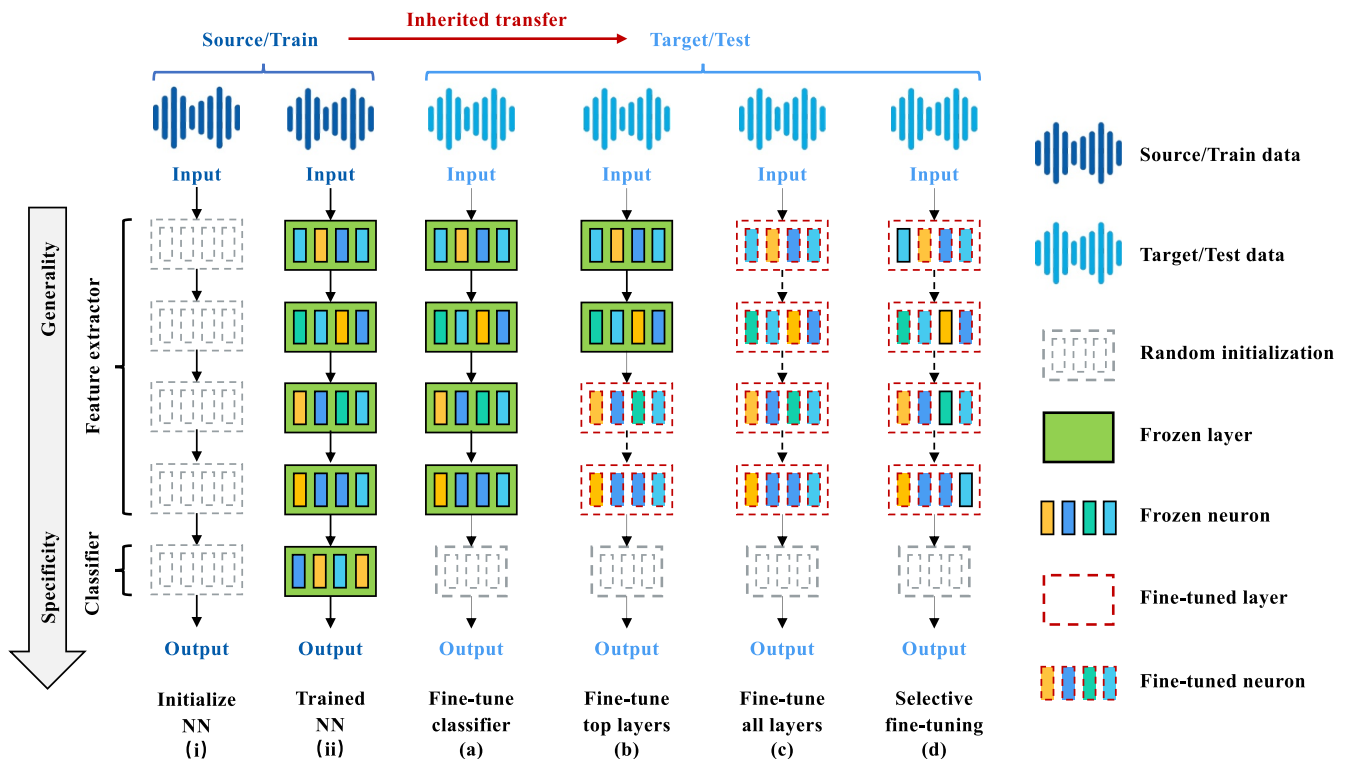


FIGURE 3 Inherited transfer approaches from the source to target task

Fine-tune all layers

All parameters of the feature extractor and classifier are initiated with the pre-trained parameters and then fine-tuned in the target training process.

Selective fine-tuning

According to the feature importance, only the partial neurons of the partial layers are inherited and frozen while the remaining neurons inherit the parameters from the trained model and are fine-tuned in the target training.

Given the inherited parameters θ_i and the fine-tuning parameters θ_j , the inherited transfer is expressed in Equation (1).

$$\begin{aligned} \left. \begin{array}{l} \{X_S, Y_S\} \\ f_S(\cdot) \end{array} \right\} &\xrightarrow{\text{Train}} f_S(\theta_i, \theta_f) \rightarrow \hat{Y}_S \\ &\downarrow \text{init.} \\ \{X_T, Y_T\} &\xrightarrow{\text{Fine-tune}} f_T(\theta_i, \theta_f) \rightarrow \hat{Y}_T \end{aligned} \quad (1)$$

In many fault diagnosis problems, the source domain $\mathcal{D}_S = \mathcal{X}_S, P(\mathbf{X}_S)$ and the target domain $\mathcal{D}_T = \{\mathcal{X}_T, P(\mathbf{X}_T)\}$ share the same feature space $\mathcal{X}_S = \mathcal{X}_T$, but the marginal probability distributions are different $P(\mathbf{X}_S) \neq P(\mathbf{X}_T)$. The feature space \mathcal{X} can be represented in the time domain, frequency domain, time-frequency domain or in expert-designed domain. The marginal probability discrepancy may stem from the changing working conditions, such as variable loads and variable rotational speed.

Training a new fault diagnosis model from scratch with limited fault data is not only time consuming but also difficult to guarantee the model's accuracy. Zhang et al. [76] applied inherited transfer to the time series data by using a sliding frame with a fixed step. The source diagnosis model \mathbf{M}_S was pre-trained with labelled data. Then, the target model \mathbf{M}_T was initialised with the pre-trained parameters. The top layers and the classifier of the target faults were replaced for the target task, and the entire model was fine-tuned. Their experimental results demonstrated that the proposed inherited transfer method improved the fault diagnosis accuracy and reduced the training time based on limited labelled target data. Xu et al. [50] applied a similar approach to a digital-twin-based system. The pre-trained model \mathbf{M}_S in virtual space was used to initialise the target model \mathbf{M}_T in the physical space for real-time prognosis and health management. A stacked sparse autoencoder was utilised to extract a meaningful representation of raw data in both virtual and physical domains. Fine-tuning was used to train \mathbf{M}_S and \mathbf{M}_T . Wen et al. [39] and Hasan et al. [77] transformed the raw time signals into images and applied CNN models to deal with the inherited transfer problems.

The performance of a fault diagnosis model depends on the input parameters. Though a CNN can extract meaningful features from the raw data, however, the deep model needs to be carefully designed for the application considered. The research community has benefited from open-source CNN models, such as AlexNet, LeNet, and VggNet that have been applied in fault diagnosis [39, 50, 51, 78, 79]. Cao et al. [50] proposed a preprocessing-free algorithm for gear fault diagnosis. They applied bicubic interpolation to transform

vibration signals into greyscale images and used AlexNet to extract high-level data representations. The classification layer was replaced for a specific gear fault diagnosis task. Classifier fine-tuning was executed to train the fault diagnosis model. Their algorithms achieved great success with a limited training data. Wen et al. and Liu et al. applied a similar concept to LeNet in the diagnosis of faults from bearings and pumps. Shao et al. [51] combined VGGNet and fine-tuning in the top layers for accurate machine fault diagnosis. The performance of their proposed approach was validated with three fault diagnosis datasets: induction motor, gearbox, and bearing.

The initial layers of a deep neural network usually learn general features, while the deeper layers focus on learning feature specifics. In summary, transfer learning is more effective if general features are inherited in the shallow layer and fine-tuned in deeper layers.

The freeze and fine-tune method needs are carefully selected, otherwise a negative transfer from \mathcal{T}_S to \mathcal{T}_T will reduce the target performance. Han et al. [78] investigated the transfer abilities of three inherited transfer methods mentioned in Figure 3a–c. The rules of thumb applied for the method selection are shown in Table 3.

Kim et al. [76] proposed a new inherited transfer method, selective parameter freezing (SPF), integrating fine-tuning and freezing in one layer as shown in Figure 3d and Figure 4. The sensitivity $\xi_i = \|\Sigma_k(\partial \hat{\mathbf{y}}_k / \partial \theta_i)^2 / N_k\|$ of NN parameter θ_i reflects its impact on the performance of the source task. The trainable parameters for the target task were selected based on hyperparameter α . The experimental results indicated that the SPF-based inherited transfer with preselected α outperformed all fine-tuning methods.

3.3 | Asynchronous transfer

Deep asynchronous transfer in fault diagnosis involves feature extraction and classifier training. Feature extraction aims at learning the general representation from the source and target domains. Classifier training aims to deal with classification tasks. An autoencoder includes an encoder and a decoder.

