Deep Multi-adversarial Conditional Domain Adaptation Networks for Fault Diagnostics of Industrial Equipment

Bingsen Wang, Piero Baraldi, and Enrico Zio, *Senior Member, IEEE*

Abstract

Deep learning methods of fault diagnostics require the availability of a large amount of labeled data for training, i.e., signal values corresponding to known degradation and fault states. Furthermore, the distribution of the training data should be similar to that of the (test) data collected in the field. Since these conditions are typically not satisfied in most industrial applications, this work develops a deep multi-adversarial conditional domain adaptation network. The main original contribution lies in a novel method to align, class by class, the weighted marginal data distributions using multiple domain discriminators. The network allows overtaking the classification underperformance caused by the problem of negative transfer, which is typically encountered when only few training data of some of the classes are available. The proposed method is shown to outperform other state-of-the-art methods on two cross-domain fault diagnostic case studies, verified by applying Friedman and Holm post-hoc tests.

Manuscript received February 04, 2022; revised July 24, 2022 and October 05, 2022; accepted October 26, 2022. This project has received funding from the Shift2Rail Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No 101015423. The information in this document is provided "as is", and no guarantee or warranty is given that the information is fit for any particular purpose. The content of this document reflects only the author's view – the Shift2Rail Joint Undertaking is not responsible for any use that may be made of the information it contains. The users use the information at their sole risk and liability. The work of B. Wang was supported by the China Scholarship Council under Grant 201801810028. Paper no. TII-22-0580.

Corresponding author: Piero Baraldi.

B. Wang and P. Baraldi are with the Department of Energy, Politecnico di Milano, Milan 20156, Italy (e-mail: bingsen.wang@polimi.it; piero.baraldi@polimi.it).

E. Zio is with the Centre de Recherche sur les Risques et les Crises (CRC), Mines ParisT-PSL, Sophia Antipolis 06904, France (e-mail: enrico.zio@mines-paristech.fr), and also with the Department of Energy, Politecnico di Milano, Milan 20156, Italy (e-mail: enrico.zio@polimi.it).

Index Terms

Fault diagnostics, deep transfer learning, multi-adversarial domain adaptation, conditional domain adaptation, distribution alignment, negative transfer, fleet of machines, bearings, railway industry, automatic doors.

I. Introduction

Within Prognostics and Health Management (PHM), fault diagnostics is the task of identifying the degradation or fault state of an industrial component using related signal measurements taken from the monitoring system [1]. Analytics based on Artificial Intelligence (AI) techniques have shown great potential in extracting component condition information from massive monitoring data [2]. Conventional AI algorithms, such as Artificial Neural Network (ANN), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN), typically require knowledge on signal processing and expertise on component degradation to manually extract and select features for the specific diagnostic task [3]. Differently, deep learning-based algorithms, such as Deep Belief Network (DBN) [4], Sparse Auto-Encoder (SAE) [5] and Convolutional Neural Network (CNN) [6], automatically extract degradation features from large-scale raw data, given their capability of mining hierarchical representations and, therefore, of self-adaptive feature learning [7]. For example, Shao *et al.* [4] developed a method based on Particle Swarm Optimization (PSO) to optimize the architecture and the hyper-parameters of a DBN employed for fault diagnostics of rolling bearings. Sun *et al.* [5] developed a deep neural network to classify induction motor faults, in which SAEs are used to learn feature representations from raw data. Wen *et al.* [6] developed a Two-Dimensional (2D) CNN which converts raw time-series signals into 2D images, and applied it to fault diagnostics of motor bearings and pumps.

Training deep learning models for fault diagnostics requires the availability of a large amount of data representative of the system behavior, i.e., signal values measured during operation in correspondence of known degradation states and types of faults. Also, the distribution of the data used to train the model should, ideally, be the same of the (test) data which the model is applied to in the field, for successful fault diagnostics in practice [8]. However, in practical engineering, the ground-truth degradation and fault state of in-service components is typically not known, and only few data labeled with the degradation and fault type are available for training [2].

An additional challenge of deep learning-based fault diagnostic models comes from the divergence between the distributions of the training and test data, which can be caused by:

a) the occurrence of concept drifts due to unpredictable modifications of the working conditions caused by various possible factors, such as seasonality effects, sensor or component aging, and changes of operating conditions [9].

b) the fact that the data for training the fault diagnostic model are obtained by performing lab experiments or generated by simulation, and not from the in-service components and systems operating in the field; this is typical of the situation when there is lack of fault data from the field, e.g., for safety-critical, high-value or new-design components and systems.

c) the use of data collected by monitoring components to train the model which is, then, applied to data from another component of the same or similar type, but which could very well behave (degrade and fail) differently because the fact that experiencing different operating conditions and maintenance during its life. This situation typically occurs for components used in a fleet of machines.

The consequence of the divergence of the training and test data distributions is that the fault diagnostic model developed using the training data characteristic of a certain context (source domain) and characterized by a given (source) distribution may provide unsatisfactory results when applied to test data from another context (target domain) characterized by a different (target) distribution.

In this respect, Transfer Learning (TL) is a promising approach to address cross-domain learning problems by leveraging knowledge from the source domain to improve the performance of the model in the target domain [10]. Here, we focus on Domain Adaptation (DA), a special case of TL in which the learning tasks in the source and target domains are the same [10]. A literature review of DA methods and their application to PHM is reported in Subsection I-A. Most of the existing works about cross-domain fault diagnostics consider the case in which the discrepancies between the source and target domains are caused by modifications of the operating conditions of a same machine. The applications usually concern components of rotating machinery, such as bearings and gearboxes [8]. Few studies focus on negative transfer, i.e., a misalignment of the class-conditional data distributions, which reduces the accuracy of the classifier in the target domain [10]. This problem is common in fault diagnostic applications characterized by imbalanced and multimodal datasets [11].

We propose a novel cross-domain fault diagnostic method based on deep TL. A deep multi-adversarial conditional DA network is developed for fault diagnostics of a fleet of machines. We realistically assume that signal values are available and the corresponding degradation and fault labels are known for a machine of a fleet (source domain), but unknown for other similar machines of the same fleet (target domain). The shared feature representations between the source and target domains are extracted from the temporal signals by a CNN. Then, a task-specific classifier of the degradation state is trained to minimize the classification loss on the source domain using the known degradation labels. Jointly, an ensemble of domain discriminators is adversarially trained over the feature extractor to minimize the divergence between the distributions of the source and target domain patterns of each class in the space of the extracted features. As a result, a feature representation common to the source and target domains is obtained. Differently from the traditional adversarial learning approach, which employs a domain discriminator based on a single network [12], we use an ensemble of domain discriminators to reduce the effect of negative transfer. Since each discriminator is responsible for the alignment of the conditional distributions of a specific class, the divergence among the distributions is expected to reduce.

The main contributions that stand out from the work are:

1) a novel cross-domain fault diagnostic method for the classification of the degradation and fault states of identical components used in a fleet of machines. Specifically, a domain-invariant feature representation across different components is extracted resorting to an adversarial learning process between a CNN-based feature extractor and an ensemble of domain discriminators.

2) an effective solution for the negative transfer problem based on the use of an ensemble of domain discriminators.

The proposed method is verified on two cross-domain fault diagnostic case studies from different industrial sectors. The former considers two bearing datasets collected from different experimental platforms, whereas the latter considers signals recorded from automatic doors used in a fleet of high-speed trains.

The rest of the paper is organized as follows. Subsection I-A reviews the main approaches to DA and discusses their application in the PHM field. Section II presents the problem statement and Section III its formulation. The proposed cross-domain fault diagnostic method is described in Section IV. Section V shows the applications of the proposed method for the cross-domain fault diagnostics of bearings of rotating machinery (case study 1) and of automatic doors used in railway industry (case study 2). Finally, conclusions are drawn in Section VI.

A. DA approaches and their application to PHM

Early approaches for DA were based on instance-transfer strategies, which reweigh or subsample groups of instances from the source domain to match the distribution of the data in the target domain [13]. Other methods seek a transformation of the feature space for mapping the source distribution into the target one [14]. Discrepancy-based methods aim at learning a domain-invariant feature representation by minimizing pre-defined metrics of the distance between the distributions of the two domains. Maximum Mean Discrepancy (MMD) [15] and Wasserstein distance [16] have been used to this purpose. Deep TL methods combine deep learning to extract features for an abstract, high-level representation of raw data with TL to perform a learning task on different, but related, data distributions. The Deep Domain Confusion (DDC) method for cross-domain classification introduces a MMD-based adaptation layer into a CNN [15]. Recently, inspired by Generative Adversarial Networks (GANs), a Domain-adversarial Neural Network (DANN) has been proposed, which builds a set of features characterized by similar data distributions in the source and target domains via an adversarial learning process [12]. Instead of minimizing the Kullback-Leibler (KL) or the Jensen-Shannon (JS) divergence metrics between source and target domains, as typically done in adversarial adaptation, the Wasserstein distance is estimated by a domain critic and, then, minimized by updating a feature extractor [16].

