This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

Federated Transfer Learning for Intelligent Fault Diagnostics Using Deep Adversarial Networks with Data Privacy

Wei Zhang and Xiang Li

***Abstract*—Intelligent data-driven machinery fault diagnosis methods have been popularly developed in the past years. While fairly high diagnosis accuracies have been obtained, large amounts of labeled training data are mostly required, which are difficult to collect in practice. The promising collaborative model training solution with multiple users poses high demands on data privacy due to conflict of interests. Furthermore, in the real industries, the data from different users can be usually collected from different machine operating conditions. The domain shift phenomenon and data privacy concern make the joint model training scheme quite challenging. To address this issue, a federated transfer learning method for fault diagnosis is proposed in this study. Different models can be used by different users to enhance data privacy. A federal initialization stage is introduced to keep similar data structures in distributed feature extractions, and a federated communication stage is further implemented using deep adversarial learning. A prediction consistency scheme is also adopted to increase model robustness. Experiments on two real-world datasets suggest the proposed federated transfer learning method is promising for real industrial applications.**

***Index Terms*—Fault diagnosis; federated learning; rotating machines; deep learning; data privacy.**

I. Introduction

The past decades have witnessed the great developments of the intelligent data-driven machinery fault diagnostic methods [1]–[4]. While high testing accuracies can be obtained in many case studies, the promising model performance generally relies on the availability of large amounts of labeled training data [5]. In the real industries, due to the significant economic and labor costs, it is usually difficult for a single user to collect sufficient high-quality data for establishing an effective data-driven fault diagnosis model by itself. That impedes the development and application of the promising intelligent techniques.

In order to enlarge the training dataset at lower costs, it is straight-forward to integrate the individual training data of multiple users which share similar operating machines.

The material in this paper is based on work supported by grants (52005086, 11902202) from the National Natural Science Foundation of China, grants (N2005010 and N180708009) from the Fundamental Research Funds for the Central Universities, grants (2020-BS-048 and 2019-BS-184) from Liaoning Provincial Department of Science and Technology, and grant (VCAME201906) from the Key Laboratory of Vibration and Control of Aero-Propulsion System Ministry of Education Northeastern University.

W. Zhang is with School of Aerospace Engineering, Shenyang Aerospace University, Shenyang 110136, China, and Department of Mechanics, Tianjin University, Tianjin 300350, China

X. Li is with Key Laboratory of Vibration and Control of Aero-Propulsion System Ministry of Education, and College of Sciences, Northeastern University, Shenyang 110819, China. E-mail: xiangli@mail.neu.edu.cn

Manuscript received xxx; revised xxx.

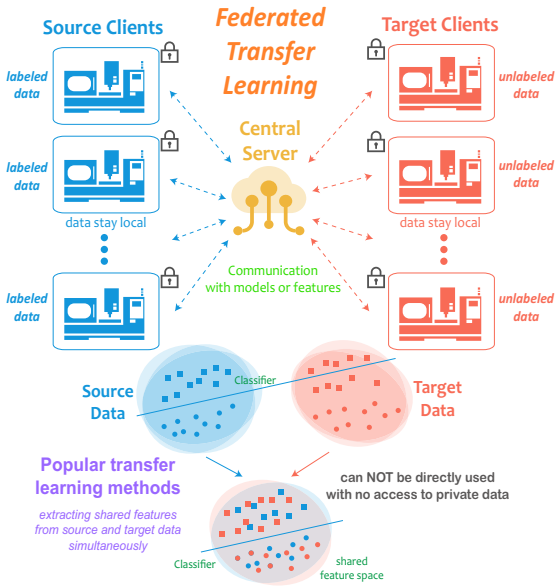


Fig. 1: Illustrations of the federated transfer learning scheme and the existing challenges.

Despite that better model performance can be achieved through exploration of more data, direct data sharing with each other is hardly feasible in practice. The industrial users generally have potential conflicts of interests, and they are basically not comfortable if their private data leave the local storage. Consequently, direct data aggregation is not preferable.

To address this issue, the federated learning strategy offers a promising approach to explore the private data at different users to build a powerful data-driven model, while ensuring the data privacy at the same time. Specifically, multiple users which are also denoted as clients, form a data federation. A central server is usually introduced for coordinating the model updates. Only the models can be communicated, and the private data are safely distributed at the local storage. The data-driven knowledge can be shared under the federation without the risk of private data leakage. Figure 1 shows the scheme of the federated learning scheme.

In most existing federated learning studies, the basic data setting is generally assumed, that the training and testing data are from the same distribution. Under this assumption, exploring more training data from different clients intuitively leads to enhanced model performance on the testing data. However,

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

in the real fault diagnosis industries, the machines of different users are generally working under different conditions. Consequently, the condition monitoring data at different clients are subject to different distributions, that makes it difficult to generalize the knowledge learned from the labeled data at some clients, denoted as source domain, to the unlabeled data at other clients, denoted as target domain. This fact is known as the domain shift phenomenon.

To bridge the domain gap, transfer learning methods have been successfully developed. However, most existing methods assume availability of both the source and target data simultaneously [6]. That is not feasible under the federated learning framework, where the private data at different clients cannot be touched by others. Furthermore, the popular federated learning methods mostly require identical models at different clients. Despite the effectiveness, two disadvantages exist. First, a large number of model parameters are included in the communications, that poses significant burden on the bandwidths. Secondly, the risk of data leakage also increases when the same model is used for different users. Therefore, it is preferred that different models are used by different clients.

In order to address these issues, a novel federated transfer learning strategy is proposed in this paper. The main contributions and novelties are listed as follows,

- The data privacy-preserving federated transfer learning problem in machinery fault diagnosis tasks is investigated, which has been seldomly studied in the current literature.
- Deep adversarial neural networks are proposed to address the federated transfer learning problem, where the domain distribution discrepancy can be effectively bridged with data privacy.
- Different models can be used at different clients. A prediction consistency scheme is proposed to enhance model generalization abilities.
- Experiments on two real-world rotating machinery datasets validate the proposed method, which offers a promising tool for federated transfer learning.