Encoder: creates a compressed latent space representation from the input data.

Decoder: reconstructs the compressed representation of the input data.

A multi-layer auto-encoder with non-linear activation functions learns complex non-linear representation from the source and target domains. It offers the following benefits:

- Supports an arbitrary network structure from dense layers to the convolutional layers.
- Unsupervised learning method without expensive and time-consuming labels.
- Multiple efficient representations at different layers with different dimensions.
- Robust denoising ability to complex application scenarios.
- Reusable pre-trained layers for transfer learning.

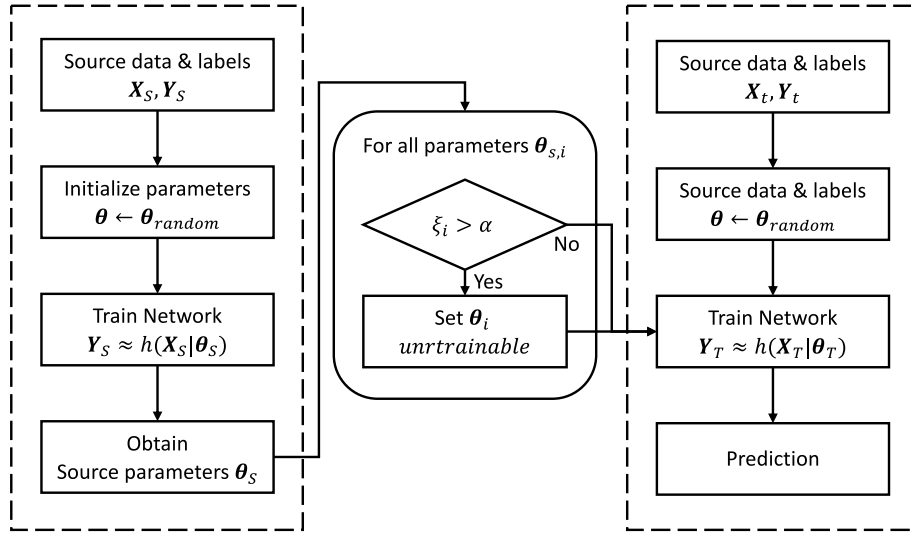


FIGURE 4 Procedure of selective parameter freezing [76]

TABLE 3 Experimental rules for selection of the fine-tuning method

		Size of the target dataset		
		Low	Medium	High
Discrepancy between source and target	Low	Freeze	Freeze or tune	Freeze
	Medium	Freeze or tune	Freeze	Freeze
	High	Freeze or tune	Freeze	Freeze

Given encoder parameters $(\mathbf{W}_1, \mathbf{b}_1)$, fault diagnosis classifier parameters $(\mathbf{W}_2, \mathbf{b}_2)$, decoder parameters $(\hat{\mathbf{W}}_1, \hat{\mathbf{b}}_1)$ and discrepancy criterion \mathcal{V} , the autoencoder-based asynchronous transfer in fault diagnosis is presented in Equation (2).

$$\begin{array}{c}
 \mathbf{X}_S \xrightarrow{(\mathbf{W}_1, \mathbf{b}_1)} \mathbf{H}_S \xrightarrow{(\mathbf{W}_2, \mathbf{b}_2)} \hat{\mathbf{Y}}_S \\
 \mathcal{V}(\mathbf{H}_S, \mathbf{H}_T) \leftarrow \begin{array}{l} \downarrow (\mathbf{W}_1, \hat{\mathbf{b}}_1) \hat{\mathbf{X}}_S \\ \uparrow (\mathbf{W}_1, \hat{\mathbf{b}}_1) \hat{\mathbf{X}}_T \end{array} \\
 \mathbf{X}_T \xrightarrow{(\mathbf{W}_1, \mathbf{b}_1)} \mathbf{H}_T \xrightarrow{(\mathbf{W}_2, \mathbf{b}_2)} \hat{\mathbf{Y}}_T
 \end{array} \quad (2)$$

Asynchronous transfer in fault diagnosis involves the objectives as below:

Reconstruction error minimisation: The reconstruction error \mathcal{L}_{Rec} between input-output pairs $(\mathbf{X}_S, \hat{\mathbf{X}}_S)$ and $(\mathbf{X}_T, \hat{\mathbf{X}}_T)$ should be minimised as close as possible.

Discrepancy criterion minimisation: $\mathcal{V}(\mathbf{H}_S, \mathbf{H}_T)$ between compressed features $(\mathbf{H}_S$ and $\mathbf{H}_T)$ from a source domain \mathcal{D}_S and target domain \mathcal{D}_T should be minimised.

Diagnosis error minimisation: The diagnosis error \mathcal{L}_{dig} between diagnosis output $\hat{\mathbf{Y}}_S$ and the ground truth label \mathbf{Y}_S should be minimised.

Therefore, the general basic objective function of pre-trained feature learning is provided in Equation (3):

$$\min J_{\text{SAE}} = \mathcal{L}_{\text{Rec}}(\mathbf{X}, \hat{\mathbf{X}}) \quad (3)$$

The overall objective function of the diagnosis classifier is shown in Equation (4):

$$\min J_C = \mathcal{L}(\mathbf{Y}_S, \hat{\mathbf{Y}}_S) + \lambda \mathcal{V}(\mathbf{H}_S, \mathbf{H}_T) \quad (4)$$

Lu et al. [80] introduced an asynchronous transfer in fault diagnosis. The trained strategy of their fault diagnosis model deep neural network for domain adaptation in fault diagnosis (DAFD) consists of three stages: (1) Initialisation: A single layer encoder and decoder is employed to obtain the compressed representation with the source data and object function $\min J_{\text{SAE}} = \mathcal{L}_{\text{Rec}}(\mathbf{X}, \hat{\mathbf{X}})$ during the initialisation training process. (2) Training DAFD: The source and target data are both utilised to achieve the objection $\mathcal{L}_{\text{DAFD}} = \mathcal{L}_{\text{ae}} + \lambda \mathcal{L}_{\text{MMD}} + \frac{\mu}{2} \mathcal{L}_{\mathcal{W}}$, where \mathcal{L}_{ae} is the reconstruction error, \mathcal{L}_{MMD} is the discrepancy distance and $\mathcal{L}_{\mathcal{W}}$ is used for keeping the weights from zero and strengthen the specific ones. The transferable features are obtained through the training process. (3) Training the domain adaptation classifier: The final classifier SVM is determined and trained from the labelled transferable source features and then deployed to make predictions on target datasets.

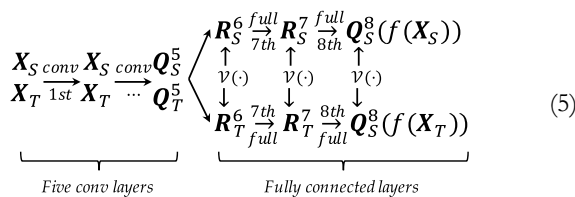
Wen et al. [40] proposed a two-stage asynchronous transfer fault diagnosis based on deep transfer learning (DTL). Compared with the DAFD algorithm, they merged transferable feature extraction and classifier training by defining the final objective function as $\mathcal{L}_{\text{DTL}} = \mathcal{L}_{\text{dig}}(\mathbf{Y}_S, \hat{\mathbf{Y}}_S) + \mu \text{MMD}(\xi^S, \xi^T)$. In the expression \mathcal{L}_{DTL} , ξ^S and ξ^T represents the sparse auto-encoder feature matrices learnt by the objection

function $\mathcal{L}_{SSAE} = \mathcal{L}(\mathbf{X}, \hat{\mathbf{X}}) + \beta KL(\rho\rho_k) + \mu R(\mathbf{W}, \mathbf{b})$, where the first term represents the reconstruction loss, the second one is a sparse penalty term using KL-divergence and the third one is a regularisation item which is used to prevent overfitting. Compared with DAFD, DTL employs the sparse autoencoder as the initialisation to extract more useful representations while deducing the redundant information in the latent layer. A similar algorithm following the same asynchronous strategy was deployed in a digital-twin-assisted fault diagnosis using deep transfer learning method [52].

Duan et al. [81] proposed an auxiliary-model-based domain adaptation algorithm for fault diagnosis under variable conditions. The basic supervised CNN functioned as an auxiliary model and provided the initialised features from a source and target domain. Then, the marginalised stacked denoising autoencoder (mSDA, a statistical transfer method) was employed to reduce the discrepancy between the source and the target features. Finally, the classifier SVM was trained with labelled source features and a fault was predicted by feeding target features into the trained classifier.