Given that labeled data are often unavailable in the target domain, most unsupervised DA approaches align only the global distributions in the feature spaces of the source and target domains (marginal DA), without considering the complex multimodal structures underlying the data distributions [17]. As a result, a misalignment between the source and target domain data in the space of the extracted features can cause misclassifications in the target domain. This problem, which is referred to as negative learning [10], is more common when the number of observations of some classes is significantly smaller than the number of observations of other classes (imbalanced classification problem), and, therefore, the alignment of the global distributions does not properly consider the under-represented class. In this respect, it has been shown that fine-grained alignments of features extracted from different domains can yield better performance in many TL tasks [17]. Several methods attempt to eliminate the effect of negative transfer by jointly aligning marginal and conditional distributions. Outcomes of the classifier have been used for the conditional adaptation of the feature representations via the multiplicative concatenation [18]. Alignments of the marginal and class-conditional distributions have been performed by introducing a joint predictor, which learns a discriminative model of class and domain labels [19]. A Batch Spectral Penalization (BSP) approach, which penalizes the eigenvectors that are associated to the largest singular values in the space of the features extracted by DANN, has been developed to enhance the feature discriminability [20].

TL methods have already been applied to PHM, especially for fault diagnostics [21]. Transfer strategies have been introduced to improve the performance of fault diagnostic models when applied to machinery under variable rotating speeds and loads [22]. Deep TL has been used for fault diagnostics of motors, gearboxes and shaft bearings [21]. CNN and TL have been combined for bearing fault diagnostics under variable working conditions [23]. In [24], the MMD between the latent common features extracted by a three-layer SAE from source and target domains has been minimized to achieve satisfactory fault diagnostic results on the bearing dataset of Case Western Reserve University (CWRU). A DANN-based method for cross-domain fault diagnostics has been developed and applied to the CWRU bearing dataset considering variations of the working load [25].

II. Problem statement

We have available n_S patterns collected during the operation of a machine of a fleet. Each pattern $\boldsymbol{x}_{i_S} \in \mathbb{R}^{m\times L}$, $i_S = 1, 2, ..., n_S$, is a multidimensional time series constituted by the values of m signals measured during a time window of L consecutive time instants. The degradation/fault state of the component at the time in which $\boldsymbol{x}_{i_S}$ is acquired is represented by the label $y_{i_S} = \{0, 1, ..., K-1\}$. Label 0 indicates an healthy state, whereas the remaining labels $1, \ldots, K-1$ indicate degraded or fault states which, depending on the diagnostic application can differ for the type of degradation mechanism, the amount of degradation or the failure mode. $X_S = \{\boldsymbol{x}_{i_S}, i_S = 1, 2, ..., n_S\}$ and $Y_S = \{y_{i_S}, i_S = 1, 2, ..., n_S\}$ represent the sets of collected signal values and corresponding classes, respectively. We assume that the time interval during which the time series $\boldsymbol{x}_{i_S}$ is acquired is much smaller than the time scale of the degradation process, and, therefore, the degradation state of the component is not varying during the acquisition of $\boldsymbol{x}_{i_S}$.

In practice, since components typically operate when they are in healthy state, the number of data of class 0 (healthy state) is expected to be much larger than the number of data in a degraded state, i.e., $n_S^{y_{i_S}=0} \gg n_S^{y_{i_S}\neq 0}$, and different degradation states typically have different frequencies of occurrence. Then, in practical applications, $\{X_S, Y_S\}$ is typically an imbalanced dataset, characterized by the presence of at least one class with a fraction of representative patterns smaller than 40% of the number of patterns that belong to the majority class [26].

In this setting, the objective of the present work is to develop a method for the classification of the degradation/fault state of a set of patterns $X_T = \{\boldsymbol{x}_{i_T} \in \mathbb{R}^{m \times L}, i_T = 1, 2, ..., n_T\}$ collected from other same-type machines of the same fleet, for which the same m signals are measured during the condition monitoring, but the true class labels $Y_T = \{y_{i_T}, i_T = 1, 2, ..., n_T\}$ are not available and, therefore, a dedicated supervised diagnostic model cannot be directly built.

Notice that the situation in which diagnostic labels are not available, is a common one in several industrial applications, due to the cost of labeling the data and/or the difficulty/danger of operating the machines under abnormal or degraded conditions, especially for safety-critical or high-value systems.

III. Problem formulation

The fault diagnostic problem stated in Section II is here formulated in the context of DA. A domain $\mathcal{D}$ consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, from which a dataset $X = \{\boldsymbol{x_i}\}_{i=1}^{n} \in \mathcal{X}$ is sampled. For a given domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a task $\mathcal{T}$ is defined by a label space $\mathcal{Y}$ and a predictive function $f(\cdot)$, which can be learned from the pairs $\{\boldsymbol{x_i}, y_i\}$, where $\boldsymbol{x_i} \in X$ and $y_i \in \mathcal{Y}$.

We consider the case of two domains: the source domain $D_S = \{(\boldsymbol{x}_{i_S}, y_{i_S})\}_{i_S=1}^{n_S}$, with an associated learning task $\mathcal{T}_S$, and the target domain $D_T = \{(\boldsymbol{x}_{i_T}, y_{i_T})\}_{i_T=1}^{n_T}$, with the associated learning task $\mathcal{T}_T$. TL aims to improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, when $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. DA is a particular case of TL, which assumes that tasks and feature spaces in the source $\mathcal{D}_S$ and target $\mathcal{D}_T$ domains are the same ($\mathcal{T}_S = \mathcal{T}_T$ and $\mathcal{X}_S = \mathcal{X}_T$), but the marginal probability distributions of the data are different, $P(X_S) \neq P(X_T)$ [10].

In this work, the source domain $D_S = \{X_S, Y_S\}$ refers to the monitored machine for which labeled data containing signal values and corresponding degradation/fault states are available, whereas the target domain $D_T = \{X_T\}$ refers to other machines of the same fleet for which the degradation/fault states are not known. Notice that the source-domain data, D_S, and the target-domain data, D_T, share the same feature and label space, i.e., $\mathcal{X}_S = \mathcal{X}_T$ and $\mathcal{Y}_S = \mathcal{Y}_T$. However, since each machine is experiencing different operating and environmental conditions and maintenance interventions during its life, it is expected that D_S and D_T are originated from different marginal probability distributions, i.e., $P(X_S) \neq P(X_T)$. The learning tasks, $\mathcal{T}_S$ and $\mathcal{T}_T$, coincide and consist in the classification of the machine degradation/fault state.

The objective of the diagnostic work is to build a classifier for associating the signal measurements collected in the target domain, $\boldsymbol{x}_{i_T}$, to the corresponding label of the machine degradation/fault state.

IV. The proposed fault diagnostic method

A DANN-based DA approach is here developed. Since the dataset $\{X_S, Y_S\}$ is imbalanced, directly using a feature extractor $G_f(\boldsymbol{x}_i)$, which receives in input a pattern $\boldsymbol{x}_i$ and produces in output the extracted feature representation $\boldsymbol{f}_i = G_f(\boldsymbol{x}_i)$ [12], for the alignment of the marginal distributions of the source and target domains can lead to the mismatch of data belonging to different classes, and, therefore, to unsatisfactory classification performances in the target domain (negative transfer).

To alleviate this problem, the class-conditional distributions of the extracted features are individually aligned to obtain the feature representations such that $P(G_f(X_S)|Y_S) \approx P(G_f(X_T)|Y_T)$. Specifically, a deep multi-adversarial conditional DA network which uses an ensemble of K domain discriminators is proposed for cross-domain fault diagnostics. Each discriminator is responsible for the alignment of the conditional distributions of a specific class. The overall architecture of the proposed network is presented in Fig. 1. It consists of three modules: a CNN-based feature extractor, a classifier made of Fully-Connected (FC) layers, and an ensemble of domain discriminators.

- CNN-based Feature Extractor

  The feature extractor aims at seeking a latent feature space, in which the feature representations of the source and target domains are characterized by similar distributions. It defines the function $G_f = G_f(\boldsymbol{x}_i; \boldsymbol{\theta}_f) : \mathbb{R}^L \rightarrow \mathbb{R}^D$, with parameters $\boldsymbol{\theta}_f$, which maps the L-dimensional temporal sequence $\boldsymbol{x}_i$ into the D-dimensional feature representation $\boldsymbol{f}_i$. In