The remainder of this paper starts with the related works in Section II. The preliminaries are presented in Section III. The proposed method is introduced in Section IV, and experimentally validated in Section V. We close the paper with conclusions in Section VI.

II. Related Works

In the past years, intelligent data-driven machinery fault diagnosis problems have largely benefited from the successful development of deep neural networks [7]–[9]. The end-to-end diagnostic framework with automatic feature extraction can be easily established, which takes the condition monitoring data as inputs, and outputs the diagnosis results [10]–[14]. That lowers the application threshold in the real industries and offers high diagnostic accuracies [15]. The vibration and acoustic data have been popularly used for identifying machine health conditions in the literature [16], [17]. Caesarendra *et al*. [18] proposed a parsimonious network based on a fuzzy inference system for time series feature prediction of low speed slew bearing prognosis. The experiments indicate promising results can be obtained using the proposed techniques.

For the basic data setting where the training and testing data are from the same distribution, different variants of neural networks have achieved satisfactory performance [19], including multi-layer perceptron (MLP), convolutional neural networks (CNN), recurrent neural networks (RNN), auto-encoders *etc*. A RNN-based auto-encoder model was proposed by Liu *et al*. [20], where the sequential time series data are explored to capture time relevant knowledge for fault diagnosis. A long short-term memory (LSTM) structure was proposed by Wu *et al*. [21] for machine degradation assessment.

As a practical industrial scenario, the transfer learning problem has also been paid much attention recently, where the training and testing data are collected in different machine operating conditions [22]. The domain adaptation techniques have been effectively applied in such cases. Specifically, the minimization of maximum mean discrepancy (MMD) was used in many related works to bridge the domain gap. Lu and colleagues [23] proposed to achieve domain fusion effects by minimizing MMD between source and target domains. Guo *et al*. [24] further integrated the adversarial learning scheme in the deep neural network framework for fault diagnosis. Through adversarial training, generalized domain-invariant features are learned for health condition classification. Adversarial learning was also used in [25], where invariant features are extracted from multiple source and target domains to achieve cross-machine fault diagnosis. While adversarial learning is an effective tool for domain adaptation, it generally requires explicit processing the data of different domains, and cannot be directly applied in the data privacy-preserving scenarios.

Federated learning has been receiving increasing attention due to its ability of collaborative training with ensured data privacy [26]. A federated averaging method was proposed by McMahan *et al*. [27] to enhance the communication efficiency between different clients and central server. A multi-objective optimization problem in federated learning was formed by Zhu *et al*. [28], where the model architecture, communication cost and global testing error are optimized at the same time.

The more challenging federated transfer learning problem has also been investigated recently. A secure federated transfer learning approach was proposed by Liu *et al*. [29], where the source clients with sufficient labeled data are explored with data privacy. Peng *et al*. [30] proposed to use adversarial domain adaptation technique to address the domain shift effect in federated learning. Sharma *et al*. [31] increased the efficiency and security of the federated transfer learning approach by utilization of secret sharing. This study further extends the federated transfer learning methodology in the machinery fault diagnosis scenario, which has received far less attention in the literature.

III. Preliminaries

A. Federated Transfer Learning Problem

In this study, the federated transfer learning problem is investigated under the machine fault diagnosis scenarios. The main assumptions are presented as follows,

1) Multiple clients with similar machines collaborate to train fault diagnosis models. The source clients hold labeled data, and the target clients have unlabeled data.

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

2) The data distributions at different clients are not the same due to variations in operating conditions.
3) The fault diagnosis task is identical for all the clients, that indicates the same fault mode set is considered.
4) Data privacy must be ensured. The local data of different clients cannot be accessed by others.
5) To further ensure privacy, different clients have different local models for feature extraction.
6) A central server is used for coordinating model optimization, which can touch high-level data representations at the clients.

Let $\mathcal{D}^{s,i} = \left\{\left(\mathbf{x}_j^{s,i}, y_j^{s,i}\right)\right\}_{j=1}^{n^{s,i}}$, $i = 1, 2, ..., N_{client}$ denote the data sample set at the i-th source client, where $\mathbf{x}_j^{s,i} \in \mathrm{R}^{N_{input}}$ and $y_j^{s,i}$ are the data and machine health label of the j-th sample, respectively. N_{client} is the number of the source clients, $n^{s,i}$ denotes the number of the samples in $\mathcal{D}^{s,i}$, and N_{input} is the sample dimension. Without loss of generalization, one target client is investigated in this paper for simplicity. It should be noted that the proposed method can be readily applied on multiple target clients. Let $\mathcal{D}^t = \left\{\mathbf{x}_j^t\right\}_{j=1}^{n_t}$ denote the data sample set at the target client, where n_t is the number of the samples.

This study aims to build a cross-domain fault diagnosis model for the target client, which is trained using the labeled source data $\left\{\mathcal{D}^i\right\}_{i=1}^{N_{client}}$ and the unlabeled target data $\mathcal{D}^t$. Different from most cross-domain fault diagnosis studies, the raw data at different clients cannot be directly accessed by the other clients or the central server. That forms a challenging federated transfer learning problem in the real industries.

B. Domain Adversarial Learning

Adversarial learning has been popularly introduced into the domain adaptation problems in the recent years [32]. In general, a feature extractor F is used to obtain high-level features from raw data in different domains, and a domain discriminator D is adopted to identify the original domains of the input high-level features. Specifically, with respect to the raw data $\mathbf{x}$, the domain discriminator takes $F(\mathbf{x})$ as input, and outputs the predicted domain label $D(F(\mathbf{x}))$. The two modules are optimized adversarially that, the domain discriminator aims to accurately recognize the domains of the data, while the feature extractor is trained to confuse the domain discriminator. In this way, generalized domain-invariant features can be learned.