Some asynchronous transfer variants and their applications can be found in Ref. [44, 66, 81–84]. Qian et al. [85] used the high-order Kullback–Leibler constraint in both fault feature extraction and fault feature classification. Gao et al. [84] proposed a new robust fault feature extraction method based on contractive stacked autoencoders. Chen et al. [83] used autoencoder to extract features, and the same parameters were inherited from source trained model to the target model except for the last layer due to the different fault diagnosis tasks. Qian et al. [82] combined autoencoder with adaptive batch normalisation (AdaBN), which enhances the domain adaptability of fault diagnosis model from a source domain to target domain. Chen et al. [44] ensembled multiple autoencoders to deal with missing data problem under multi-rate sampling.

The discrepancy criterion defines the difference between source features and target features. Various discrepancy criteria have been widely used in transfer learning. Among all the criteria, the maximum mean discrepancy (MMD) was the most popular one adopted in both discussed asynchronous transfer and upcoming synchronous transfer.



3.3.1 | Synchronous transfer

Deep synchronous transfer integrates feature leaning and classifier learning. The domain transfer and discrepancy alignment of data distribution are implemented in the training process.

The synchronous transfer usually embeds the feature discrepancy into its objective function and minimises the

distribution discrepancy between both domains in the learning process. The synchronous transfer is expressed in Equation (5):

The first shallow layers in Equation (5) extract the general features in source and target domains layer by layer. The distribution discrepancy constraint \mathcal{V} are deployed to reduce the specific feature difference between high-level layers. It should be noted that the discrepancy constraint can also be applied in the shallow layers if additional time costing for training is acceptable, or the accuracy of fault diagnosis needs to be improved.

A general basic objective function for the feature-extractor transfer includes two terms: a diagnosis error loss \mathcal{L}_C and a discrepancy constraint term \mathcal{L}_V . A trade-off coefficient λ is set to maintain the balance between \mathcal{L}_C and \mathcal{L}_V . Finally, the final objective function can be defined in Equation (6):

$$\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_V \quad (6)$$

Xiao et al. [86] proposed a domain adaptive motor fault diagnosis algorithm based on feature-extractor transfer. They employed a CNN framework as a multi-level feature extractor. The divergence of features, from generality to speciality, among all intermediate layers were measured by maximum mean discrepancy (MMD). They incorporated this regularisation term with cross-entropy loss for fault diagnosis in the training process and imposed constraints on weight updating to minimise the distribution mismatch in both domains. The experimental comparison with several mentioned statistical transfer (such as TCA and JDA) and deep transfer (such as deep convolutional transfer learning network [DCTLN] and feature-based transfer neural network [FTNN]) demonstrated its effectiveness and superiority.

Two similar synchronous transfer methods can be found in Ref. [87, 88]. Li et al. [87] used multi-kernel maximum mean discrepancies (MK-MMD) to measure the mismatch between the source and target domains, and their bearing fault diagnosis model was constrained to reduce the discrepancy criterion during the training process. An et al. [88] replaced the MMD discrepancy with a kernel method in reproducing the kernel Hilbert space, multiple radial basis function (RBF) kernels in the intermediate layers, and different layer weights from the discrepancy term in the final objective function.

Han et al. [67] introduced a JDA into deep synchronous transfer. The domain adaptation aims to optimise the combined objective function $\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_V(J_S, J_T)$, where J_S and J_T are the joint probability distributions of \mathcal{D}_S and \mathcal{D}_T , respectively. An iterative optimisation strategy was utilised in the transfer algorithm. The network was first initiated with a source dataset for generating pseudo labels in target samples. Then, the regularisation term of JDA and the network optimisation were updated sequentially, and new pseudo labels were updated based on the optimised network. The iterative optimisation ended until the convergence or the last two generated pseudo labels equals $\hat{\mathbf{Y}}_i = \hat{\mathbf{Y}}_{i-1}$ in the target domains. A similar synchronous transfer with JDA applied in power equipment fault diagnosis can be found in Ref. [89].

All the four mentioned synchronous transfer fault diagnosis algorithms were similar in structure and objective

function. The main difference lies in the different selected discrepancy criteria and coefficients among criteria and layers. Some variants based on the general synchronous transfer were proposed with auxiliary regularisation term \mathcal{L}_A in Equation (7).

$$\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_V + \beta \mathcal{L}_A \quad (7)$$

Yang et al. [38, 90] incorporated pseudo label learning in the FTNN domain adaptation process. The final objective functions contained three regularisation terms: $J_{\text{FTNN}} = \mathcal{L}_C(\mathbf{Y}_S, \hat{\mathbf{Y}}_S) + \lambda \mathcal{L}_V + \beta \mathcal{L}(\hat{\mathbf{Y}}'_T, \hat{\mathbf{Y}}_T)$. In this expression, the labelled prediction error, the discrepancy difference, and the pseudo learning error with trade-off coefficients were defined sequentially. During the training process, the pseudo labels of target unlabelled samples were evaluated with the maximum predicted probability based on the currently trained classifier. The auxiliary regularisation term with pseudo labels enlarges the inter-class distance while reducing the intra-class distance of target features. The discrepancy alignment and the fault diagnosis were improved with the help of pseudo-label learning.

Li et al. [91] introduced distance metric learning in their bearing fault diagnosis algorithm. A representation clustering algorithm, like FTNN, was proposed to improve the cross-domain ability. They embedded a cluster regularisation term $\mathcal{L}_{\text{Cluster}}$ into the object function as $\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_V + \beta \mathcal{L}_{\text{Cluster}}$. The discrepancy alignment between the source and target domains was implemented through the last two terms: \mathcal{L}_D and $\mathcal{L}_{\text{Cluster}}$. The MK-MMD discrepancy \mathcal{L}_D in this research study was measured on the fully connected layer with specific features. The representation clustering term was formulated as $\mathcal{L}_{\text{Cluster}} = -\mathcal{V}_{\text{inter}} + \eta \mathcal{V}_{\text{intra}}$. $\mathcal{V}_{\text{inter}}$ and $\mathcal{V}_{\text{intra}}$ measure the inter-class separability and intra-class compactness, respectively, with Euclidean distance, and the coefficient η was set as a trade-off between these two clustering metrics. Experimental results have validated the robustness against environmental noises and variability of working conditions.

Guo et al. [92] proposed a DCTLN including two modules: condition recognition and domain adaptation. The condition recognition was trained to recognise the health conditions, while the domain adaptation was trained to simultaneously improve the discrepancy alignment. Three items, classifier loss \mathcal{L}_C , distribution discrepancy \mathcal{L}_V , and domain classifier \mathcal{L}_d , were included in the objective function $\mathcal{L} = \mathcal{L}_C + \lambda - \beta \mathcal{L}_d$. A Gaussian RBF was selected to estimate the MMD between domains in the final fully connected layer. It should be noted that an adversarial concept was employed in the domain classifier. If a domain classifier is not able to discriminate between the source and the target domain features, the features are domain invariant. The third component \mathcal{L}_d in \mathcal{L} of the objective function optimisation maximises the domain classification error, and the coefficient is negative. The effectiveness of DCTLN was verified with six transfer tasks across three different datasets.

The deep transfer methods in fault diagnosis are categorised as follows: (1) Inherited transfer might be the

simplest to be implemented and avoids training from scratch under limited data. Fine-tuning is essential for better performance and selective fine-tuning is promising. (2) Synchronous transfer separates feature learning and classifier learning for better initialisation of the feature extractor. Different autoencoders have been used to improve the robustness of feature representation in different domains. Aiming at improving fault diagnosis performance, some discrepancy criteria were incorporated in the classifier learning. (3) Asynchronous transfer combines feature learning and classifier learning in deep network optimisation. Some discrepancy alignment algorithms have emerged. Without enlisting an autoencoder for a specific fault diagnosis task, asynchronous transfer shows its generalisation abilities in fault diagnosis of rotary machines.

3.4 | Adversarial transfer

Deep learning-based GANs have gained popularity in recent years. The original GAN, including a generator G and a discriminator D , was inspired by game theory aiming at solving the min-max two-player game problem. The generator G is responsible for generating fake samples \mathbf{X}_F from a random noise vector \mathbf{v}_R approximating the true samples, while the discriminator D plays a role to distinguish between the fake or true samples \mathbf{X}_T and predict its label $\hat{\mathbf{Y}}$. The generator G and discriminator D are trained alternately until the equilibrium is attained. The GAN is illustrated in Equation (8).

$$\begin{array}{ccccc} & & \mathbf{X}_T & \xrightarrow{D} & \mathbf{Y}_T \\ & & \uparrow & & \\ \mathbf{v}_R & \xrightarrow{G} & \mathbf{X}_F & & \end{array} \quad (8)$$

Adversarial transfer has been researched in fault diagnosis. Based on the adversarial strategies of GAN, two categories are considered: (a) gradient reversal transfer (GRT) and (b) adversarial network transfer (ANT).