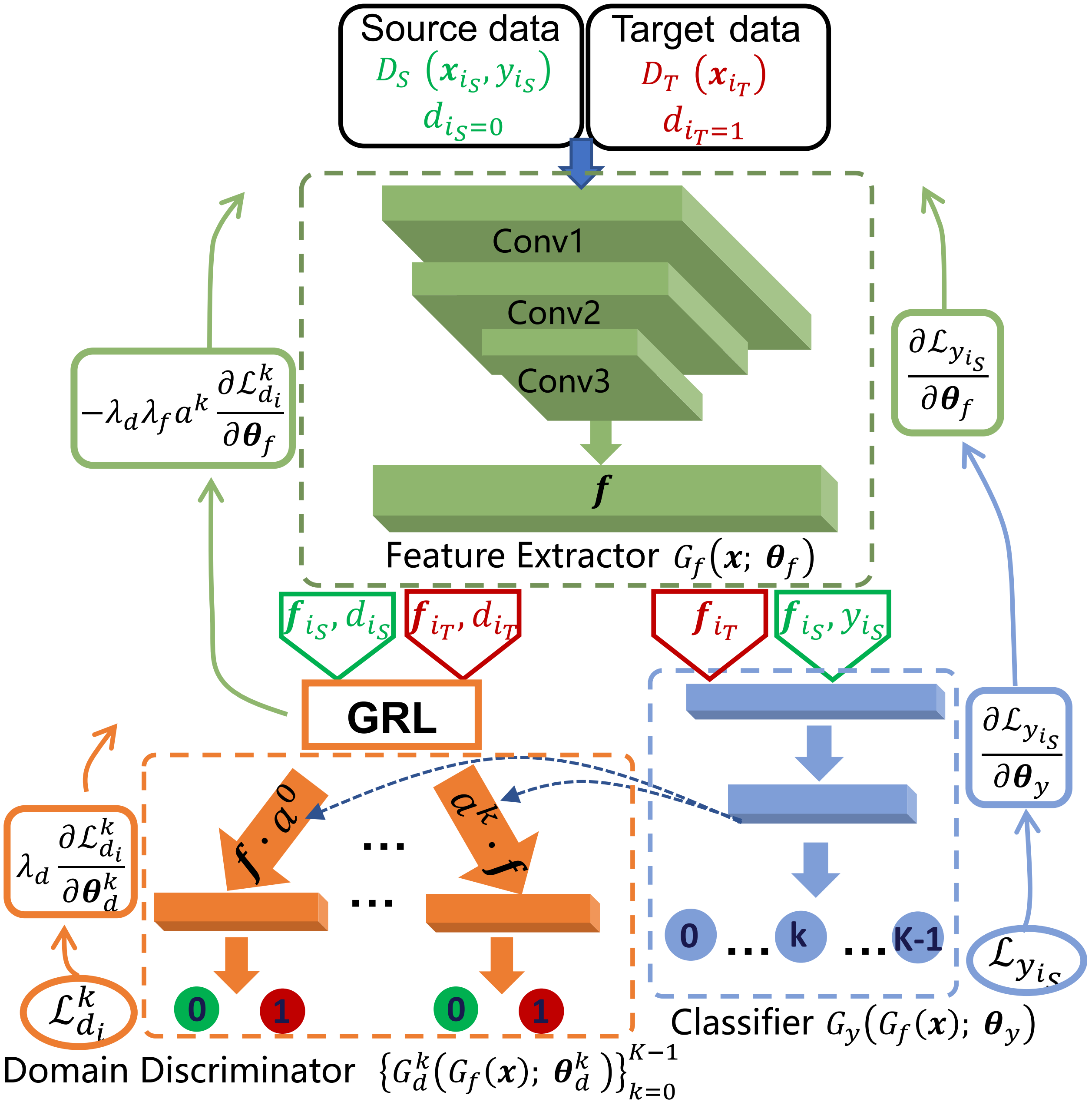

Source data
$D_S\ (\boldsymbol{x}_{i_S}, y_{i_S})$
$d_{i_S=0}$
Target data
$D_T\ (\boldsymbol{x}_{i_T})$
$d_{i_T=1}$
Conv1
Conv2
Conv3
$\boldsymbol{f}$
Feature Extractor $G_f(\boldsymbol{x};\ \boldsymbol{\theta}_f)$
$-\lambda_d\lambda_f a^k \frac{\partial \mathcal{L}^k_{d_i}}{\partial \boldsymbol{\theta}_f}$
$\frac{\partial \mathcal{L}_{y_{i_S}}}{\partial \boldsymbol{\theta}_f}$
$\boldsymbol{f}_{i_S}, d_{i_S}$
$\boldsymbol{f}_{i_T}, d_{i_T}$
$\boldsymbol{f}_{i_T}$
$\boldsymbol{f}_{i_S}, y_{i_S}$
GRL
$\boldsymbol{f}\cdot a^0$
...
$a^k\cdot \boldsymbol{f}$
...
0
1
0
1
$\lambda_d \frac{\partial \mathcal{L}^k_{d_i}}{\partial \boldsymbol{\theta}^k_d}$
$\mathcal{L}^k_{d_i}$
0
...
k
...
K-1
$\frac{\partial \mathcal{L}_{y_{i_S}}}{\partial \boldsymbol{\theta}_y}$
$\mathcal{L}_{y_{i_S}}$
Classifier $G_y(G_f(\boldsymbol{x});\ \boldsymbol{\theta}_y)$
Domain Discriminator $\{G^k_d(G_f(\boldsymbol{x});\ \boldsymbol{\theta}^k_d\ )\}_{k=0}^{K-1}$

Fig. 1. The architecture of the proposed network.

this work, the feature extractor is a Convolutional Neural Network (CNN) made by three convolutional blocks, each one formed by a convolutional layer and a pooling layer. This configuration has been already successfully applied in [12], [25] to learn discriminative features.

- Classifier

It receives in input the D-dimensional feature representation $\boldsymbol{f}_i$ and produces output a K-dimensional vector $\hat{\boldsymbol{y}}_i = [\hat{y}_i^0, \ldots, \hat{y}_i^k, \ldots, \hat{y}_i^{K-1}]$, with $\hat{y}_i^k$ indicating the probability that $\boldsymbol{x}_i$ belongs to the degradation class k: $G_y = G_y(G_f(\boldsymbol{x}_i); \boldsymbol{\theta}_y) : \mathbb{R}^D \rightarrow \mathbb{R}^K$. The classifier is formed by three FC layers followed by a softmax activation function, which produces the normalized probability distribution $\hat{\boldsymbol{y}}_i$ with values in the range [0,1] and whose sum is 1 [7]. The loss function used to train the classifier is the cross entropy:

$$\mathcal{L}_y(\boldsymbol{x}_i, y_i) = -\sum_{k=0}^{K-1} \mathbf{1}\{y_i = k\} \log G_y(G_f(\boldsymbol{x}_i)) \tag{1}$$

computed considering the labeled source domain data $D_S = \{X_S, Y_S\}$.

- Multiple Domain Discriminators

  An ensemble of discriminators made by K class-wise domain discriminators $G_d^0, \ldots, G_d^k, \ldots, G_d^{K-1}$ is developed to ensure the class-by-class alignment of the data distributions extracted from the source and target domains. The objective of the k-th discriminator $G_d^k = G_d^k(G_f(\boldsymbol{x}_i); \boldsymbol{\theta}_d^k) : \mathbb{R}^D \rightarrow [0, 1]$, with parameters $\boldsymbol{\theta}_d^k$, is to distinguish whether the feature representation $\boldsymbol{f}_i$ has been originated from a pattern $\boldsymbol{x}_i$ of class k of the source domain D_S (domain label $d_{i_S} = 0$) or from the target domain D_T (domain label $d_{i_T} = 1$). G_d^k is adversarially trained over the feature extractor G_f to reduce the divergence between the distributions of the source and target data of the k-th class.

  The training of the discriminators G_d^k requires, in principle, patterns of class k of both the source and target domains. This raises the issue that the degradation states y_{i_T} of the target patterns $\boldsymbol{x}_{i_T} \in D_T$ are not known and, therefore, the feature $\boldsymbol{f}_{i_T}$ extracted by G_f cannot be sent to the specific discriminator dedicated to the patterns of class y_{i_T}. To overcome this problem, we have developed a mechanism based on the use of an attention term, $a_i^k \in [0, 1]$, which indicates how much attention should be given by the discriminator G_d^k to a generic pattern, $\boldsymbol{x}_i$, during its training. Specifically, the attention a_i^k is used to weight the extracted feature representations, $\boldsymbol{f}_i = G_f(\boldsymbol{x}_i)$, i.e., the k-th discriminator, G_d^k, receives in input $a_i^k \cdot G_f(\boldsymbol{x}_i)$. The effect of this weighting is that the updating of the model weights performed by the error backpropagation method during the training of the discriminator G_d^k is proportional to a_i^k [27].

  With respect to the source domain, the attention $a_{i_S}^k$ given by the discriminator G_d^k to the pattern $\boldsymbol{x}_{i_S} \in D_S$ of class y_{i_S} is:

$$a_{i_S}^k = \begin{cases} 0 & \text{if } k \neq y_{i_S} \\ 1 & \text{if } k = y_{i_S} \end{cases}, \quad i_S = 1, 2, \ldots, n_S \tag{2}$$

and, therefore, $\boldsymbol{x}_{i_S}$ contributes to the training of only the discriminator dedicated to the pattern of its true class y_{i_S}.

With respect to the target domain, the attention $a_{i_T}^k$ given by the discriminator G_d^k to the pattern $\boldsymbol{x}_{i_T} \in D_T$ is:

$$a_{i_T}^k = \begin{cases} 0 & \text{if } \hat{y}_{i_T}^k < \frac{1}{K} \\ \hat{y}_{i_T}^k & \text{if } \hat{y}_{i_T}^k \geq \frac{1}{K} \end{cases}, \quad i_T = 1, 2, \ldots, n_T \tag{3}$$

where $\hat{y}_{i_T}^k$ is the probability, estimated by the classifier G_y, that $\boldsymbol{x}_{i_T}$ is of class k. The motivation of the assignment is that, since the true class of the target patterns is not known, its estimation provided by the classifier is used. The effect is that G_d^k will pay more attention on the patterns $\boldsymbol{x}_{i_T}$ with larger attentions $a_{i_T}^k$ during its training. The threshold $\frac{1}{K}$ used in Eq. (3), which corresponds to the probability that $\boldsymbol{x}_{i_T}$ is of class k for a naive classifier, allows limiting the propagation of the errors from G_y to G_d^k.