To ensure learning representative features of different machine health conditions during training, a classifier C can be used for fault classification, which takes $F(\mathbf{x})$ as input. Let θ_F, θ_C and θ_D denote the parameters of F, C and D, respectively. The optimization problem can be formulated as,

$$\begin{aligned} \hat{\theta}_F &= \arg\{\min_{\theta_F} L_c(\hat{\theta}_C, \theta_F), \max_{\theta_F} L_d(\hat{\theta}_D, \theta_F)\}, \\ \hat{\theta}_C &= \arg\min_{\theta_C} L_c(\theta_C, \hat{\theta}_F), \\ \hat{\theta}_D &= \arg\min_{\theta_D} L_d(\hat{\theta}_F, \theta_D), \end{aligned} \tag{1}$$

where L_c denotes fault diagnosis error, and L_d represents the domain prediction error. $\hat{\theta}_F$, $\hat{\theta}_C$ and $\hat{\theta}_D$ denote the optimal values of θ_F, θ_C and θ_D, respectively.

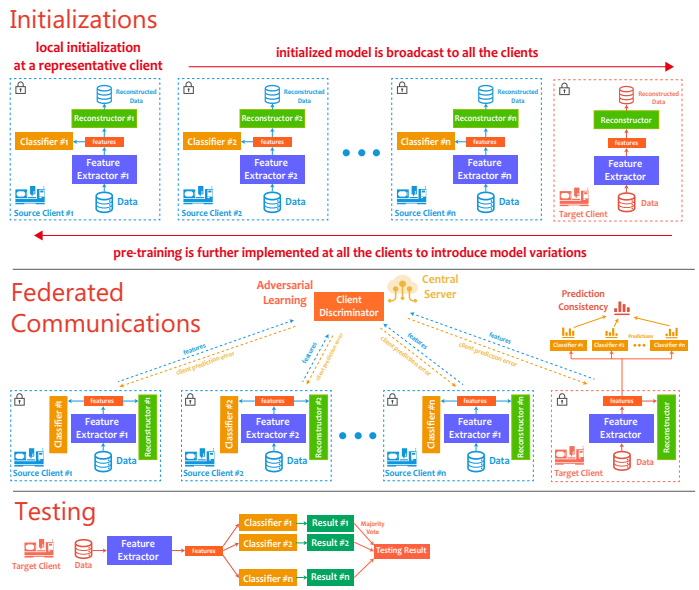


Fig. 2: Overview of the proposed federated transfer learning method in fault diagnosis.

IV. Proposed Methodology

A. Method Overview

Figure 2 shows the overview of the proposed federated transfer learning method. In model training process, an initialization stage is firstly introduced. A representative source client can be selected, and local model initialization is implemented. The local model consists of a feature extractor, a reconstructor and a classifier. The initialized model is then broadcast to the other source and target clients. Local pre-training at all the clients is further carried out to introduce model variations for data privacy.

Next, the federated communication stage starts. In each round, the high-level features of different clients are uploaded to the central server with a client discriminator module, and the server downloads the client prediction error to different clients. All the model parameters are optimized based on the communicated information. Through multiple iterations with adversarial learning, the shared features from different clients can be learned by different feature extractors.

After the training process is finished, the model testing can be readily implemented in the framework. The high-level features of the testing samples at the target client are fed into the source classifiers to obtain the prediction results. The majority vote of all the classifiers is used as the final fault diagnosis result in testing. The specific deep neural network architectures are presented in Figure 3.

B. Initialization

First, all the modules in the proposed framework are initialized. Recall that in this study, different models are used at different clients. While marginal data distribution alignments can be achieved at the high-level representations for transfer learning, overfitting may easily occur, that results in class-level mis-matches. In order to keep the feature structure at the high-level representations for different clients, we propose to initialize all the feature extractors with that of a single

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

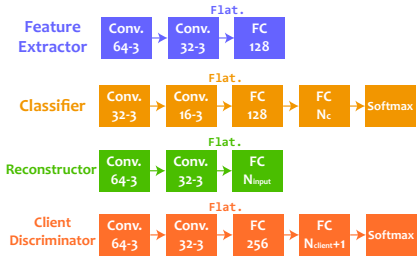


Fig. 3: The deep neural network architectures of different modules in the proposed method.

client. While domain shifts still exist, the conditional data distributions of the health states can be better aligned for different clients.

Specifically, a representative source client can be selected, and local initialization is implemented, including supervised learning and data reconstruction. Without loss of generality, the i-th source client is considered for instance. First, the labeled local data are explored for model training. The cross-entropy based classification loss $L_c^{s,i}$ is minimized, which is defined as,

$$L_c^{s,i} = -\frac{1}{n^{s,i}} \sum_{j=1}^{n^{s,i}} \sum_{k=1}^{N_c} 1\left\{y_j^{s,i} = k\right\} \log \frac{e^{x_{j,k}^{i,hi}}}{\sum_{m=1}^{N_c} e^{x_{j,m}^{i,hi}}}, \quad (2)$$

where N_c denotes the number of the machine health states. $x_{j,k}^{i,hi}$ represents the k-th output value at the last layer in the classifier of the i-th client, which takes the j-th labeled sample as input.

Meanwhile, potential information loss is supposed to be minimized through data transformations in the hidden layers. Therefore, the auto-encoder structure is adopted including the feature extractor and reconstructor modules, and the reconstructor aims to rebuild the original data sample from the high-level features. Specifically, the reconstruction error $L_r^{s,i}$ is minimized, which is defined as,

$$L_r^{s,i} = \frac{1}{n^{s,i} N_{input}} \sum_{j=1}^{n^{s,i}} \left\| R^i(F^i(\mathbf{x}_j)) - \mathbf{x}_j \right\|_1, \quad (3)$$

where F^i and R^i denote the feature extractor and reconstructor modules in the i-th client, respectively. In summary, the general optimization function in local initialization can be formulated as,

$$\min L_{init}^{s,i} = L_c^{s,i} + L_r^{s,i}. \quad (4)$$

After initialization at the i-th client for N_{epoch}^{init} epochs, the same parameters of the three modules are used to initialize the models at the other clients. In this way, similar data structures at the high-level representations can be obtained at different clients. In order to introduce model variations for data privacy, pre-training can be further implemented for all the client models before federal communication. To be concrete, Equation (4) is carried out for N_{epoch}^{pre} additional training epochs at all the source clients respectively, and Equation (3) is used at the target client with unlabeled data for N_{epoch}^{pre} epochs. That leads to different model parameters under the federation, which further enhances data security.

