FedAlign: Federated Model Alignment via Data-Free Knowledge Distillation for Machine Fault Diagnosis

Wenjun Sun, Ruqiang Yan, *Fellow, IEEE,* Ruibing Jin, Rui Zhao, and Zhenghua Chen, *Senior Member, IEEE*

***Abstract*—Due to privacy issues, the data island problem of machine fault diagnosis widely exists in real industry. Federated Learning (FL) has received much attention as a decentralized machine-learning paradigm that learns a global model in the server by iteratively aggregating the local model parameters in a privacy-preserving scheme to address the data island problem. However, the fault data in industry is inevitably scarce and class imbalanced, so that the fault data in different clients are heterogeneous from each other in FL. The existing federated machine fault diagnosis methods are mostly based on the federated averaging (FedAvg) algorithm, which ignore the data heterogeneity issue for machine fault diagnosis. To address this issue for machine fault diagnosis, we propose a new model alignment method (called FedAlign) via data-free knowledge distillation, where a compact generator is trained in the server in a data-free manner to estimate the data feature distribution in a global view without accessing to the local data. Then the generator can produce the pseudo features to convey the distribution knowledge to both the server and the clients sides. Our proposed FedAlign utilizes the pseudo features to align the global model and the local models to finally address the inherent heterogeneity of local data for fault diagnosis in FL. Experiments performed on fault diagnosis datasets of non-iid settings indicate that our proposed method facilitates FL for machine fault diagnosis with favorable effectiveness and achieves significant performance gains compared with state-of-the-art methods.**

***Index Terms*—data-free knowledge distillation, data heterogeneity, fault diagnosis, federated learning, generator.**

I. Introduction

MACHINE health monitoring and fault diagnosis are indispensable processes in modern industry to ensure the safety of machinery for efficient manufacturing. With the rapid development of sensing technology, intelligent machine fault diagnosis based on deep learning approaches [1]–[4] has achieved great success in modern industrial applications. Although these data-driven approaches have brought promising performance gains, they heavily rely on a large amount of high-quality labeled data for supervised learning. However, it is challenging to collect such a large number of high-quality labeled data [5], [6], which is labor-intensive and time-consuming for a company. And a small and insufficient dataset in a company may cause unsatisfactory diagnosis performance, which is known as the data island problem. This problem limits the application of intelligent fault diagnosis methods in industry.

Fortunately, the types of machines for industrial manufacturing in some companies are similar and these machines generally suffer from similar fault diagnosis issues. Under this scenario, we expect to collect the data from these companies to establish a large-scale dataset for model training. However, it is impossible for data sharing between different companies due to privacy concerns and the conflicts of interest between companies.

Recently, Federated Learning (FL) [7]–[9] has become a cooperative and decentralized learning paradigm to train models in a privacy-preserving scheme. It aims to develop a global model on a centralized server, while keeping the local data distributed over the local clients safely and privately. FL offers a practical solution to data island problems in various fields, such as healthcare [10], smart city [11], and edge computing [12], etc. FL recently has also received much attention in fault diagnosis [13]–[15] for the data island problem in machine fault diagnosis.

Despite the promising prospect of FL, it remains challenges to apply FL to machine fault diagnosis in real industry. One major challenge is caused by data heterogeneity [16]–[18], where the data of clients in industry are almost non-identically and independently (non-iid) distributed. In addition, the class imbalance [5], [19], data scarcity [5], [20] and domain shift caused by varied operating conditions [21], [22] in machine fault diagnosis further aggravate the data heterogeneity among clients, leading to large differences between local data distributions. Accordingly, the corresponding feature distributions of local data in clients are different and also different from that of the global data, which causes the model drift issue [23] and inherently deflects the local optimization away from the global optimization [17], [18], [23]. As a result, the classical FL algorithm FedAvg [7], which simply performs the local model aggregation, may lead to a drifted global model with degraded performance in non-iid scenarios [17], [18], [23].

There have existed approaches [13]–[15] to address the data island problem for fault diagnosis in federated learning.

This work was supported in part by the National Natural Science Foundation of China under Grant 51835009 and in part by the National Research Foundation of Singapore, AME Young Individual Research Grant A2084c0167. (Corresponding author: Ruqiang Yan; Ruibing Jin.)