Each discriminator is formed by a binary classifier with two FC layers. Given a pattern $\boldsymbol{x}_i \in D_S \cup D_T$, the domain classification loss function of the k-th domain discriminator is the binary cross entropy:

$$\begin{aligned} \mathcal{L}_d^k(\boldsymbol{x}_i, d_i) = -[&d_i \log G_d^k(a_i^k \cdot G_f(\boldsymbol{x}_i)) \\ &+ (1 - d_i) \log(1 - G_d^k(a_i^k \cdot G_f(\boldsymbol{x}_i)))] \end{aligned} \tag{4}$$

The proposed method is built based on the Domain-Adversarial Neural Network (DANN) [12], from which the feature extractor G_f and the classifier G_y derive. The main novelty is the use of an ensemble of discriminators and their training using Eq. (4). The feature extractor G_f, the classifier G_y and the K domain discriminators $\{G_d^k\}_{k=0}^{K-1}$ are jointly trained in a deep feed-forward network. Specifically, the parameters $\boldsymbol{\theta}_f$, $\boldsymbol{\theta}_y$ of G_f, G_y are synergistically optimized to extract discriminative feature representations for the given fault diagnostic task, through minimizing the classification loss of degradation states $\mathcal{L}_y(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y)$ on the labeled source domain data D_S. The

Algorithm 1 Training procedure of the proposed deep multi-adversarial conditional DA network

```
Input: source dataset: $D_S = \{(\boldsymbol{x}_{i_S}, y_{i_S})\}_{i=1}^{n_S}$; target dataset: $D_T = \{\boldsymbol{x}_{i_T}\}_{i=1}^{n_T}$; mini-batch size for
    source and target datasets: $h$; hyper-parameters: $\lambda_d$, $\lambda_f$; learning rate: $\mu$
Output: optimal $\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_y, \{\hat{\boldsymbol{\theta}}_d^k\}_{k=0}^{K-1}$
initialize the parameters $\boldsymbol{\theta}_f, \boldsymbol{\theta}_y$ and $\{\boldsymbol{\theta}_d^k\}_{k=0}^{K-1}$ of the feature extractor $G_f$, the classifier $G_y$
    and the $K$ domain discriminators $\{G_d^k\}_{k=0}^{K-1}$;
repeat
    sample mini-batch $\{\boldsymbol{x}_{i_S}, y_{i_S}\}_{i=1}^{h}$ from $D_S$;
    sample mini-batch $\{\boldsymbol{x}_{i_T}\}_{i=1}^{h}$ from $D_T$;
    update $\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_y, \{\hat{\boldsymbol{\theta}}_d^k\}_{k=0}^{K-1}$ by:
    $\boldsymbol{\theta}_f \longleftarrow \boldsymbol{\theta}_f - \mu \left( \frac{\partial \mathcal{L}_y}{\partial \boldsymbol{\theta}_f} - \lambda_d \lambda_f a^k \frac{\partial \mathcal{L}_d^k}{\partial \boldsymbol{\theta}_f} \right)$
    $\boldsymbol{\theta}_y \longleftarrow \boldsymbol{\theta}_y - \mu \frac{\partial \mathcal{L}_y}{\partial \boldsymbol{\theta}_y}$
    $\boldsymbol{\theta}_d^k \longleftarrow \boldsymbol{\theta}_d^k - \mu \lambda_d \frac{\partial \mathcal{L}_d^k}{\partial \boldsymbol{\theta}_d^k}$
until $\boldsymbol{\theta}_f, \boldsymbol{\theta}_y$ and $\{\boldsymbol{\theta}_d^k\}_{k=0}^{K-1}$ converge.
```

parameters $\{\boldsymbol{\theta}_d^k\}_{k=0}^{K-1}$ of $\{G_d^k\}_{k=0}^{K-1}$ are adversarially optimized over $\boldsymbol{\theta}_f$ to enable obtaining domain-invariant feature representations for each degradation state, resorting to multiple two-player minimax games in which $\{G_d^k\}_{k=0}^{K-1}$ minimize the domain classification loss $\mathcal{L}_d(\boldsymbol{\theta}_f, \{\boldsymbol{\theta}_d^k\}_{k=0}^{K-1})$, whereas G_f maximizes it oppositely, using domain labels of D_S and D_T. The overall loss function, which is obtained by combining the loss function of the classifier of the degradation state, $\mathcal{L}_y$, reported in Eq. (1), and the loss function of the domain discriminators, $\mathcal{L}_d^k$ with $k = 0, 1, ..., K-1$, reported in Eq. (4), is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y, \{\boldsymbol{\theta}_d^k\}_{k=0}^{K-1}) &= \frac{1}{n_S} \sum_{\boldsymbol{x}_i \in D_S} \mathcal{L}_y \left(G_y(G_f(\boldsymbol{x}_i)), y_i \right) \\ -\frac{1}{n} \sum_{\boldsymbol{x}_i \in D_S \cup D_T} & \sum_{k=0}^{K-1} \mathcal{L}_d^k \left(\lambda_d \cdot G_d^k(\lambda_f \cdot a_i^k \cdot G_f(\boldsymbol{x}_i)), d_i \right)\end{aligned} \tag{5}$$

where $n = n_S + n_T$, n_S is the number of patterns in the labeled source-domain dataset $D_S = \{\boldsymbol{x}_{i_S}, y_{i_S}\}_{i=1}^{n_S}$ and n_T is the number of patterns in the unlabeled target-domain dataset $D_T = \{\boldsymbol{x}_{i_T}\}_{i=1}^{n_T}$. The hyper-parameters λ_d and λ_f are introduced to weight the loss functions in Eq. (5). Specifically, the updating of the weights $\boldsymbol{\theta}_d^k$ of the domain discriminators $\{G_d^k\}_{k=0}^{K-1}$ is influenced by λ_d, whereas the updating of the weights $\boldsymbol{\theta}_f$ of the feature extractor G_f is influenced by both λ_d and λ_f, as shown in Fig. 1 and reported in Algorithm 1. The parameters $\boldsymbol{\theta}_f, \boldsymbol{\theta}_y, \{\boldsymbol{\theta}_d^k\}_{k=0}^{K-1}$ can be optimized by seeking a saddle point solution $\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_y, \{\hat{\boldsymbol{\theta}}_d^k\}_{k=0}^{K-1}$ so that

$$(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_y) = \underset{\boldsymbol{\theta}_f, \boldsymbol{\theta}_y}{\arg\min}\, \mathcal{L}(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y, \{\hat{\boldsymbol{\theta}}_d^k\}_{k=0}^{K-1}) \tag{6}$$

$$\{\hat{\boldsymbol{\theta}}_d^k\}_{k=0}^{K-1} = \underset{\{\boldsymbol{\theta}_d^k\}_{k=0}^{K-1}}{\arg\max}\, \mathcal{L}(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_y, \{\boldsymbol{\theta}_d^k\}_{k=0}^{K-1}) \tag{7}$$

which can be trained by the standard backpropagation algorithm. Algorithm 1 summarizes the pseudo-code of the proposed method. Fig. 1 shows the Gradient Reversal Layer (GRL), which generates the minus sign (–) between the gradient from G_y, $\frac{\partial \mathcal{L}_y}{\partial \boldsymbol{\theta}_f}$, and the gradient from G_d^k, $\frac{\partial \mathcal{L}_d^k}{\partial \boldsymbol{\theta}_f}$. The GRL, which is widely applied in adversarial learning [12], acts as an identity transformation during the forward propagation, but reverses the sign of the gradients from G_d^k before passing them to G_f during the backpropagation.

V. Case studies

The proposed method is applied to fault diagnostics of bearings of rotating machines (case study 1), and of automatic doors used in the railway industry (case study 2).

A. Case study 1

Since the failures in bearings are the primary cause of rotating machinery unavailability, their health-degradation-fault state monitoring is of paramount importance in many industrial sectors. In this work, we consider two bearing datasets, IMS and CWRU [28], containing data patterns of four classes (healthy (H) and degraded due to defects at the Inner-race (I), Outer-race (O) and Ball (B)), collected from two different experimental platforms. The objective is to show that the novel method proposed in this work can correctly classify the degradation state of bearings working in an experimental platform different from the one from which the labeled patterns used to train the model have been collected.

The IMS bearing dataset has been generated using an experiment platform made by four double-row bearings on a shaft [28]. Vibration signals are measured by accelerometers installed on the bearing housings at a frequency of 20 kHz. The CWRU bearing dataset provided by the

TABLE I

CONFIGURATION OF THE DEVELOPED MODELS IN CASE STUDIES 1 AND 2

G_f	kernel/pooling size		filter	output size		G_y	input size		output size	G_d^k	input size		output size
case study	1	2	1=2	1	2		1	2	1=2		1	2	1=2
conv1/pooling1	197×1/2×1	57×1/2×1	64	502×64	222×64	FC_1	1760	2720	256	FC_1	1760	2720	1024
conv2/pooling2	123×1/2×1	33×1/—	64	190×64	190×64	FC_2	256	256	64	FC_2	1024	1024	2
conv3/pooling3	81×1/2×1	21×1/2×1	32	55×32	85×32	FC_3	64	64	4				
FC	—	—	—	1760	2720								

Case Western Reserve University (CWRU) has been generated on another experimental platform made by a 2 hp motor and test bearings [28]. Also in this case, vibration signals are measured by accelerometers attached to the housing, at a frequency of 12 kHz.