C. Federated Adversarial Learning

After initialization, federated communication starts to achieve knowledge sharing at different clients, and the deep adversarial learning scheme is implemented. Specifically, the central server receives the uploaded high-level features from different clients, and the cross-entropy based client prediction loss L_d can be defined as,

$$L_d = \sum_{i=1}^{N_{client}} L_d^{s,i} + L_d^t,$$

$$L_d^{s,i} = -\frac{1}{n^{s,i}} \sum_{j=1}^{n^{s,i}} \sum_{k=1}^{N_{client}+1} 1\left\{d_j^{s,i} = k\right\} \log \frac{e^{x_{j,k}^{i,dis}}}{\sum_{m=1}^{N_{client}+1} e^{x_{j,m}^{i,dis}}}, \quad (5)$$

$$L_d^t = -\frac{1}{n^t} \sum_{j=1}^{n^t} \sum_{k=1}^{N_{client}+1} 1\left\{d_j^t = k\right\} \log \frac{e^{x_{j,k}^{t,dis}}}{\sum_{m=1}^{N_{client}+1} e^{x_{j,m}^{t,dis}}}.$$

where $L_d^{s,i}$ and L_d^t denote the client prediction loss for the data at the i-th source and the target clients, respectively. $d_j^{s,i}$ denotes the client label of the j-th sample at the i-th source client, and d_j^t is that of the j-th sample at the target client. $x_{j,k}^{i,dis}$ denotes the k-th output at the last layer in the client discriminator that takes the j-th sample at the i-th source client as input and similarly, $x_{j,k}^{t,dis}$ is that with respect to the target client.

In this study, the client discriminator follows the conventional training approach with L_d, since it can get access to the high-level features of different clients. In the perspective of adversarial learning, the local feature extractor can be trained with the local data alone, as presented in Equation (5).

D. Prediction Consistency

In order to improve the model generalization ability, a prediction consistency scheme is proposed in this study. As shown in Figure 2, the testing fault diagnosis results of the target client data are obtained through majority votes on the prediction results by multiple source classifiers. Considering the fact that the source classifiers have different parameters, the prediction consistency generally leads to enhanced model generalization ability with lower risk of overfitting. Therefore, the variations in the fault diagnosis results by the source classifiers are supposed to be minimized. Specifically, the loss function L_p is defined as,

$$\min L_p = \frac{1}{n^t N_{client}} \sum_{i=1}^{n^t} \sum_{j=1}^{N_{client}} \left\| C^{s,j}(F^t(\mathbf{x}_i^t)) - \bar{C}(F^t(\mathbf{x}_i^t)) \right\|_1, \quad (6)$$

$$\bar{C}(F^t(\mathbf{x}_i^t)) = \frac{1}{N_{client}} \sum_{j=1}^{N_{client}} C^{s,j}(F^t(\mathbf{x}_i^t)),$$

where $C^{s,j}$ is the classifier module at the j-th source client. $\bar{C}(F^t(\mathbf{x}_i^t))$ denotes the mean vector of the predictions at the source classifiers, i.e. the representations at the last fully-connected layers.

E. Proposed Optimization Scheme

Through integration of supervised learning, data reconstruction, adversarial learning and prediction consistency schemes,

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

the model optimization problem in federated communication can be formulated as,

$$\begin{aligned}
\hat{\theta}_F^{s,i} &= \arg\{\min_{\theta_F^{s,i}} L_c^{s,i} + L_r^{s,i}, \max_{\theta_F^{s,i}} L_d^{s,i}\}, \\
\hat{\theta}_F^{t} &= \arg\{\min_{\theta_F^{t}} L_r^{t}, \max_{\theta_F^{t}} L_d^{t}\}, \\
\hat{\theta}_R^{s,i} &= \arg\min_{\theta_R^{s,i}} L_r^{s,i}, \hat{\theta}_R^{t} = \arg\min_{\theta_R^{t}} L_r^{t}, \\
\hat{\theta}_C^{s,i} &= \arg\min_{\theta_C^{s,i}} L_c^{s,i} + L_p, \hat{\theta}_D = \arg\min_{\theta_D} L_d,
\end{aligned} \tag{7}$$

where $\theta_F^{s,i}$, $\theta_R^{s,i}$ and $\theta_C^{s,i}$ denote the parameters of the feature extractor, reconstructor and classifier at the i-th source client respectively. θ_F^t and θ_R^t are the parameters of the feature extractor and reconstructor at the target client respectively. $\hat{\theta}_F^{s,i}$, $\hat{\theta}_R^{s,i}$, $\hat{\theta}_C^{s,i}$, $\hat{\theta}_F^t$ and $\hat{\theta}_R^t$ are the optimal values of $\theta_F^{s,i}$, $\theta_R^{s,i}$, $\theta_C^{s,i}$, θ_F^t and θ_R^t respectively.

By using the gradient reversal layer, the optimization problem can be readily solved with the back-propagation optimization method [25]. At each training step, the parameter can be updated as,

$$\theta_F^{s,i} \leftarrow \theta_F^{s,i} - \delta\left(\frac{\partial L_c^{s,i}}{\partial \theta_F^{s,i}} + \frac{\partial L_r^{s,i}}{\partial \theta_F^{s,i}} - \frac{\partial L_d^{s,i}}{\partial \theta_F^{s,i}}\right), \tag{8}$$

$$\theta_R^{s,i} \leftarrow \theta_R^{s,i} - \delta\frac{\partial L_r^{s,i}}{\partial \theta_R^{s,i}}, \tag{9}$$

$$\theta_C^{s,i} \leftarrow \theta_C^{s,i} - \delta\left(\frac{\partial L_c^{s,i}}{\partial \theta_C^{s,i}} + \frac{\partial L_p}{\partial \theta_C^{s,i}}\right), \tag{10}$$

$$\theta_F^{t} \leftarrow \theta_F^{t} - \delta\left(\frac{\partial L_r^{t}}{\partial \theta_F^{t}} - \frac{\partial L_d^{t}}{\partial \theta_F^{t}}\right), \tag{11}$$

$$\theta_R^{t} \leftarrow \theta_R^{t} - \delta\frac{\partial L_r^{t}}{\partial \theta_R^{t}}, \tag{12}$$

$$\theta_D \leftarrow \theta_D - \delta\frac{\partial L_d}{\partial \theta_D}, \tag{13}$$

where δ is the learning rate. The general model training process is summarized in Figure 4. Let θ_i and θ_t denote the general model parameters of the i-th source client and the target client, respectively. θ_D is the parameters of the client discriminator. N_{round} and N_{epoch}^{fed} represent the numbers of the federated training rounds and the local epochs in federated communication, respectively. The detailed training implementations of the proposed method are presented in Algorithms 1 and 2.