3.4.1 | Gradient reversal transfer

Compared with the traditional deep learning that involves a feature extractor and a label predictor, gradient reversal transfer incorporates a domain classifier after the feature extractor. Traditional deep learning usually involves pre-training. Then, the feature extractor functions as a generator G that attempts to align the distribution discrepancy from both domains and generates confusing features to the domain classifier. At the same time, the domain classifier acts as a discriminator D distinguishing the feature origin. A gradient reversal layer is usually employed to facilitate the backpropagation. After reaching the equilibrium following the adversarial training, the distribution discrepancy from the source and target domains is reduced. Finally, the label predictor is used as a classifier to diagnose

machine health conditions. The diagram of gradient reversal is illustrated in Equation (9).

$$\begin{array}{c} \mathbf{X}_S^L \xrightarrow{f_F} \mathbf{Q}_S^L \xrightarrow{f_D} \mathbf{Y}_S^L \text{ Class label} \\ \mathbf{X}_T^U \xrightarrow{f_F} \mathbf{Q}_T^U \xrightarrow{f_D} \mathbf{Y}_T^U \text{ Domain label} \end{array} \quad (9)$$

Han et al. [93] introduced an adversarial convolutional neural network (DACNN) into fault diagnosis in rotary machines. They adopted the adversarial learning strategy on the last one fully connected layer in the above diagram. By adding an additional discriminative classifier into the CNN-based deep network, their fault diagnosis model learns a more robust feature representation and becomes more generalised to the domains with different distributions and limited labelled samples.

Guo et al. [92] also brought a similar adversarial strategy DCTLN into the synchronous transfer as we have discussed above. Compared with DACNN, DCTLN not only integrates gradient reversal transfer into the deep networks but also includes the distribution discrepancy into its objective function, so the robust features extracted from both source and target domains are constrained from both adversarial learning and discrepancy alignment learning.

Recently, Zhang et al. [94] proposed a novel fault diagnosis algorithm based on Wasserstein distance guided multi-adversarial networks (WDMAN). Compared with DACNN and DCTLN, WDMAN applies multiple gradient adversarial learning and Wasserstein distance constraints on the last two fully connected layers. The shared domain invariant features are learnt during adversarial learning.

Zhang et al. [95] constructed a parallel network structure named A2CNN for source and target tasks separately, which differs from the shared network in DACNN, DCTLN, and WDMAN. The source network is first trained with massive labelled data in the source domain. Then, the target network inherits parameters from the pre-trained one and keeps the parameters frozen in shallow layers for general feature representations. Unfrozen top layers in both source and target networks are trained adaptatively with a domain discriminator.

Wang et al. [96] made a benchmark study among DANN-based, MMD-based and AdaBN-based transfer learning methods. They found that the MMD-based method can achieve the best diagnosis accuracy with the most training time while the AdaBN-based method can speed up twice with a small diagnosis accuracy decreasing. The gradient reversal transfer-driven DANN diagnosis algorithm can keep a good balance between diagnosis accuracy and training time.

3.4.2 | Adversarial network transfer

Aforementioned gradient reversal transfer methods assume that sufficient testing data under known fault types are available for training, which is unrealistic in real machine condition monitoring. An imbalanced monitoring environment, where massive healthy data are available while faulty data are rare, makes it impossible to use gradient reversal transfer methods. Fortunately, the generator in GAN can generate realistic data

under varying environments. Some research based on ANT brings promising achievements to overcome the problem when testing faulty data are unavailable.

Li et al. [97] proposed a two-stage ANT for fault diagnosis when only healthy target data are available. Assume that there are N_C-1 types of faulty conditions, and one healthy condition in labelled training data. The j th types of faulty data and healthy data can be remarked as \mathbf{X}_{S,F_j}^L and $\mathbf{X}_{S,H}^L$, respectively. In the first stage, a feature extractor f_F , a basic fault diagnosis classifier f_C and N_C-1 generators f_{G_j} , which map generated faulty data from the healthy condition to each fault class, were trained. The diagram of the first stage can be presented in Equation (10). \mathbf{P}_j denotes the high-level representation matrix of j th real faulty type while \mathbf{Q}_j denotes the j th generated faulty representation matrix by generator f_{G_j} . A generator loss is defined with multi-kernel MMD discrepancy criterion as $\mathcal{L}_g = \sum_{j=1}^{N_C-1} \alpha_j \text{MMD}_k(\mathbf{P}_j, \mathbf{Q}_j)$, where $\mathbf{P}_j = f_F(\mathbf{X}_{S,F_j}^L, F_j)$ and $\mathbf{Q}_j = f_{G_j}(f_F(\mathbf{X}_{S,H}^L))$, and the objective function can be represented as $\mathcal{L}_1 = \mathcal{L}_{C1} + \lambda_1 \mathcal{L}_g$, where the coefficient λ keeps the balance between the diagnosis loss \mathcal{L}_{C1} and generator loss \mathcal{L}_g .

$$\begin{array}{c} \mathbf{X}_{S,H}^L \xrightarrow{f_F} \mathbf{H}_{S,H}^L \xrightarrow{f_{G_j}} \mathbf{Q}_j \\ \mathbf{X}_{S,F_j}^L \xrightarrow{f_F} \mathbf{H}_{S,F_j}^L \longleftrightarrow \mathbf{P}_j \end{array} \quad (10)$$

In the second stage, they defined an accuracy loss \mathcal{L}_{C2} of fault diagnosis classifier f_{C2} and a distribution discrepancy loss \mathcal{L}_d between source and target domains in the objective function as $\mathcal{L}_2 = \mathcal{L}_{C2} + \lambda_2 \mathcal{L}_d$. The distribution discrepancy loss is designed to reduce the distribution misalignment by minimising $\text{MMD}_K(\mathbf{P}_f, \mathbf{Q}_f)$, where \mathbf{P}_f and \mathbf{Q}_f denote the integrated representation matrix from source and target domains, respectively. An integrated target domain dataset G consists of high-level representations of real healthy data and generated faulty representations by f_{G_i} , and the integrated target representations can be described as $\mathbf{G} = \{f_{G_i}(f_F(\mathbf{X}_{T,H}^U)); f_F(\mathbf{X}_{T,H}^U)\}$ while the source representation matrix can be described as $\mathbf{O} = \{f_F(\mathbf{X}_S^L)\}$. Given cross-domain classifier f_D , the aforementioned integrated representations \mathbf{P}_f and \mathbf{Q}_f can be represented as $\mathbf{P}_f = f_D(\mathbf{O})$ and $\mathbf{Q}_f = f_D(\mathbf{G})$, respectively.

Through the above two-stage optimisation, integrated representations in the target domain can be achieved and used for target fault diagnosis tasks.

Xie et al. [48] proposed to use cycle-consistent GAN to generate target samples for different working conditions. Given two working conditions C_X , and faulty data $X \in C_X$, $Y \in C_Y$, the goal is to find a generator function G that can map X to Y when Y is unavailable in the real monitoring scene. The diagram of cycle-consistent GAN can be represented in (11.a), and it can be exploded into (11.b) and (11.c).

$$\begin{array}{c} \mathcal{D}_X \quad \mathcal{D}_Y \quad \mathcal{D}_Y \quad \mathcal{D}_X \\ \uparrow \quad \uparrow \quad \uparrow \quad \uparrow \\ \mathbf{X} \xrightarrow{G} \mathbf{Y} \Leftrightarrow \mathbf{X} \xrightarrow{G} \mathbf{\hat{Y}} \xrightarrow{F} \mathbf{\hat{X}} + \mathbf{Y} \xrightarrow{F} \mathbf{\hat{X}} \xrightarrow{G} \mathbf{\hat{Y}} \end{array} \quad (11)$$

The generator function G and F shown in Equation (11) are inverse transformations of each other. The process $\mathbf{X} \xrightarrow{G} \mathbf{Y}$ in (11.b) aimed to learn a generator G for generated samples $\hat{\mathbf{Y}}$ in condition C_Y , and the discriminator D_Y was responsible for distinguishing real and generated fake samples in C_Y . The process $\hat{\mathbf{Y}} \xrightarrow{F} \hat{\mathbf{X}}$ was designed to avoid overfitting of the process $\mathbf{X} \xrightarrow{G} \mathbf{Y}$ and a reconstructed cycle-consistency loss \mathcal{L}_{cyc} between \mathbf{X} and $\hat{\mathbf{X}}$ was included in the objective function. However, the generator F in (11.a) never sees the true \mathbf{X} in C_X , so its generated $\hat{\mathbf{X}}$ may differ far away from the \mathbf{X} in C_X . So, an inverse reconstruction from \mathbf{Y} to $\hat{\mathbf{Y}}$ was integrated in (11.c), which constraints both generator G and F generates realistic samples in condition C_Y and C_X , respectively.