We consider a one-dimensional time series constituted by $L = 1200$ consecutive acceleration measurements as input pattern $\boldsymbol{x}_i \in \mathbb{R}^{1\times 1200}$ of the diagnostic model. Both datasets are characterized by imbalanced multi-class distributions of the majority class H (500 patterns in the IMS dataset and 400 in the CWRU dataset) and the under-represented classes I, O, and B (100 patterns of class I, 100 of class O and 200 of class B in the IMS dataset; 80 patterns of class I, 80 of class O and 100 of class B in the CWRU dataset). In both datasets, the most under-represented classes (I and O) are present with an imbalance ratio (the ratio between the number of patterns of the minority and majority classes) of 20%, which is smaller than the limit of 40% defining an imbalanced dataset [26].

The labeled IMS dataset is firstly used as source dataset $D_S = \{X_S^{IMS}, Y_S^{IMS}\}$ and the unlabeled CWRU dataset as target dataset $D_T = \{X_T^{CWRU}\}$. The input data are min-max normalized and, then, fed into our proposed deep multi-adversarial conditional DA network. The configuration of the overall network architecture is reported in Table I. The CNN-based feature extractor G_f uses wide convolution kernels to capture low-frequency features and suppress high-frequency noises.

The hyper-parameter, λ_d, used for updating the domain discriminators $\{G_d^k\}_{k=0}^{K-1}$ is set equal to 1 to ensure that G_d^k is trained as fast as the classifier G_y [12]. To avoid the heavy computational cost of a grid search for the selection of the other model hyper-parameters, the hyper-parameter, λ_f, used for updating the feature extractor G_f, and the learning rate, μ, are adaptively adjusted [12]. Regarding the adaptation of λ_f:

$$\lambda_f = \frac{2}{1 + exp(-10 \cdot p)} - 1 \tag{8}$$

where p is set equal to the ratio between the current epoch, $epoch_i$, and the maximum number of epochs, $epoch_{max}$:

$$p = \frac{epoch_i}{epoch_{max}} \tag{9}$$

As a result, the optimization of the feature extractor G_f is dominated by the error of the classifier G_y at the early stage of the training procedure, when λ_f is small, which allows extracting features more discriminative for the classification and, therefore, to propagate less errors to the domain discriminators $\{G_d^k\}_{k=0}^{K-1}$. With respect to the setting of the learning rate μ, the Cyclical Learning Rate (CLR) policy, which cyclically varies μ within a band of values $[base\_lr, max\_lr]$, is used to eliminate the need of manual tuning [29]. Specifically, the *LR range test* [29] is applied to set $base\_lr$ and max_lr, which result equal to 5e-4 and 3e-3, respectively, and max_lr exponentially decreases until it reaches $base\_lr$. The mini-batch size, h, is set equal to 16 to achieve a robust convergence of the training algorithm [30], and the maximum training epoch $epoch_{max}$ is set equal to 2500. Since the target domain data are unlabeled, the training procedure is not terminated until the classification loss (Eq. (1)) on the labeled test data of the source domain stabilizes and does not decrease for 100 successive epochs, as suggested in [12].

The results of the proposed method are compared to the results obtained by the following state-of-the-art methods:

- a CNN trained using only the source dataset $D_S = \{X_S, Y_S\}$ and directly tested on the target input data X_T; this method, which will be referred to as “M1”, provides a lower performance bound for the DA-based methods.
- two marginal DA methods based on DANN [12] (denoted as “M2”) and Wasserstein Distance Guided Representation Learning (WDGRL) [16] (denoted as “M3”), respectively.
- four conditional DA methods based on multi-feature concatenation ($f \oplus \hat{y}$) [31] (denoted as “M4”), multilinear mapping ($f \otimes \hat{y}$) [18] (denoted as “M5”), multilinear mapping combined

with entropy weight ($f \otimes \hat{y}$ + entropy weight) [18] (denoted as "M6") and Batch Spectral Penalization (BSP) [20] (denoted as "M7").

To ensure the fairness of comparison, the feature extractor G_f and the condition predictor G_y of all methods adopt the same network architecture of the proposed method (Table I), and the same strategies are applied for the adaptive setting of the hyper-parameters and the stopping criterion. The models of all methods are trained using the Adam optimization method [32]. A 5-fold cross validation approach is applied to evaluate the diagnostic performance. Specifically, the target datasets D_T is split into five smaller subsets (folds), then, each of the five folds is used as a test set to compute the classification accuracy, and the remaining four folds are merged and used for the model development. Stratified sampling is applied to ensure that the fractions of patterns of each class are approximately preserved in each train and test fold. Given the imbalanced distribution of the patterns among the classes, the performances of the methods are evaluated considering the *F-score* metric for each class $k = 0, 1, \ldots, K-1$:

$$F - score_k = \frac{2P_k R_k}{P_k + R_k} \tag{10}$$

where P_k and R_k are the class k precision and recall, respectively:

$$P_k = \frac{TP_k}{TP_k + FP_k} \tag{11}$$

$$R_k = \frac{TP_k}{TP_k + FN_k} \tag{12}$$

where TP_k is the number of test patterns of class k that have been correctly assigned to class k, FP_k is the number of test patterns that are not class k but have been erroneously assigned to class k, and FN_k is the number of test patterns of class k that have been erroneously assigned to other classes.

The obtained results and the corresponding computational times on a GeForce RTX 2070 SUPER GPU are reported in Table II. Notice that: 1) as expected, all methods based on DA outperform M1; 2) the overall classification accuracy of the conditional DA methods tends to be close to the accuracy of DANN, and it is more satisfactory than the accuracy of WDGRL; 3) all methods, except the proposed one, provide very unsatisfactory performances on class O; 4) the proposed method achieves the best classification accuracy on the whole test data, and

TABLE II

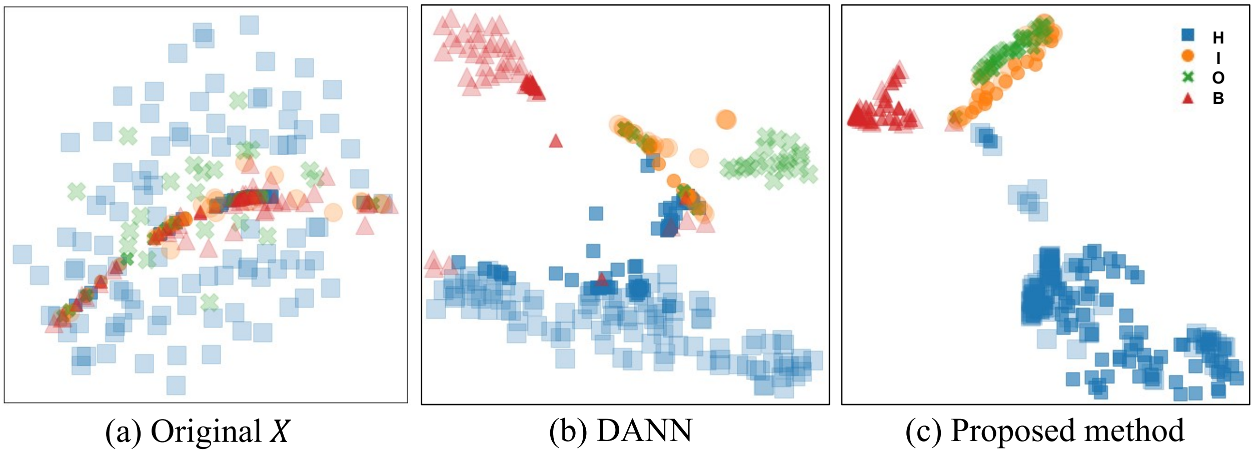
CLASSIFICATION RESULTS OF CASE STUDY 1; COMPUTATIONAL TIME IS REPORTED IN (MINUTE:SECOND)

Method	**case study 1**													
	source: IMS → target: CWRU							source: CWRU → target: IMS						
	mean *F-score*				overall accuracy		time	mean *F-score*				overall accuracy		time
	H	I	O	B	*mean*	*std*	*(m:s)*	H	I	O	B	*mean*	*std*	*(m:s)*
M1	0.80	0.39	0.11	0.00	0.63	0.08	03:01	0.74	0.05	0.02	0.13	0.57	0.02	02:34
M2	**1.00**	0.63	0.00	0.97	0.86	0.01	08:45	0.87	0.13	0.25	0.47	0.68	0.09	09:51
M3	0.98	0.48	0.07	0.49	0.78	0.05	22:20	0.91	0.11	0.02	0.78	0.74	0.01	20:51
M4	0.99	0.66	0.00	0.96	0.87	0.02	09:53	0.95	0.30	0.44	0.64	0.75	0.04	10:40
M5	**1.00**	**0.72**	0.11	0.91	0.88	0.01	10:58	0.92	0.25	0.43	0.68	0.73	0.06	10:44
M6	**1.00**	0.70	0.22	0.86	0.87	0.01	12:18	0.90	0.23	0.33	0.55	0.71	0.04	11:52
M7	0.99	0.68	0.08	0.92	0.87	0.01	20:01	0.95	0.20	0.36	0.80	0.80	0.03	20:55
Proposed	0.99	**0.72**	**0.68**	**0.99**	**0.93**	0.02	15:36	**0.98**	**0.31**	**0.57**	**0.85**	**0.84**	0.02	14:05

significantly improves the classification accuracy on classes I and O; 5) the proposed method requires more computational effort than M1, M2, M4, M5 and M6, due to the increase of the model complexity, but still smaller computational effort than M3 and M7.