It should be noted that while the high-level features of different clients are assumed available under the federation in this study, the data privacy can be still ensured [33], [34]. Using specific types of neural networks, the input data generally cannot be reproduced using the output features, which thus help preserve the data privacy.

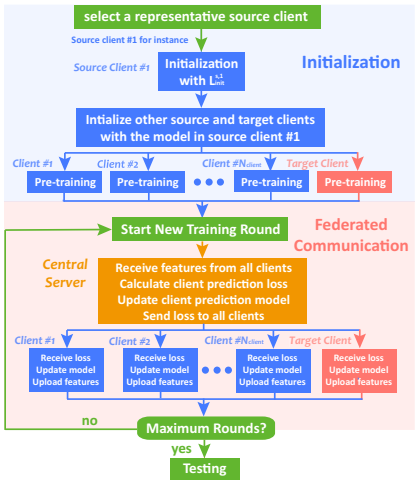


Fig. 4: Flow chart of the model training process in the proposed method.

Algorithm 1: Initialization

Input: Network structures, N_{epoch}^{init}, N_{epoch}^{pre}, N_{client}, δ
Select a representative u-th source client for initialization
for each epoch $i = 1, 2, ..., N_{epoch}^{init}$ **do**
  $\theta_u \leftarrow \theta_u - \delta\frac{\partial L_{init}^{s,u}}{\partial \theta_u}$
end for
Initialize the models at the other clients with θ_u
for each epoch $j = 1, 2, ..., N_{epoch}^{pre}$ **do**
  for each source client i in $1, 2, ..., N_{client}$ **do**
    $\theta_i \leftarrow \theta_i - \delta\frac{\partial L_{init}^{s,i}}{\partial \theta_i}$
  end for
  At target client $\theta_t \leftarrow \theta_t - \delta\frac{\partial L_r^t}{\partial \theta_t}$
end for
Output: Initialized models $\{\theta_i\}_{i=1}^{N_{client}}$ and θ_t

Algorithm 2: Federated Communication

Input: Initialized $\{\theta_i\}_{i=1}^{N_{client}}$ and θ_t, N_{round}, N_{epoch}^{fed}, N_{client}
Client Side:
for each training round $j = 1, 2, ..., N_{round}$ **do**
  for each source client $i = 1, 2, ..., N_{client}$ **do**
    Receive client prediction loss from central server
    Update model with Equations (8), (9) and (10) for N_{epoch}^{fed} epochs
    Upload high-level features of local data to central server
  end for
  Target client receives client prediction loss from central server
  Update target client model with Eq. (11) and (12) for N_{epoch}^{fed} epochs
  Upload high-level features of target client data to central server
end for
Server Side:
for each training round $j = 1, 2, ..., N_{round}$ **do**
  Receive uploaded data features from all clients
  Update client discriminator with Equation (13)
  Send client prediction loss to all clients
end for
Output: Trained models $\{\theta_i\}_{i=1}^{N_{client}}$, θ_t and θ_D

V. Experimental Study

A. Fault Diagnosis Tasks

1) Train Bogie Data

First, a practical rotating machine condition monitoring dataset is used for validations in this study, which is collected from the experimental setup of a high-speed multi-unit train bogie system. The health conditions of the roller bearings are considered. The vibration acceleration data of the component are measured for analysis. Besides the normal bearing state

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

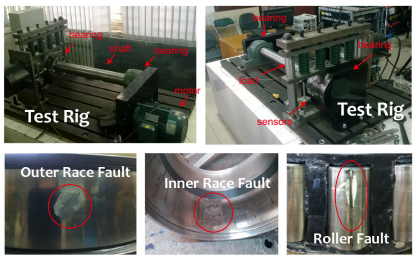


Fig. 5: Experimental setup in the train bogie dataset, and different bearing faults.

TABLE I: Detailed information of the Train Bogie and CWRU datasets.

Train Bogie Dataset			CWRU Dataset		
Label	Fault Mode	Severity	Label	Fault Mode	Fault Size (mil)
B1	N	/	C1	N	/
B2	I	Minor	C2	I	7
B3	I	Medium	C3	I	14
B4	I	Significant	C4	I	21
B5	R	Minor	C5	B	7
B6	R	Medium	C6	B	14
B7	R	Significant	C7	B	21
B8	O	Minor	C8	O	7
B9	O	Medium	C9	O	14
B10	O	Significant	C10	O	21

(N), different fault modes are created at different locations of the bearings, including inner race fault (I), roller fault (R) and outer race fault (O). Three operating conditions with rotating speeds of 1590, 1770 and 1950 rpm are implemented. Furthermore, for each fault mode, three fault severities are also generated as minor, medium and significant levels. In summary, ten bearing health states are considered with three domains. Figure 5 shows the experimental setup in this dataset, as well as the concerned bearing faults. The specific dataset descriptions are presented in Table I.

In this study, based on the dataset information, the federated learning scenario is simulated. Specifically, in accordance with the federated transfer learning problem setting, the non-independent and identically distributed (Non-IID) cases are considered, where the data at different clients are subject to different distributions. Correspondingly, it is assumed that each client as a domain holds the condition monitoring data from each rotating speed respectively. Table II shows the detailed information of different fault diagnosis tasks on the Train Bogie dataset.

TABLE II: Descriptions of the federated transfer learning tasks for machine fault diagnostics in this study. The numbers denote the rotating speeds (rpm) in the datasets.