Through a cycle-consistent structure, realistic samples, which are unavailable in target working conditions can be generated from source conditions.

3.5 | Adaptive transfer

Here, we name adaptive transfer as a special transfer learning method applied in fault diagnosis since implied transfer abilities and strong robustness for new fault diagnosis and detection tasks are achieved through no transfer learning methods mentioned above are applied. Here are two perspectives: (1) adaptive structure and (2) ensemble strategy.

3.5.1 | Adaptive structure

The above-discussed transfer learning algorithms are carefully designed and validated on special datasets under different distribution conditions. Compared with those, some deep adaptive structures themselves are robust to the distribution divergence, which deserves priority when the data discrepancy is unknown.

As a representative work, Zhang et al. [41] embedded a domain adaptive structure named adaptive batch normalisation (AdaBN) [96] in their deep fault diagnosis model. As an extension of batch normalisation (BN) [97], AdaBN implemented domain adaptation through linear scale and shift transformation in each neuron. Before applying the trained model to the target task, the mean value $\mu_t^i = E[\mathbf{X}_T^{(i)}]$ and variance value $\sigma_t^i = \text{Var}[\mathbf{X}_T^{(i)}]$ of the i th neuron based on all the target feature matrix \mathbf{X}_T were calculated. Then in the application stage, a linear scale $\hat{\mathbf{X}}_t^{(i)}(p) = (\mathbf{X}_t^{(i)}(p) - \mu_t^i) / \sigma_t^i$ and shift transformation $\hat{\mathbf{Y}}_t^{(i)}(p) = \gamma^{(i)} \hat{\mathbf{X}}_t^{(i)}(p) \beta^{(i)}$ can be implemented in the BN layers. Then, $\hat{\mathbf{y}}_t(p)$ was the final output of the i th BN layer and acts as the input of the next neuron block. Here, $\gamma^{(i)}$ and $\beta^{(i)}$ were the scale and shift coefficients, respectively, which are used to maintain the

representation power of the network. AdaBN transfers sample representations into a similar distribution using scale and shift parameters in statistics. The AdaBN is also found in deep asynchronous transfer [82].

Up to date, most of the adaptive structures such as AdaBN are presented in an intuitive way and verified by experiments. For example, Jia et al. [110] and Zhang et al. [111] determined that the first convolutional layer act as a band-pass filter. A reasonable length selection of the first convolutional layer retains the useful frequency information while filtering out the redundant information. In addition, the wide convolution kernel structure filters out the low-frequency features that support the fault frequency [112]. The deep Siamese structure has been demonstrated to be effective in fault diagnosis with limited training data of variable distribution [47].

3.5.2 | Ensemble strategy















Deep ensemble learning is an effective and robust strategy to improve the transferability of a model under different data distributions. According to the *no free lunch theorem* [113, 114], there is no algorithm that can achieve the best performance on all problems. Therefore, integrating multiple models/algorithms into real machine health monitoring is essential when the working condition changes. Some classical ensemble strategies such as Bagging, Boosting and Stacking [115] have been widely applied in fault diagnosis. Different multi-sensor fusion-based deep ensemble strategies (at data/feature/decision level) are emerging in recent years [42].

Xia et al. [116] proposed a data ensemble strategy at the data level. Multiple time-series sensor signals were stacked to construct a 2-D input matrix to a deep network. The stacked data coming from different sensors reflected the machine health condition from different perspectives. Taking CNN as a feature extractor, effectively avoided the manual feature design or selection and combined all information from different sensors.

Li et al. [117] proposed a feature ensemble strategy at feature-level. They employed the inception module from GoogLeNet [104]. The inception module was used to concatenate and extract multiple features learnt by different convolutional kernels. A similar feature ensemble work based on ResNet was proposed by Wen et al. [118]. ResNet can effectively deal with gradient vanishing problem and combine general features with specific features together from different layers.

Li et al. [43] proposed a decision ensemble strategy at decision-level. Multiple CNN models with different structure parameters were designed and trained using data from different sensors. An improved D-S (IDS) evidence fusion algorithm was proposed to deal with fusion conflicts problem. In the test stage, IDS was responsible for merging all the probabilistic decisions from the pre-trained models to a final fault diagnosis decision.

TABLE 4 Open-sourced datasets for research on knowledge transfer in fault diagnosis of rotary machines (*RTF: run to failure; the dataset names come from the corresponding references or the abbreviation of institutes for brevity)

Idx.	Year	Equipment	Dataset	Application	Component	Generation	Parameter (#)	Ref.
1	2006		IMS	Prognosis	Bearing	RTF	Acceleration (2)	[98]
2	2008		C-MAPSS	Prognosis	Turbofan engine	Simulation	Systematic sensors (26)	[99]
3	2009		PHM2009	Diagnosis	Gearbox	Industrial	Acceleration (2) and speed (1)	[100]
4	2012		EEMTO-ST	Prognosis	Bearing	RTF	Acceleration (1) and temperature (1)	[101]
5	2013		MFPT	Diagnosis	Bearing	Artificial and industrial	Acceleration (1)	[102]
6	2015		CWRU	Diagnosis	Bearing	Artificial	Acceleration (2)	[103]
7	2016		KAT	Diagnosis	Bearing	Artificial and RTF	Acceleration (2), current (2), force (1), speed (1), temperature (1), and torque (1)	[104]
8	2016		RSE-BU	Diagnosis	Pump	Artificial or RTF	Acceleration (1)	[105]
9	2018		ME-UCONN	Diagnosis	Gearbox	Artificial	Acceleration (1)	[50]
10	2018		XJTU	Prognosis	Bearing	RTF	Acceleration (2)	[106]
11	2018		MAFAULDA	Diagnosis	Bearing	Artificial	Acceleration (6), sound-wave (1), and speed (1)	[107]
12	2018		MFS-PK5M	Diagnosis	Bearing	Artificial	Acceleration (1) and speed (1)	[108]
13	2019		DIGR	Prognosis diagnosis	Bearing	Artificial and RTF	Acceleration (2)	[109]
14	2019		SEU	Diagnosis	Bearing gearbox	Artificial	Acceleration (7) and torque (1)	[51]

4 | CHALLENGES AND FUTURE RESEARCH DIRECTIONS

4.1 | Challenges

4.1.1 | Data issues

The complexity of the data issues makes it difficult to apply transfer learning in many real scenarios. Most of current transfer learning research efforts in fault diagnosis of rotary machines are from laboratory equipment (see Table 4) while few are tested in real applications. Under experimental environments, significant and limited faulty data can be artificially designed and easily selected; however, unlabelled and imbalanced attributes widely exist in historical data in real scenarios. Faults in actual production rarely occur, and it is time consuming and expensive to record balanced faulty data from real applications. It is also a challenging task to select useful data for training a promising diagnosis model from massive unlabelled database, which may contain giga byte, tera byte, peta byte or even more memorial units. In addition, there are issues with missing data and wrong data in almost all industrial monitoring systems, which may reduce the performance of diagnosis models.

4.1.2 | Modelling matters

There is no unified learning strategy to solve all transfer learning problems in fault diagnosis, and different cases should

be analysed specifically. Expert knowledge-based algorithms show the mechanism behind the failure. Machine learning-based algorithms combine expert knowledge with robust classification abilities. Deep learning-based algorithms help researchers save time to extract useful features in a complex big-data context. Transfer learning relieves the requirement of massive essential task data and leverages knowledge from a source domain to a target domain. Although the automated learning capabilities of machines have increased over the past three decades, it still requires the involvement of human beings in algorithm design, model selection, and parameter tuning for a promising fault diagnosis performance in real scenarios. Further, the new data distribution discrepancy emerges quickly in complex working environments. Many time-consuming deep transfer learning algorithms have to be re-trained from scratch and can hardly meet the responsive requirements. Investing the computation complexity of the knowledge transfer algorithms and building an online fault diagnosis model are essential.