Fig. 2 shows the two-dimensional representations provided by t-SNE embedding [33] of the original features X and the features $\boldsymbol{f}$ extracted by G_f. It can be observed that the original features are not discriminative among different classes and that there is a mismatch between the distributions of the data in the source and target domains. Also, the divergence between the distributions of the source and target data of each class is still remarkably larger in the space of the features extracted by DANN, where the target data of class O completely overlap with the source data of class I, than in the feature space obtained by the proposed method. This result confirms the capability of the class-specific domain discriminators employed by the proposed method of mitigating the effects of negative transfer on the two most under-represented classes I and O.

The proposed approach for the training of the multiple domain discriminators, which is based on the use of the attention mechanism described in Section IV, is here compared with



(a) Original X (b) DANN (c) Proposed method

Fig. 2. Two-dimensional representation of the spaces of: *a*) the original features, *b*) the features extracted by DANN and *c*) the features extracted by the proposed method. The source and target test data are represented using less and more opacity, respectively.

another possible alternative strategy, which assigns each pattern to the training set of only one discriminator. Specifically, the generic pattern $\boldsymbol{x}_{i_S}$ of class k of the source domain is assigned to the training set of the discriminator G_d^k, whereas the generic pattern $\boldsymbol{x}_{i_T}$ of the target domain is assigned to the training set of a discriminator G_d^k randomly sampled from the discrete probability distribution of the classes, $\hat{\boldsymbol{y}}_{i_T}$, provided by the classifier G_y. Fig. 3 shows the superior performance of the proposed method, especially on the under-represented classes I and O, which confirms the capability of the attention mechanism of providing class specific information for the training of the class-specific domain discriminators $\{G_d^k\}_{k=0}^{K-1}$.

Furthermore, the effect of the amount of data used for the development of the proposed model on the classification performance is investigated in Fig. 4 considering the transfer task IMS → CWRU. The total number of training data in the source and target domains are reduced to 50% and 20%, respectively, by using stratified sampling to preserve the imbalanced distributions among the data of the different classes. It can be observed that: 1) a 50% reduction of the data causes the significant decrease of the classification performance of only the minority class O; 2) when the number of data is further reduced to 20% of the original dataset with only 100 patterns of class H, 20 of class I, 20 of class O and 40 of class B in the IMS dataset, and 80 patterns of class H, 16 of class I, 16 of class O and 20 of class B in the CWRU dataset, the overall classification performance becomes unsatisfactory due to the insufficient number of

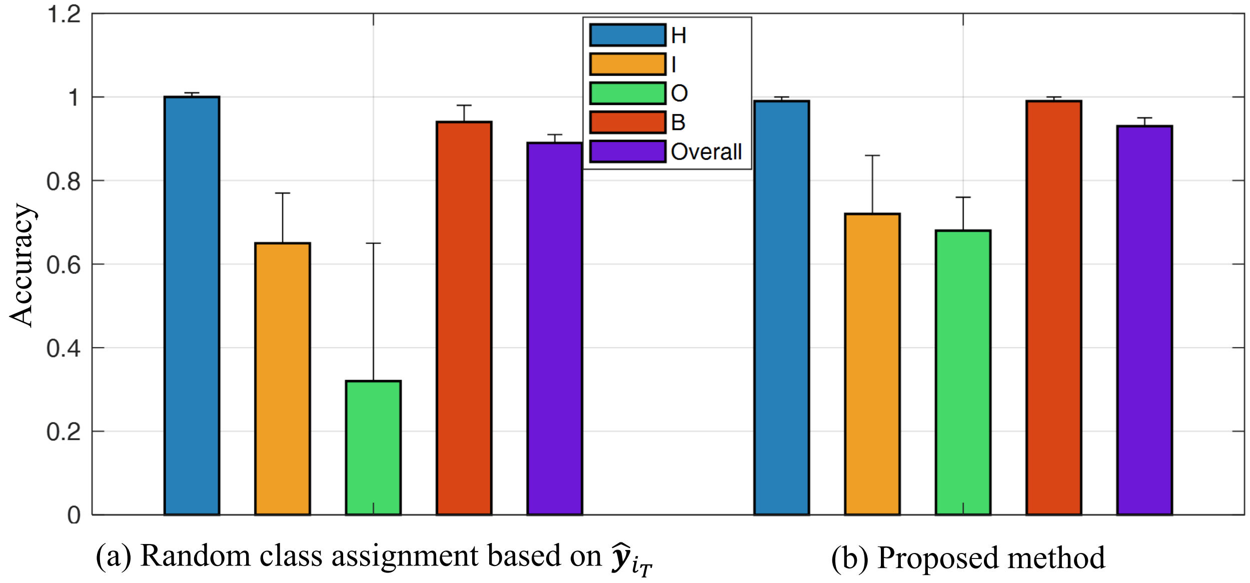


Fig. 3. Comparison of different strategies for the training of the multiple domain discriminators: *a*) random assignment based on $\hat{\boldsymbol{y}}_{i_T}$ and *b*) the proposed attention mechanism.

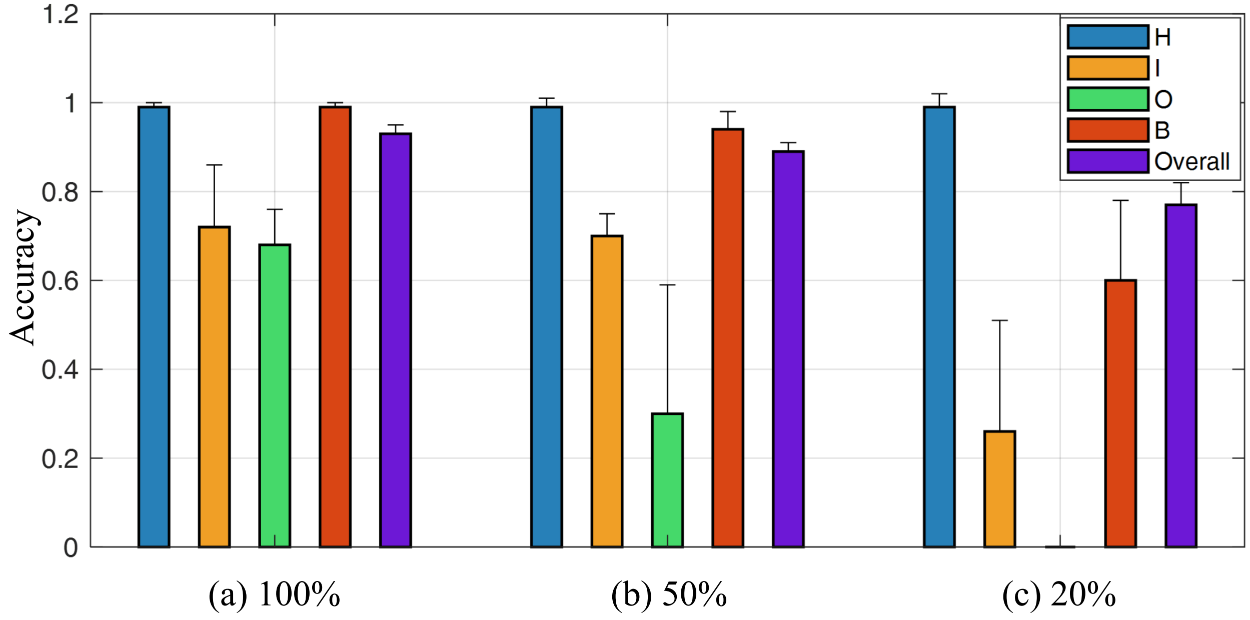


Fig. 4. Classification performance when 100%, 50% and 20% of the available data are used for model development, considering the transfer task IMS → CWRU.

training patterns.

The proposed method is also verified considering the reverse DA task: the CWRU dataset is used as source dataset $D_S = \{X_S^{CWRU}, Y_S^{CWRU}\}$ and the IMS dataset as target dataset $D_T = \{X_T^{IMS}\}$. The performances obtained using the same architecture and strategy for hyper-parameters setting are reported in Table II.

Since the CWRU dataset $(n_S = 660)$ contains less data than the IMS dataset $(n_T = 900)$, the classification accuracy on the DA task from CWRU to IMS is less satisfactory than the classification accuracy on the DA task from IMS to CWRU. These results are due to the negative transfer problem, which is more relevant when the source domain contains few and class-imbalanced training patterns, and, therefore, the domain discrepancy cannot be effectively reduced by only aligning the marginal distributions. Still, also in this case, the proposed method remarkably improves the classification performance of class O.