Train Bogie Data				CWRU Data				
Task Name	Source Clients #1	#2	Target Client	Task Name	Source Clients #1	#2	#3	Target Client
S1	1950	1770	1590	T1	1797	1772	1750	1730
S2	1950	1590	1770	T2	1797	1772	1730	1750
S3	1590	1770	1950	T3	1797	1750	1730	1772
				T4	1730	1750	1772	1797

TABLE III: Parameters used in this study.

Parameter	Value	Parameter	Value
Batch size	256	N_{epoch}^{init}	500
N_{input}	256	N_{round}	2000
δ in Initialization	1e-4	N_{epoch}^{pre}	10
δ in Federated Comm.	1e-5	N_{epoch}^{fed}	1

2) CWRU Dataset

To further evaluate the performance of the proposed method, the popular rolling element bearing dataset from the Case Western Reserve University (CWRU) is also adopted in this study [35]. In general, the CWRU dataset is similar with the train bogie dataset. The vibration acceleration data at the drive end of the motor are collected for analysis. Four bearing health states are considered, including normal state (N), inner race fault (I), ball fault (B) and outer race fault (O). For each fault location, three severities are also created with fault diameters of 7, 14 and 21 mils respectively. Four operating conditions are implemented with rotating speeds of 1797, 1772, 1750 and 1730 rpm respectively. The specific dataset information is shown in Table I. Similarly, multiple fault diagnosis tasks under the federated learning scenarios are investigated, and the detailed information is presented in Table II.

B. Experimental Settings

In this study, the collected raw machinery vibration acceleration data are firstly pre-processed. Specifically, fast Fourier transformation is applied to obtain the frequency spectra, and the min-max scaling approach is used to reduce the effects of the peak values on the subsequent data processing. For each frequency-domain sample $\mathbf{x} = [x_1, x_2, ..., x_{N_{input}}]$, the logarithmic function is applied as,

$$x_i = \ln(x_i - x_{\min} + 1), x_{\min} = \min x_i. \tag{14}$$

The pre-processed samples are used as the model inputs for fault diagnosis afterwards. In each concerned task as presented in Table II, it is assumed that 200 labeled samples of each machine health state are available for training at each source client. For the target client, 200 unlabeled samples of each health state are available for testing. The detailed experimental parameters are listed in Table III. The testing accuracy is generally used as the evaluation metric in this study, which is defined as,

$$Accuracy = \frac{n_{correct}}{n_{test}} \times 100\%, \tag{15}$$

where n_{test} denotes the number of all the testing samples, and $n_{correct}$ represents the number of the testing samples that are correctly classified by the model.

C. Compared Methods

In this study, different approaches are implemented to evaluate the effectiveness and superiority of the proposed method. Specifically, the following methods are carried out for comparisons.

1. Baseline. First, a baseline approach is considered for comparisons, where only the well-established supervised learning paradigm is implemented. Specifically, at the source

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

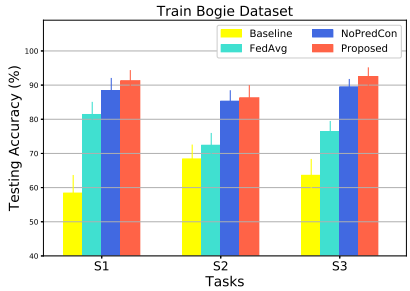


Fig. 6: Testing accuracies by different methods on the fault diagnosis tasks on the train bogie dataset.

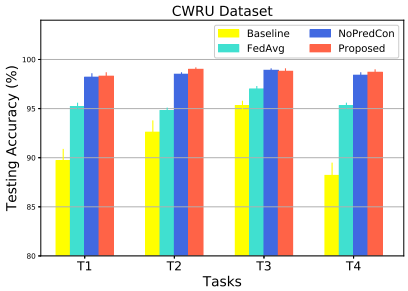


Fig. 7: Testing accuracies by different methods on the fault diagnosis tasks on the CWRU dataset.

client, the private labeled data are used to train the fault diagnosis model locally, and only the classification loss function in Equation (2) is considered in optimization.

2. FedAvg. The federated averaging method is one of the most popular federated learning algorithms in the current literature [27]. In general, the models are trained locally at different clients, and the central server aggregates all the models with the averaging operation at each training round. Afterwards, the obtained global model is sent back to the clients for further local training. Through multiple iterations, the globally optimized model can be obtained.

3. NoPredCon. In order to examine the benefits of the proposed prediction consistency scheme as described in Section IV-D, the NoPredCon method is implemented, where the scheme is removed from the proposed framework. Specifically, the objective L_p in Equation (6) is not considered in model optimization, and the other experimental settings are the same with the proposed method.

D. Experimental Results

Figure 6 shows the testing results of different methods in the concerned fault diagnosis tasks on the Train Bogie dataset. It can be observed that the proposed method significant outperforms the other compared approaches in different cases. The Baseline method basically fails in the cross-domain fault diagnosis problems, and the testing accuracies are generally below 70%. That validates the negative effects of the domain shift phenomenon, and the learned data-driven fault diagnosis knowledge cannot be well generalized across different machine operating conditions. Compared with the baseline, noticeable improvements in the model performance are observed for the FedAvg method, which explores the data from all the source clients to build the global model. The federated averaging method is a widely used approach for federated learning with data privacy, and the knowledge can be effectively learned from multiple source domains. As a result, the global model can be more generalized with higher testing accuracies. However, no knowledge transfer is considered in FedAvg.

On the other hand, the proposed federated transfer learning method explicitly aims at transferring the data-driven knowledge across clients, and the shared features can be learned where no access to local private data is permitted. Therefore, around 90% testing accuracies can be obtained in different tasks. The NoPredCon method is also competitive, while the proposed approach still achieves clear improvements than NoPredCon. That indicates the effectiveness of the proposed prediction consistency scheme in the federated learning framework.

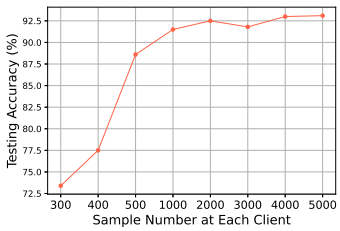


Fig. 8: Effects of the sample number at each client on the model performance in the task S1.