4.1.3 | Application problems

A promising transfer strategy in fault diagnosis should consider application attributes such as real-time, reliability, interpretability and explainability. Due to the deeper and complex structure with more parameters, some deep-based transfer algorithms may not meet extreme real-time requirements like monitoring the health condition of lithography. Many of the

current transfer learning (TL) research efforts take accuracy as the only criterion to compare the advantage while ignoring the false-negative error and so on, which may lead to catastrophic losses in processing. Interpretability bridges the cause and effect while explainability helps researchers to understand the internal faulty mechanics of a machine or a model, which are quite essential and challenging in the current deep learning era. In addition, the selection and deployment of transfer learning in fault diagnosis should combine soft algorithms with hardware implementation.

4.2 | Future research directions

4.2.1 | Digital twin-based enhancement

Digital twin is a general concept that maps a physical space into a virtual space. Monitoring of data by GANs may allow to deal with unlabelled and imbalanced data in a physical space. Analysis of the noise mechanisms in virtual models allows to reduce the noise in a physical model. In general, a virtual model can be employed to provide generated data to assist in training of a diagnosis model under real conditions while a physical model can return the feedback of its real conditions to iteratively fine-tune the virtual model.

4.2.2 | Automated incremental learning

It is a promising trend to Automation of all aspects of incremental transfer learning in health management, which is of great interest. Current automated machine learning (Auto-ML) methods have surpassed traditional machine learning designed by human experts in several tasks [119]. In the last few years, Auto-ML has achieved great success in both theoretical research and commercial applications [120]. Several general techniques (such as meta-learning, neural architecture search, and hyperparameter optimisation) have laid a foundation of Auto-ML frameworks such as AutoKeras (Google), Azure Machine Learning (Microsoft), and AutoWeka (the University of Waikato). AutoML can also help to implement high-quality transfer learning through auto algorithms selection, hyperparameters tuning, model optimisation, which reduces the requirement for solid knowledge through trials and errors. At the same time, transfer-based fault diagnosis models should be updated according to the emerging series data. An efficient incremental learning (IL) paradigm learns and updates models timely from new data streams while keeping the health information from the old. Automated incremental learning (AIL), which combines AutoML and IL, could be of great interest to transfer-based fault diagnosis of rotary machines in the future.

4.2.3 | Algorithm-chip integration

Dedicated chips integrated with TL algorithms applied in fault diagnosis could be developed in the future. Besides accuracy,

software and hardware should be integrated to satisfy specific fault diagnosis requirements such as real-time, energy saving, reliability and so on. Most of current TL algorithms are trained and deployed on general chips such as GPU/TPU/FPGA and so on, which are time consuming in big-data context. In recent years, hardware based on optical components [121] or memristor [122] has accelerated training of models. In addition to the improved computing power, reduced power consumption and lower hardware costs have been realised. Special edge transfer algorithms and edge computation chips could be the next advances in fault diagnosis of rotary machines.

4.2.4 | Interpretability and explainability

Interpretability and explainability are essential in understanding the mechanisms behind the black box effects of complex transfer learning algorithms. Current transfer learning in fault diagnosis is expected to combine faulty mechanisms with deep learning to enhance interpretability and explainability of the models and results. To bridge the gap, four research algorithms can be considered: (a) integrating interpretable algorithms: integrate rigorous algorithms such as logistic regression, decision tree, and k-nearest neighbour to justify the decision made by a TL algorithm; (b) employ surrogate algorithms: a surrogate algorithm is an interpretable algorithm that is trained to approximate a black-box model, where the interpretability is implemented by training multiple interpretable machines to learn an incomprehensible one; (c) data visualisation: visualisation techniques such as maximising the activation and feature inversion have been researched to offer better understanding of what a model learns, and they may have a potential to interpret how a transfer learning algorithm works; (d) adversarial example-based explanation: humans learn from mistakes. Adversarial examples have a potential to provide a better understanding of why an algorithm makes wrong predictions and indirectly enhance interpretability.

4.2.5 | Open source dataset and platform

Open source dataset and platform: Data-driven intelligent fault diagnosis contains four key elements: data, expertise, algorithms, and computation resources. Most of the current research studies use simulated data to conduct benchmarking studies and lack expertise in combining physical modelling with data-driven algorithms. Further, the undisclosed algorithms and limited computation resources hinder the development of their industrial applications in big-data environments. Thus, high-quality industrial failure data are in need to validate the advanced KTFD algorithms. Detailed systematic parameters beyond sensor data are in need to drive theoretical developments. A platform is in need to collect, manage and share open-source datasets and relevant repositories. Researchers are encouraged to share and compare their developed algorithms under unified resources. Benchmark studies with different datasets and algorithms are

suggested to avoid repetitive work. Special cases, such as noisy data, missing data, and continuous varying working conditions in real applications deserve consideration in an industrial benchmarking platform.

5 | CONCLUSION

Knowledge transfer has received attention in fault diagnosis and prognostics. This paper offered a systematic review of knowledge transfer in fault diagnosis of rotary machines. A problem-oriented taxonomy, with 5 scenarios and 13 transfer tasks, was proposed. Based on the data and model levels, knowledge transfer was classified into four paradigms: statistics transfer, deep transfer, adversarial transfer, and adaptive transfer. The advanced transfer algorithms and representative research in fault diagnosis were introduced in each category. Future challenges and research directions were explored from data, learning and application perspectives.

ACKNOWLEDGEMENTS

This work was supported in part by the China Scholarship Council (CSC) during a research visit of Guokai Liu to the University of Iowa under Grant 201906160078 and by the Fundamental Research Funds for the Central Universities of China (Grant Number: 2021GCRC058).

CONFLICT OF INTEREST STATEMENT

No conflict of interest exists in the submission of this manuscript, and it is approved by all authors for publication.

DATA AVAILABILITY STATEMENT

Data derived from public domain resources.