B. Case study 2

We consider the problem of classifying the degradation state of automatic doors used in a fleet of high-speed trains, in the situation in which labeled data are available only for the doors of one train.

Two signals, whose names are not reported in this work for confidentiality reasons, are measured during the opening and closing of the door. Fig. 5 shows an example of pattern $\boldsymbol{x}_i \in \mathbb{R}^{2\times 500}$, collected during a door opening (left) and closing (right), which last for a period of approximately 5 seconds ($L = 500$ measurements). Two datasets have been collected from two different trains, hereafter referred to as train#1 and train#2. They contain patterns of four classes, which correspond to the healthy state (N0) and three degraded states caused by different degradation mechanisms (F1, F2 and F3). The classification problem is imbalanced, being the number of patterns of the healthy state class significantly larger (100 patterns of class N0 for train#1 and 86 for train#2) than the number of patterns of the classes corresponding to the degraded states (42 patterns of class F1, 38 of class F2 and 74 of class F3 for train#1; 20 patterns of class F1, 18 of class F2 and 22 of class F3 for train#2). Notice that the most under-represented class F2 is present in the dataset with an imbalance ratio of 38$\%$ for train#1 and of 21$\%$ for train#2, which are smaller than the limit of 40$\%$ defining an imbalanced dataset [26]. Also, the dataset collected from train#1 contains more patterns than the dataset collected from train#2. This is due to the fact that the doors installed in the two trains operate in different operating conditions due to different mounting settings and characteristics of the train journeys

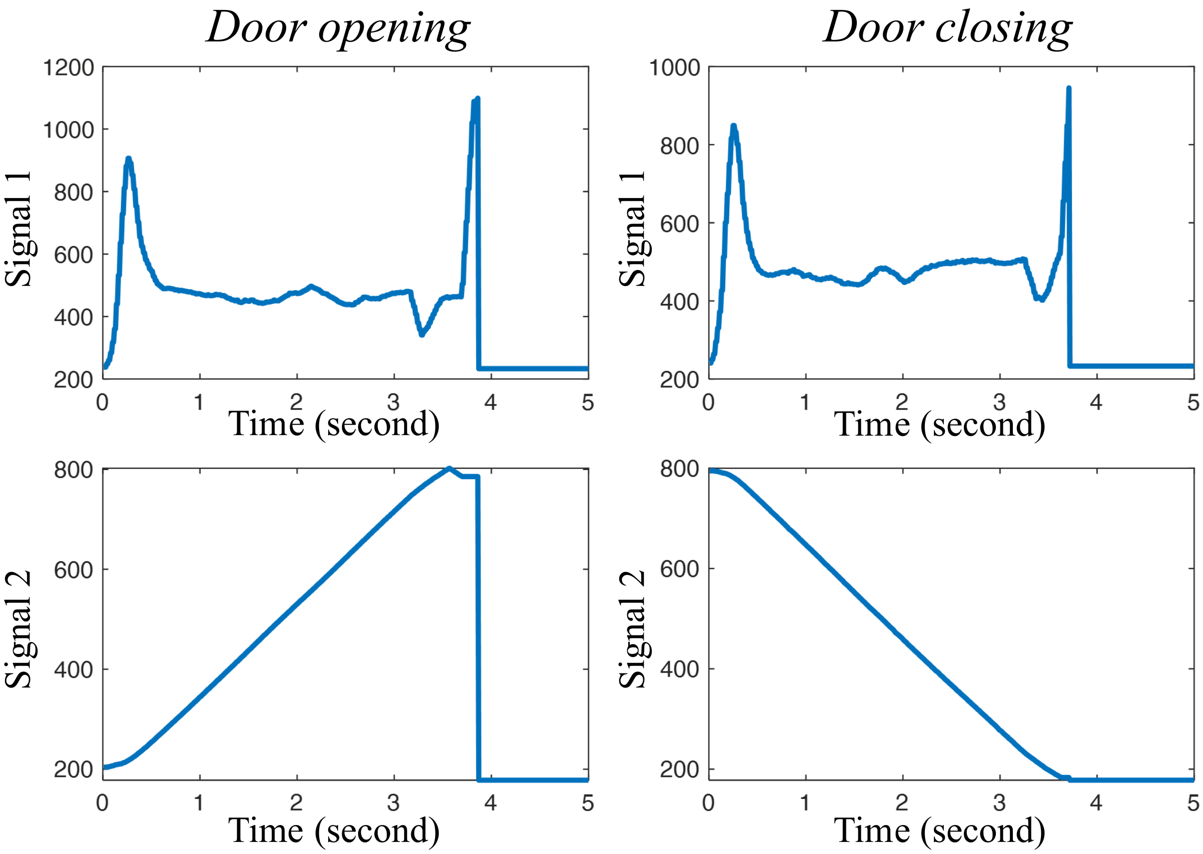


Fig. 5. Signal measurements during a door opening (left) and closing (right). Measurements units are not reported for confidentiality reasons.

and, therefore, the corresponding datasets are characterized by different marginal and class-conditional distributions.

We consider two different cross-domain fault diagnostic tasks: $train\#1 \rightarrow train\#2$, in which the labeled $train\#1$ dataset is used as source domain $D_S^A = \{X_S^{train\#1}, Y_S^{train\#1}\}$ and the unlabeled $train\#2$ dataset is used as target domain $D_T^A = \{X_T^{train\#2}\}$, and $train\#2 \rightarrow train\#1$, in which the labeled $train\#2$ dataset is used as source domain $D_S^B = \{X_S^{train\#2}, Y_S^{train\#2}\}$ and the unlabeled $train\#1$ dataset is used as target domain $D_T^B = \{X_T^{train\#1}\}$. As for the first case study, the input data are min-max normalized and the same strategy of adaptively adjusting the hyper-parameter λ_f and the learning rate μ is applied. Also, all methods are based on the same network architectures of the feature extractor G_f and the classifier G_y (Table I) and the same stopping criterion is used. The mini-batch size h is set equal to 16, and the maximum training epoch $epoch_{max}$ to 2500. The obtained performances and the corresponding computational times

TABLE III

CLASSIFICATION RESULTS OF CASE STUDY 2; COMPUTATIONAL TIME IS REPORTED IN (MINUTE:SECOND)

Method	**case study 2**												
	source: train#1 → target: train#2							source: train#2 → target: train#1						
	mean *F-score*				overall accuracy		time	mean *F-score*				overall accuracy		time
	N0	F1	F2	F3	*mean*	*std*	*(m:s)*	N0	F1	F2	F3	*mean*	*std*	*(m:s)*
M1	0.86	0.00	0.26	0.66	0.70	0.07	01:15	0.93	0.58	0.30	0.68	0.72	0.04	01:34
M2	0.94	0.67	0.86	0.88	0.89	0.03	04:10	0.88	0.18	0.70	0.78	0.76	0.03	04:05
M3	0.96	0.76	0.71	0.89	0.90	0.03	08:37	0.91	0.58	0.70	0.82	0.82	0.02	10:50
M4	0.93	0.71	0.88	0.96	0.90	0.04	04:54	0.86	0.14	0.75	0.82	0.76	0.02	04:34
M5	0.94	0.79	0.82	**0.97**	0.91	0.02	05:13	0.94	0.34	0.56	0.76	0.78	0.05	04:46
M6	0.95	0.70	0.75	0.89	0.89	0.03	05:38	**0.96**	0.56	0.69	0.77	0.81	0.02	04:50
M7	**0.97**	0.78	0.76	0.89	0.91	0.03	09:31	0.89	0.27	0.82	0.81	0.79	0.03	09:05
Proposed	0.96	**0.86**	**0.92**	0.91	**0.94**	0.03	07:27	0.94	**0.60**	**0.94**	**0.88**	**0.87**	0.02	08:27

within a 5-fold cross validation are reported in Table III.

Considering the cross-domain diagnostic task ($train\#1 \rightarrow train\#2$), all the methods performing DA provide more satisfactory results than the "M1" method, which confirms that environments and operating conditions in the two trains are remarkably different, and cause large divergence in the data distributions. All conditional DA methods outperform the method DANN, although, as expected, the computational burden is increased. The proposed method provides the best overall accuracy, which is mainly due to the improvement in the classification of the minority classes F1 and F2. This result confirms the capability of the proposed multi-adversarial conditional adaptation model of performing class-by-class transfer. Considering the cross-domain diagnostic task ($train\#2 \rightarrow train\#1$), characterized by the availability of less data in the source domain than in the target domain, negative transfer causes an overall worsening of the performances due to the difficulty of aligning the class-conditional distributions. Notice, however, that the proposed method still provides the most satisfactory results.