The experimental results on the CWRU dataset are presented in Figure 7. It should be pointed out that the CWRU dataset is relatively easier for fault diagnosis since the fault features are clearer with weak noise. Therefore, the testing accuracies of different methods on the CWRU cases are mostly higher. In general, similar display patterns are still observed with those in the Train Bogie dataset. The proposed framework achieves the best performances in all the cases, and the other approaches are less effective. The results in this section numerically validate the effectiveness and superiority of the proposed federated transfer learning method.

In the experiments in this study, it is generally assumed that 200 samples of each machine health state are available at different clients. Since ten classes are considered in the datasets, a total number of 2000 samples are thus used at each client. Based on the current understanding of deep learning in the literature, more data samples generally lead to better model performance, and the model can be ineffective if insufficient samples are used. In this section, the effects of the sample number on the model performance are evaluated, and the task S1 is considered. Figure 8 shows the experimental results.

It can be observed that in general, the influence of the sample number of the model performance is not very significant when more than 1000 samples are adopted at each client. While slight improvements in testing accuracies are noted with more data, the performance is mostly stable in different cases. On the other hand, when insufficient samples are used such as less than 500 samples at each client, the performance of the

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

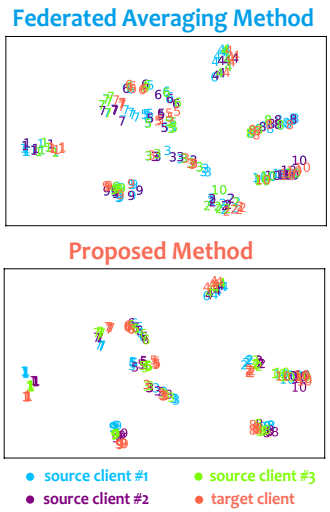


Fig. 9: Feature visualization results by the FedAvg and proposed methods in the task T1. The numbers denote the machine health state labels. For simplicity, the letter 'C' is omitted in the labels.

proposed method remarkably deteriorates. Therefore, it has been validated that the proposed method is generally robust with respect to the sample number in the common scenarios.

E. Visualization Analysis

In this section, the learned features from the feature extractors at different clients are visualized for validations. Specifically, the t-SNE method is used and the feature distributions in the task T1 are presented in Figure 9. It can be observed that using both the methods, different machine health conditions can be well clustered. However, the federated averaging method only focuses on distributed supervised learning for building a global model, and the intra-class compactness of the samples is less satisfactory. That indicates the model robustness against noises and domain shift is not strong.

The proposed federated transfer learning method aims at extracting shared features from multiple clients with data privacy. Through adversarial training, the data of different domains at different clients can be well projected into the same regions in the high-level representation sub-space, that leads to promising intra-class compactness and inter-class separarity. Therefore, the effectiveness of the proposed method is intuitively validated through visualizations in this section.

VI. Conclusions and Discussions

In this study, the data privacy-preserving federated learning problem in machinery fault diagnosis is investigated. Multiple clients collaborate to build effective data-driven models under the condition that the local private data cannot be touched by others. The domain shift phenomenon under the data federation is considered where the data at different clients are collected from different operating conditions, and a novel federated transfer learning method is proposed to achieve knowledge transfer with ensured data privacy. A federal initialization scheme is introduced to ensure extraction of similar features across different clients in data transformations, which is followed by a federated communication stage using deep adversarial learning. A prediction consistency technique is proposed to further enhance the model robustness.

The case studies are implemented for validations, which well simulate the federated transfer learning framework. While the actual data are collected from the same test rig, both data privacy and domain shift issues are strictly considered. For instance, the model at each client cannot touch the data at different clients, and it is only trained in the supervised approach using its own data. The data at different clients are also subject to different distributions, which are practically prepared from different rotating speeds in the cases. Therefore, the data privacy-preserving collaborative learning scenario is well simulated.

As a general methodology, the proposed algorithm can be potentially applied in different kinds of machine condition monitoring scenarios, such as multiple types of rotating machines and the components. For instance, robot arms, rolling bearings, gearboxes, shafts, motors and so forth. Furthermore, different data-driven applications can also benefit from the proposed method, including collaborative image processing tasks, audio analysis problem, video pattern learning cases *etc*.

For potential practical implementations, the central server can be at the cloud, and the communications between the central server and different clients can be through the industrial internet of things (IoT) devices. The central server is supposed to have sufficient memory and computing capabilities in order to process model optimization under the data federation. Since this paper is mostly focused on the methodology part, more investigations should be carried out in terms of actual deployment in the real industrial scenarios.

Despite the promising results, one of the main limitations of this study is the assumption of identical machine health state sets at all the clients. In the real industries, the data at different clients can be possibly collected from different fault modes, and the alignment of marginal distributions may not work well with mis-matched label spaces using the current methodology. While some latest works have paid attention to this challenging problem and achieved promising results [36], the data privacy issue under the framework has not been considered in the literature. A potentially promising solution lies in the introduction of additional weight items for different fault modes, in order to achieve conditional domain adaptation without exposing local data. However, that still requires sufficient validation for research purpose.

Moreover, the recent researches have also focused on the cross-dataset fault diagnosis problem [24], where the source and target data are from different test beds. This scenario is also possible in the federated learning framework. However, since the same machine health sets are assumed in this paper, this practical case cannot be directly addressed using the proposed method. The aforementioned challenges will be further investigated in the following research works.

References

[1] A. Abid, M. T. Khan, H. Lang, and C. W. d. Silva, "Adaptive system identification and severity index-based fault diagnosis in motors," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 4, pp. 1628–1639, 2019.

[2] G. Jiang, P. Xie, H. He, and J. Yan, "Wind turbine fault detection using a denoising autoencoder with temporal information," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 1, pp. 89–100, 2018.

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2021.3065522, IEEE/ASME Transactions on Mechatronics

[3] W. Zhang, X. Li, H. Ma, Z. Luo, and X. Li, "Open set domain adaptation in machinery fault diagnostics using instance-level weighted adversarial learning," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2021.