ORCID

Guokai Liu  <https://orcid.org/0000-0002-6746-7137>

Weiming Shen  <https://orcid.org/0000-0001-5204-7992>

Liang Gao  <https://orcid.org/0000-0002-1485-0722>

Andrew Kusiak  <https://orcid.org/0000-0003-4393-1385>

REFERENCES

- Frank, P.M.: Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: a survey and some new results. *Automatica*. 26(3), 459–474 (1990)
- Liao, S.-H.: Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert Syst. Appl.* 28(1), 93–103 (2005)
- Yang, Z.-L., et al.: Expert system of fault diagnosis for gear box in wind turbine. *Syst. Eng. Precedia*. 4, 189–195 (2012)
- Cui, H., et al.: A novel advancing signal processing method based on coupled multi-stable stochastic resonance for fault detection. *Appl. Sci.* 11(12), 5385 (2021)
- Liu, R., et al.: Artificial intelligence for fault diagnosis of rotating machinery: a review. *Mech. Syst. Signal Process.* 108, 33–47 (2018)
- Wuest, T., et al.: Machine learning in manufacturing: advantages, challenges, and applications. *Prod. Manuf. Res.* 4(1), 23–45 (2016)
- Lei, Y., et al.: Applications of machine learning to machine fault diagnosis: a review and roadmap. *Mech. Syst. Signal Process.* 138, 106587 (2020)
- Jia, F., et al.: Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* 72, 303–315 (2016)
- Zhang, S., et al.: Deep learning algorithms for bearing fault diagnostics—a comprehensive review. *IEEE Access*. 8, 29857–29881 (2020)
- Wang, J., et al.: Deep learning for smart manufacturing: methods and applications. *J. Manuf. Syst.* 48, 144–156 (2018)
- Yan, R., et al.: Knowledge transfer for rotary machine fault diagnosis. *IEEE Sensor. J.* 20(15), 8374–8393 (2019)
- Zheng, H., et al.: Cross-domain fault diagnosis using knowledge transfer strategy: a review. *IEEE Access*. 7, 129260–129290 (2019)
- Li, W., et al.: A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: theories, applications and challenges. *Mech. Syst. Signal Process.* 167, 108487 (2022)
- Muralidharan, V., Sugumaran, V.: A comparative study of naïve bayes classifier and bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis. *Appl. Soft Comput.* 12(8), 2023–2029 (2012)
- Cerrada, M., et al.: Fault diagnosis in spur gears based on genetic algorithm and random forest. *Mech. Syst. Signal Process.* 70, 87–103 (2016)
- Widodo, A., Yang, B.-S.: Support vector machine in machine condition monitoring and fault diagnosis. *Mech. Syst. Signal Process.* 21(6), 2560–2574 (2007)
- Li, G., et al.: Review on fault detection and diagnosis feature engineering in building heating, ventilation, air conditioning and refrigeration systems. *IEEE Access*. 9, 2153–2187 (2020)
- Zhang, Z.-H., et al.: Tri-partition state alphabet-based sequential pattern for multivariate time series. *Cognit. Comput.* 1–19 (2021)
- Deng, W., et al.: An enhanced fast non-dominated solution sorting genetic algorithm for multi-objective problems. *Inf. Sci.* 585, 441–453 (2022)
- Ran, X., et al.: A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots. *Appl. Sci.* 11(23), 11202 (2021)
- Kusiak, A.: Smart manufacturing must embrace big data. *Nature*. 544(7648), 23–25 (2017)
- Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105 (2012)
- Guo, X., Chen, L., Shen, C.: Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis. *Measurement*. 93, 490–502 (2016)
- Turker Ince, T., et al.: Real-time motor fault detection by 1-d convolutional neural networks. *IEEE Trans. Ind. Electron.* 63(11), 7067–7075 (2016)
- Guo, L., et al.: A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*. 240, 98–109 (2017)
- Han, L., et al.: Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. *ISA Trans.* 77, 167–178 (2018)
- Zhao, R., et al.: Machine health monitoring with lstm networks. In: 2016 10th international conference on sensing technology (ICST), pp. 1–6. IEEE (2016)
- Lei, J., Liu, C., Jiang, D.: Fault diagnosis of wind turbine based on long short-term memory networks. *Renew. Energy*. 133, 422–432 (2019)
- Shao, S., Wang, P., Yan, R.: Generative adversarial networks for data augmentation in machine fault diagnosis. *Comput. Ind.* 106, 85–93 (2019)
- Wang, Z., Wang, J., Wang, Y.: An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition. *Neurocomputing*. 310, 213–222 (2018)
- Li, X., et al.: A threshold-control generative adversarial network method for intelligent fault diagnosis. *Complex Syst. Modeling Simulation*. 1(1), 55–64 (2021)
- Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22(10), 1345–1359 (2009)
- Weiss, K., Khoshgoftaar, T.M., Wang, D.D.: A survey of transfer learning. *J. Big Data*. 3(1), 1–40 (2016)
- Zhuang, F., et al.: A comprehensive survey on transfer learning. *Proc. IEEE*. 109(1), 43–76 (2020)

35. Zhang, J., et al.: Recent advances in transfer learning for cross-dataset visual recognition: a problem-oriented perspective. *ACM Comput. Surv.* 52(1), 1–38 (2019)
36. Wenyuan, Y.Q.D., Xue, G., Yu, Y.: Boosting for transfer learning. In: *Proceedings of the 24th international conference on machine learning*, pp. 193–200. Corvallis (2007)
37. Yao, Y., Doretto, G.: Boosting for transfer learning with multiple sources. In: *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1855–1862. IEEE (2010)
38. Yang, B., et al.: An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings. *Mech. Syst. Signal Process.* 122, 692–706 (2019)
39. Wen, L., et al.: A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Trans. Ind. Electron.* 65(7), 5990–5998 (2017)
40. Wen, L., Gao, L., Li, X.: A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Transac. syst. man, cybernetics: Syst.* 49(1), 136–144 (2017)
41. Zhang, W., et al.: A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors.* 17(2), 425 (2017)
42. Jing, L., et al.: An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox. *Sensors.* 17(2), 414 (2017)
43. Li, S., et al.: An ensemble deep convolutional neural network model with improved ds evidence fusion for bearing fault diagnosis. *Sensors.* 17(8), 1729 (2017)
44. Chen, D., Yang, S., Zhou, F.: Transfer learning based fault diagnosis with missing data due to multi-rate sampling. *Sensors.* 19(8), 1826 (2019)
45. Wang, C., et al.: Heterogeneous transfer learning based on stack sparse auto-encoders for fault diagnosis. In: *2018 Chinese automation congress (CAC)*, pp. 4277–4281. IEEE (2018)
46. Zhang, R., et al.: Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. *IEEE Access.* 5, 14347–14357 (2017)
47. Zhang, A., et al.: Limited data rolling bearing fault diagnosis with few-shot learning. *IEEE Access.* 7, 110895–110904 (2019)
48. Yuan, X., Zhang, T.: A transfer learning strategy for rotation machinery fault diagnosis based on cycle-consistent generative adversarial networks. In: *2018 Chinese automation congress (CAC)*, pp. 1309–1313. IEEE (2018)
49. Wang, T., et al.: A self-learning fault diagnosis strategy based on multi-model fusion. *Information.* 10(3), 116 (2019)
50. Cao, P., Zhang, S., Tang, J.: Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning. *IEEE Access.* 6, 26241–26253 (2018)
51. Shao, S., et al.: Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Trans. Ind. Inf.* 15(4), 2446–2455 (2018)
52. Xu, Y., et al.: A digital-twin-assisted fault diagnosis using deep transfer learning. *IEEE Access.* 7, 19990–19999 (2019)
53. Zhang, Y., Jombo, G., Latimer, A.: A knowledge transfer platform for fault diagnosis of industrial gas turbines. In: *2018 IEEE 22nd international conference on intelligent engineering systems (INES)*, pp. 000347–000352. IEEE (2018)
54. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55(1), 119–139 (1997)
55. Shen, F., et al.: Bearing fault diagnosis based on svd feature extraction and transfer learning classification. In: *2015 prognostics and system health management conference (PHM)*, pp. 1–6. IEEE (2015)
56. Xiao, D., et al.: Transfer learning with convolutional neural networks for small sample size problem in machinery fault diagnosis. *Proc. IME C J. Mech. Eng. Sci.* 233(14), 5131–5143 (2019)
57. Pan, S.J., et al.: Domain adaptation via transfer component analysis. *IEEE Trans. Neural Network.* 22(2), 199–210 (2010)
58. Xie, J., et al.: On cross-domain feature fusion in gearbox fault diagnosis under various operating conditions based on transfer component analysis. In: *2016 IEEE international conference on prognostics and health management (ICPHM)*, pp. 1–6. IEEE (2016)
59. Wang, J., et al.: Stratified transfer learning for cross-domain activity recognition. In: *2018 IEEE international conference on pervasive computing and communications (PerCom)*, pp. 1–10. IEEE (2018)
60. Yu, X., et al.: Conditional adversarial domain adaptation with discrimination embedding for locomotive fault diagnosis. *IEEE Trans. Instrum. Meas.* 70(1–12) (2020)
61. Long, M., et al.: Transfer feature learning with joint distribution adaptation. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2200–2207 (2013)
62. Wang, J., et al.: Balanced distribution adaptation for transfer learning. In: *2017 IEEE international conference on data mining (ICDM)*, pp. 1129–1134. IEEE (2017)
63. Guo, J., et al.: Fault diagnosis of delta 3d printers using transfer support vector machine with attitude signals. *IEEE Access.* 7, 40359–40368 (2019)
64. Tong, Z., et al.: Bearing fault diagnosis under variable working conditions based on domain adaptation using feature transfer learning. *IEEE Access.* 6, 76187–76197 (2018)
65. Chen, C., et al.: A cross domain feature extraction method based on transfer component analysis for rolling bearing fault diagnosis. In: *2017 29th Chinese control and decision conference (CCDC)*, pp. 5622–5626. IEEE (2017)
66. Qian, W., et al.: A novel transfer learning method for robust fault diagnosis of rotating machines under variable working conditions. *Measurement.* 138, 514–525 (2019)
67. Han, T., et al.: Deep transfer network with joint distribution adaptation: a new intelligent fault diagnosis framework for industry application. *ISA Trans.* 97, 269–281 (2020)
68. Cao, N., et al.: Bearing state recognition method based on transfer learning under different working conditions. *Sensors.* 20(1), 234 (2020)
69. Gu, J., Wang, Y.: A cross domain feature extraction method for bearing fault diagnosis based on balanced distribution adaptation. In: *2019 prognostics and system health management conference (PHM-Qingdao)*, pp. 1–5. IEEE (2019)
70. Fernando, B., et al.: Unsupervised visual domain adaptation using subspace alignment. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2960–2967 (2013)
71. Gong, B., et al.: Geodesic flow kernel for unsupervised domain adaptation. In: *2012 IEEE conference on computer vision and pattern recognition*, pp. 2066–2073. IEEE (2012)
72. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: an unsupervised approach. In: *2011 international conference on computer vision*, pp. 999–1006. IEEE (2011)
73. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. *Proc. AAAI Conf. Artif. Intell.* 30 (2016)
74. Zheng, H., et al.: Intelligent fault identification based on multisource domain generalization towards actual diagnosis scenario. *IEEE Trans. Ind. Electron.* 67(2), 1293–1304 (2019)
75. Wang, X., He, H., Li, L.: A hierarchical deep domain adaptation approach for fault diagnosis of power plant thermal system. *IEEE Trans. Ind. Inf.* 15(9), 5139–5148 (2019)
76. Kim, H., Youn, B.D.: A new parameter repurposing method for parameter transfer with small dataset and its application in fault diagnosis of rolling element bearings. *IEEE Access.* 7, 46917–46930 (2019)
77. Hasan, M.J., Islam, M.M.M., Kim, J.-M.: Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. *Measurement.* 138, 620–631 (2019)
78. Xu, G., et al.: Online fault diagnosis method based on transfer convolutional neural networks. *IEEE Trans. Instrum. Meas.* 69(2), 509–520 (2019)
79. Liu, Q., Huang, C.: A fault diagnosis method based on transfer convolutional neural networks. *IEEE Access.* 7, 171423–171430 (2019)
80. Lu, W., et al.: Deep model based domain adaptation for fault diagnosis. *IEEE Trans. Ind. Electron.* 64(3), 2296–2305 (2016)
81. Duan, L., et al.: Auxiliary-model-based domain adaptation for reciprocating compressor diagnosis under variable conditions. *J. Intell. Fuzzy Syst.* 34(6), 3595–3604 (2018)