TABLE IV

RESULTS OF THE HOLM POST-HOC TEST

Method	Average rank	Z-value	p-value	Adjusted p-value	Hypothesis ($\alpha = 0.05$)
M1	7.28	7.488	<0.00001	<0.00007	Rejected
M2	5.78	5.551	<0.00001	<0.00007	Rejected
M3	5.00	4.544	<0.00001	<0.00007	Rejected
M4	4.38	3.744	0.000181	0.00054	Rejected
M5	3.73	2.905	0.003673	0.00550	Rejected
M6	4.58	4.002	0.000063	0.00025	Rejected
M7	3.80	2.995	0.002744	0.00550	Rejected
Proposed	1.48	—	—	—	—

C. Performance comparison

Friedman and Holm post-hoc tests [34] have been performed to verify whether the overall accuracy of the proposed method in the two case studies is superior to that of the other state-of-the-art methods.

Considering the four class-specific $F-scores$ and the overall accuracy on the whole test data in the four performed domain adaptation tasks (IMS → CWRU and CWRU → IMS in case study 1, train#1 → train#2 and train#2 → train#1 in case study 2), the null-hypothesis of no significant difference in the performance of all methods is rejected by the Friedman test with a level of significance $\alpha = 0.05$, being Friedman test statistic $\mathcal{X}_F^2$ equal to 66.13 which is larger than $\mathcal{X}_{.05}^2 = 14.07$. Then, according to the Holm post-hoc test, all the null-hypotheses of no significant difference between the proposed method and any of the other methods are rejected since the adjusted p-values are smaller than the level of significance $\alpha = 0.05$ (Table IV). This allows concluding that the proposed method outperforms the other comparison methods in terms of the overall accuracy on the two case studies.

VI. CONCLUSION

A novel deep multi-adversarial conditional DA network is developed for fault diagnostics of a fleet of machines. Differently from conventional deep learning approaches, the method is able to reduce the divergence among distributions of data measured from different machines, by extracting domain-invariant feature representations through an adversarial learning process between the CNN-based feature extractor and an ensemble of domain discriminators. The problem of negative transfer, which is typical of fault diagnostic applications characterized by imbalanced multi-class datasets, has been overtaken by developing an innovative solution based on an ensemble of discriminators, which allows aligning the weighted data distributions class by class. The method has been verified considering two cross-domain fault diagnostic case studies from different industrial sectors. The obtained results show the superior diagnostic performance of the proposed method with respect to other state-of-the-art methods, which is confirmed by applying Friedman and Holm post-hoc tests. Future work will include the heterogeneous TL scenario, in which the feature spaces between the source and target domains are different, and possibly characterized by different dimensionality.

References

[1] G. J. Vachtsevanos and G. J. Vachtsevanos, *Intelligent fault diagnosis and prognosis for engineering systems*. Wiley Online Library, 2006, vol. 456.

[2] P. Baraldi, F. Di Maio, and E. Zio, "Unsupervised clustering for fault diagnosis in nuclear power plant components," *International Journal of Computational Intelligence Systems*, vol. 6, no. 4, pp. 764–777, 2013.

[3] M. He and D. He, "Deep learning based approach for bearing fault diagnosis," *IEEE Transactions on Industry Applications*, vol. 53, no. 3, pp. 3057–3065, 2017.

[4] H. Shao, H. Jiang, X. Zhang, and M. Niu, "Rolling bearing fault diagnosis using an optimization deep belief network," *Measurement Science and Technology*, vol. 26, no. 11, p. 115002, 2015.

[5] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171–178, 2016.

[6] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5990–5998, 2018.

[7] S. R. Saufi, Z. A. B. Ahmad, M. S. Leong, and M. H. Lim, "Gearbox fault diagnosis using a deep learning model with limited data sample," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6263–6271, 2020.

[8] H. Zheng, R. Wang, Y. Yang, J. Yin, Y. Li, Y. Li, and M. Xu, "Cross-domain fault diagnosis using knowledge transfer strategy: A review," *IEEE Access*, vol. 7, pp. 129 260–129 290, 2019.

[9] Z. Yang, S. Al-Dahidi, P. Baraldi, E. Zio, and L. Montelatici, "A novel concept drift detection method for incremental learning in nonstationary environments," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 309–320, 2020.

[10] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[11] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, *Transfer learning*. Cambridge University Press, 2020.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[13] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning*. PMLR, 2013, pp. 222–230.

[14] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[15] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[16] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[17] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[18] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[19] S. Cicek and S. Soatto, "Unsupervised domain adaptation via regularized conditional alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1416–1425.

[20] X. Chen, S. Wang, M. Long, and J. Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2019, pp. 1081–1090.

[21] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2446–2455, 2018.

[22] W. Qian, S. Li, and J. Wang, "A new transfer learning method and its application on rotating machine fault diagnosis under variant working conditions," *IEEE access*, vol. 6, pp. 69 907–69 917, 2018.

[23] M. J. Hasan and J.-M. Kim, "Bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning," *Applied Sciences*, vol. 8, no. 12, p. 2357, 2018.

[24] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 136–144, 2017.

[25] Q. Wang, G. Michau, and O. Fink, "Domain adaptive transfer learning for fault diagnosis," in *2019 Prognostics and System Health Management Conference (PHM-Paris)*. IEEE, 2019, pp. 279–285.

[26] A. Fernández, V. LóPez, M. Galar, M. J. Del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-based systems*, vol. 42, pp. 97–110, 2013.

[27] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[28] X. Li, W. Zhang, Q. Ding, and X. Li, "Diagnosing rotating machines with weakly supervised data using deep transfer learning," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 1688–1697, 2020.

[29] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.

[30] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 2018.

[31] D. She, N. Peng, M. Jia, and M. Pecht, “Wasserstein distance based deep multi-feature adversarial transfer diagnosis approach under variable working conditions,” *Journal of Instrumentation*, vol. 15, no. 06, p. P06002, 2020.

[32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[33] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[34] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

Bingsen Wang received the M.Sc. degree in engineering mechanics from the Dalian University of Technology, Dalian, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Energy, Politecnico di Milano, Milan, Italy.

His current research interests include the development of methods and techniques for prognostics and health management of industrial equipment based on transfer learning and domain adaptation.

Piero Baraldi received the B.S. degree in nuclear engineering and the European Ph.D. degree in radiation science and engineering from the Politecnico di Milano, Milan, Italy, in 2002 and 2006, respectively.

He has been a Full Professor of nuclear engineering with the Department of Energy, Politecnico di Milano since September 2021 and an Associated Professor from 2015 to 2021. He is the author or coauthor of 2 books and more than 200 papers on international journals and proceedings of international conferences. His main research efforts are currently devoted to the development of methods and techniques for system health monitoring, fault diagnostics, prognostics, and maintenance. He is also interested in methodologies for rationally handling the uncertainty and ambiguity in the information.

Dr. Baraldi has been an Invited Keynote Lecture at the plenary sessions of the European Safety and Reliability Conference, ESREL 2014, Wroclaw, Poland, of the 2016 Prognostics and System Health Management Conference, Chengdu, China, and of the 4th International Conference on System Reliability and Safety, ICSRS 2019, Rome, Italy. He has been invited to present 4 tutorials at international conferences. He has been functioning as Technical Programme Chair of the Prognostics and System Health Management Conference, PHM 2013, Milan, Italy, of the ESREL2020PSAM15 Conference, Venice, Italy, and as Technical Committee Co-Chair of the European Safety and Reliability Conference, ESREL 2014, Wroclaw, Poland. He is serving as an Editorial Board Member of 3 international scientific journals and he is an Associated Editor of the *Journal of Risk and Reliability*. He has been an Treasurer of the European Safety and Reliability Association (ESRA) from 2014 to 2018 and he is the Chairman of the ESRA Technical Committee on “Prognostics and System Health Management”.

Enrico Zio (M'06-SM'09) received the M.Sc. degree in nuclear engineering from the Politecnico di Milano, Milan, Italy, in 1991, the M.Sc. degree in mechanical engineering from the University of California, Los Angeles (UCLA), Los Angeles, CA, USA, in 1995, the Ph.D. degree in nuclear engineering from the Politecnico di Milano, in 1996, and the Ph.D. degree in probabilistic risk assessment from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1998.

He is currently a Full Professor with the Centre de Recherche sur les Risques et les Crises (CRC), Mines ParisT-PSL, Sophia Antipolis, France, a Full Professor and the President of the Alumni Association at the Politecnico di Milano, a Distinguished Guest Professor with Tsinghua University, Beijing, China, an Adjunct Professor with the City University of Hong Kong, Hong Kong, Beihang University, Beijing, China, and Wuhan University, Wuhan, China, and the Co-Director of the Center for Reliability and Safety of Critical Infrastructures and the Sino-French Laboratory on Risk Science and Engineering, Beihang University. He is the author or coauthor of 7 books and more than 600 papers on international journals. His current research interests include the modeling of the failure-repair-maintenance behavior of components and complex systems, the analysis of their reliability, maintainability, prognostics, safety, vulnerability, resilience, and security characteristics, and the development and use of the Monte Carlo simulation methods, artificial techniques, and optimization heuristics.

Dr. Zio has been the Chairman and Co-Chairman of several international conferences, a Board Member of several international journals, and a referee of more than 20. He is a Fellow of the Prognostics & Health Management Society and an IEEE and Sigma Xi Distinguished Lecturer. He received the prestigious Humboldt Research Award in 2020. In 2021, he has been appointed as 4TU.Resilience Ambassador by the 4TU Centre for Resilience Engineering of the four Dutch Technical Universities.