[4] K. Yu, T. R. Lin, H. Ma, X. Li, and X. Li, "A multi-stage semi-supervised learning approach for intelligent fault diagnosis of rolling bearing using data augmentation and metric learning," *Mechanical Systems and Signal Processing*, vol. 146, p. 107043, 2021.

[5] J. Wang, W. Qiao, and L. Qu, "Wind turbine bearing fault diagnosis based on sparse representation of condition monitoring signals," *IEEE Transactions on Industry Applications*, vol. 55, no. 2, pp. 1844–1852, 2019.

[6] X. Wang, H. He, and L. Li, "A hierarchical deep domain adaptation approach for fault diagnosis of power plant thermal system," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 9, pp. 5139–5148, 2019.

[7] W. Zhang, X. Li, and X. Li, "Deep learning-based prognostic approach for lithium-ion batteries with adaptive time-series prediction and on-line validation," *Measurement*, vol. 164, no. 108052, 2020.

[8] X. Li and W. Zhang, "Deep learning-based partial domain adaptation method on intelligent machinery fault diagnostics," *IEEE Transactions on Industrial Electronics*, vol. DOI: 10.1109/TIE.2020.2984968, pp. 1–1, 2020.

[9] J. Wang, G. Du, Z. Zhu, C. Shen, and Q. He, "Fault diagnosis of rotating machines based on the EMD manifold," *Mechanical Systems and Signal Processing*, vol. 135, p. 106443, 2020.

[10] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," *ISA Transactions*, vol. 95, pp. 295–305, 2019.

[11] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, 2020.

[12] M. Xia, T. Li, L. Xu, L. Liu, and C. W. d. Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 1, pp. 101–110, 2018.

[13] J. Dai, J. Wang, W. Huang, J. Shi, and Z. Zhu, "Machinery health monitoring based on unsupervised feature learning via generative adversarial networks," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 5, pp. 2252–2263, 2020.

[14] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," *Measurement*, vol. 152, p. 107377, 2020.

[15] X. Li, X. Jia, Y. Wang, S. Yang, H. Zhao, and J. Lee, "Industrial remaining useful life prediction by partial observation using deep learning with supervised attention machinery," *IEEE/ASME Transactions on Mechatronics*, 2020.

[16] A. Glowacz, "Recognition of acoustic signals of commutator motors," *Applied Sciences*, vol. 8, no. 12, p. 2630, 2018.

[17] A. G. et al., "Detection of deterioration of three-phase induction motor using vibration signals," *Measurement Science Review*, vol. 19, no. 6, pp. 241–249, 2019.

[18] W. Caesarendra, M. Pratama, B. Kosasih, T. Tjahjowidodo, and A. Glowacz, "Parsimonious network based on a fuzzy inference system (PANFIS) for time series feature prediction of low speed slew bearing prognosis," *Applied Sciences*, vol. 8, no. 12, 2018.

[19] C. Lu, Z. Y. Wang, W. L. Qin, and J. Ma, "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification," *Signal Processing*, vol. 130, pp. 377–388, 2017.

[20] H. Liu, J. Zhou, Y. Zheng, W. Jiang, and Y. Zhang, "Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders," *ISA Transactions*, vol. 77, pp. 167–178, 2018.

[21] Y. Wu, M. Yuan, S. Dong, L. Lin, and Y. Liu, "Remaining useful life estimation of engineered systems using vanilla LSTM neural networks," *Neurocomputing*, vol. 275, pp. 167–179, 2018.

[22] P. Liang, C. Deng, J. Wu, G. Li, Z. Yang, and Y. Wang, "Intelligent fault diagnosis via semi-supervised generative adversarial nets and wavelet transform," *IEEE Transactions on Instrumentation and Measurement*, pp. 1–1, 2019.

[24] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 9, pp. 7316–7325, 2018.

[23] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 3, pp. 2296–2305, 2017.

[25] 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, 2019.

[26] W. Zhang, X. Li, H. Ma, Z. Luo, and X. Li, "Federated learning for machinery fault diagnosis with dynamic validation and self-supervision," *Knowledge-Based Systems*, vol. 213, p. 106679, 2021.

[27] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[28] H. Zhu and Y. Jin, "Multi-Objective evolutionary federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2019.

[29] Y. Liu, T. Chen, and Q. Yang, "Secure federated transfer learning," *arXiv preprint arXiv:1812.03337*, 2018.

[30] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, "Federated adversarial domain adaptation," *arXiv preprint arXiv:1911.02054*, 2019.

[31] S. Sharma, C. Xing, Y. Liu, and Y. Kang, "Secure and efficient federated transfer learning," *arXiv preprint arXiv:1910.13271*, 2019.

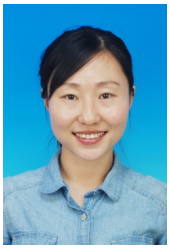
[32] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2962–2971, 2017.

[33] W.-C. Ma, S. Wang, J. Gu, S. Manivasagam, A. Torralba, and R. Urtasun, "Deep feedback inverse problem solver," in *Proceedings of ECCV*, pp. 229–246, Cham, 2020.

[34] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[35] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mechanical Systems and Signal Processing*, vol. 64-65, pp. 100–131, 2015.

[36] J. Jiao, M. Zhao, J. Lin, and C. Ding, "Classifier inconsistency based domain adaptation network for partial transfer intelligent diagnosis," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2020.



Wei Zhang received the Ph.D. degree in mechanics from Tianjin University in 2017. She was a visiting scholar with Texas A&M University, US in 2015. She is currently an assistant professor with School of Aerospace Engineering, Shenyang Aerospace University, China, and a postdoc with Tianjin University, China. Her research interests include rotor dynamics, multi-objective optimization algorithm, and squeeze film damper.

Xiang Li received the B.S. degrees in engineering mechanics and engineering management and the Ph.D. degree in mechanics from Tianjin University in 2012 and 2017, respectively. He was a visiting scholar with University of California at Merced, US in 2015. He is currently an associate professor with Northeastern University, China. His research interests include industrial AI, deep learning, fault diagnosis, prognostics and health management, and multi-objective optimization algorithm.

1083-4435 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 27,2021 at 15:36:23 UTC from IEEE Xplore. Restrictions apply.