82. Qian, W., et al.: A new deep transfer learning network for fault diagnosis of rotating machine under variable working conditions. In: 2018 prognostics and system health management conference (PHM-Chongqing), pp. 1010–1016. IEEE (2018)
83. Chen, D., Yang, S., Zhou, F.: Incipient fault diagnosis based on dnn with transfer learning. In: 2018 international conference on control, automation and information sciences (ICCAIS), pp. 303–308. IEEE (2018)
84. Gao, Y., et al.: A zero-shot learning method for fault diagnosis under unknown working loads. *J. Intell. Manuf.* 31(4), 899–909 (2020)
85. Qian, W., Li, S., Wang, J.: A new transfer learning method and its application on rotating machine fault diagnosis under variant working conditions. *IEEE Access.* 6, 69907–69917 (2018)
86. Xiao, D., et al.: Domain adaptive motor fault diagnosis using deep transfer learning. *IEEE Access.* 7, 80937–80949 (2019)
87. Xiang, L., et al.: Multi-layer domain adaptation method for rolling bearing fault diagnosis. *Signal Process.* 157, 180–197 (2019)
88. An, Z., et al.: Generalization of deep neural network for bearing fault diagnosis under different working conditions using multiple kernel method. *Neurocomputing.* 352, 42–53 (2019)
89. Wang, K., Wu, B.: Power equipment fault diagnosis model based on deep transfer learning with balanced distribution adaptation. In: International conference on advanced data mining and applications, pp. 178–188. Springer (2018)
90. Yang, B., et al.: A transfer learning method for intelligent fault diagnosis from laboratory machines to real-case machines. In: 2018 international conference on sensing, diagnostics, prognostics, and control (SDPC), pp. 35–40. IEEE (2018)
91. Xiang, L., Zhang, W., Ding, Q.: A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning. *Neurocomputing.* 310, 77–95 (2018)
92. Guo, L., et al.: Deep convolutional transfer learning network: a new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Trans. Ind. Electron.* 66(9), 7316–7325 (2018)
93. Han, T., et al.: A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. *Knowl. Base Syst.* 165, 474–487 (2019)
94. Zhang, M., et al.: A deep transfer model with wasserstein distance guided multi-adversarial networks for bearing fault diagnosis under different working conditions. *IEEE Access.* 7, 65303–65318 (2019)
95. Zhang, B., et al.: Adversarial adaptive 1-d convolutional neural networks for bearing fault diagnosis under varying working condition. *arXiv preprint arXiv:1805.00778* (2018)
96. Wang, Q., Gabriel, M., Fink, O.: Domain adaptive transfer learning for fault diagnosis. In: 2019 prognostics and system health management conference (PHM-Paris), pp. 279–285. IEEE (2019)
97. Xiang, L., Zhang, W., Ding, Q.: Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks. *IEEE Trans. Ind. Electron.* 66(7), 5525–5534 (2018)
98. Qiu, H., et al.: Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *J. Sound Vib.* 289(4–5), 1066–1090 (2006)
99. Saxena, A., et al.: Damage propagation modeling for aircraft engine run-to-failure simulation. In: 2008 international conference on prognostics and health management, pp. 1–9. IEEE (2008)
100. Zhang, B., et al.: Intelligent fault diagnosis under varying working conditions based on domain adaptive convolutional neural networks. *IEEE Access.* 6, 66367–66384 (2018)
101. Nectoux, P., et al.: Pronostia: an experimental platform for bearings accelerated degradation tests. In: IEEE international conference on prognostics and health management, PHM'12, pp. 1–8. IEEE Catalog Number: CPF12PHM-CDR (2012)
102. Bechhoefer, E.: A quick introduction to bearing envelope analysis. *Green Power Monitor System* (2016)
103. Smith, W.A., Randall, R.B.: Rolling element bearing diagnostics using the case western reserve university data: a benchmark study. *Mech. Syst. Signal Process.* 64, 100–131 (2015)
104. Lessmeier, C., et al.: Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: a benchmark data set for data-driven classification. *PHM Soc. European Conf.* 3 (2016)
105. Lu, C., et al.: Fault diagnosis for rotating machinery: a method based on image processing. *PLoS One.* 11(10), e0164111 (2016)
106. Wang, B., et al.: A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Trans. Reliab.* 69(1), 401–412 (2018)
107. Marins, M.A., et al.: Improved similarity-based modeling for the classification of rotating-machine failures. *J. Franklin Inst.* 355(4), 1913–1930 (2018)
108. Huang, H., Baddour, N.: Bearing vibration data collected under time-varying rotational speed conditions. *Data Brief.* 21, 1745–1749 (2018)
109. Daga, A.P., et al.: The politecnico di torino rolling bearing test rig: description and analysis of open access data. *Mech. Syst. Signal Process.* 120, 252–273 (2019)
110. Jia, F., et al.: Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mech. Syst. Signal Process.* 110, 349–367 (2018)
111. Zhang, W., et al.: A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mech. Syst. Signal Process.* 100, 439–453 (2018)
112. Peng, D., et al.: A novel deeper one-dimensional cnn with residual learning for fault diagnosis of wheelset bearings in high-speed trains. *IEEE Access.* 7, 10278–10293 (2018)
113. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. *Neural Comput.* 8(7), 1341–1390 (1996)
114. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1(1), 67–82 (1997)
115. Zhou, Z.-H.: Ensemble methods: foundations and algorithms. Chapman and Hall/CRC (2019)
116. Xia, M., et al.: Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks. *IEEE/ASME Trans. Mechatron.* 23(1), 101–110 (2017)
117. Li, J., Qu, W.: Aero-engine sensor fault diagnosis based on convolutional neural network. In: 2018 37th Chinese control conference (CCC), pp. 6049–6054. IEEE (2018)
118. Wen, L., Li, X., Gao, L.: A transfer convolutional neural network for fault diagnosis based on resnet-50. *Neural Comput. Appl.* 32(10) (2020)
119. Hutter, F., Kotthoff, L., Vanschoren, J.: Automated machine learning: methods, systems, challenges. Springer Nature (2019)
120. Elshawi, R., Maher, M., Sakr, S.: Automated machine learning: state-of-the-art and open challenges. *arXiv preprint arXiv:1906.02287* (2019)
121. Xing, L., et al.: All-optical machine learning using diffractive deep neural networks. *Science.* 361(6406), 1004–1008 (2018)
122. Yao, P., et al.: Fully hardware-implemented memristor convolutional neural network. *Nature.* 577(7792), 641–646 (2020)

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Liu, G., et al.: Knowledge transfer in fault diagnosis of rotary machines. *IET Collab. Intell. Manuf.* 4(1), 17–34 (2022). <https://doi.org/10.1049/cim2.12047>