Wenjun Sun is with the School of Instrument Science and Engineering at the Southeast University, Nanjing 210096, Jiangsu, China. (e-mail: swjstudent@163.com)

Ruqiang Yan is with the School of Instrument Science and Engineering at the Southeast University, Nanjing 210096, Jiangsu, China, and the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China. (e-mail: ruqiang@seu.edu.cn)

Ruibing Jin and Zhenghua Chen are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632. (e-mail: jin_ruibing@i2r.a-star.edu.sg, chen0832@e.ntu.edu.sg)

Rui Zhao is with the Company of Pluang, #10-55 1 George Street 049145, Singapore. (e-mail: rui91seu@gmail.com)

However, most of these approaches are the applications of FedAvg [7] in machine fault diagnosis, which could easily lead to the drifted model with degraded diagnostic performance in non-iid scenarios. These methods do not address the data heterogeneity issues well for fault diagnosis in FL. Recently, massive researches in FL for image and text tasks [24]–[26] focus on the challenge of data heterogeneity, which can mainly be divided into two complementary perspectives. One mainly pays attention to the local training [24], [26] by regulating the deviation of local model from the global model. Nevertheless, these methods ignore the knowledge forgetting and generalization problem in the global model after model aggregation. The other focuses on improving the model aggregation, where knowledge distillation [27], [28] is utilized as a solution. However, these approaches heavily rely on a proxy dataset for performance improvement. The proxy dataset generally refers to an additional public dataset used in FL and its distribution should be consistent to that of the specific task. Due to the lack of the related public dataset in industrial scenarios, it is usually infeasible to adopt the proxy dataset in FL for fault diagnosis.

In FL settings, the local data of clients are also not accessible to each other and the server. It is challenge to distill the knowledge in each client to the other clients and the server when there is no proxy dataset shared among clients and the local data in clients are not accessible. Thus, we propose a new approach, namely **Fed**erated **Align**ment via Data-Free Knowledge Distillation in this paper to tackle the data heterogeneity challenge for machine fault diagnosis. Our proposed FedAlign develops a data-free knowledge distillation method, which trains a compact generator in the server to learn the global data distribution in a data-free manner without accessing to the local data and generate the pseudo features in a global view for knowledge distillation instead of using a proxy dataset. The generator in our proposed FedAlign will be utilized as a bridge for producing the pseudo features to convey the learned distribution knowledge to both the client and server sides for model alignment to solve the model drift and the global knowledge forgetting issues. With our proposed FedAlign, the gap between the local feature distributions and global feature distribution can be alleviated, and the model drift and knowledge forgetting issues in FL caused by data heterogeneity are addressed.

To clearly show the advantages of our FedAlign, we compare our FedAlign with the classic method FedAvg [7] and illustrate the data boundary produced by these two approaches in Fig. 1. In our FedAlign, with the help of pseudo features generated by our proposed generator, the local models can observe beyond its own training data and regulate its prediction boundaries to approach the true boundary for model alignment, meanwhile, the global model in the server can also be refined with the pseudo features to further reduce the boundary discrepancy for model alignment. Compared with the boundary produced in FedAvg, the boundary produced by our FedAlign is closer to the true boundary. This demonstrates that our FedAlign can heavily alleviate the inherent heterogeneity of local data, and improves the performance of the global model for machine fault diagnosis. By utilizing the generator as a bridge, our proposed FedAlign links and aligns the global and the local models without releasing private local data.

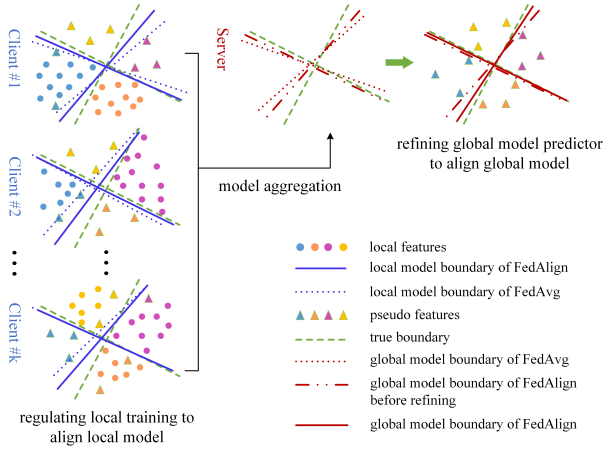

Fig. 1. Diagram of how our FedAlign solves the issue of data distributions. Models trained with different data will predict different boundaries. In our FedAlign, pseudo features are produced and used to regulate the local training and refine the global model to align the models. After introducing these pseudo features, the gap between local data and global data is alleviated and both models are aligned. With our FedAlign, the boundaries predicted by models are improved and approach the true boundary.

Apart from it, our proposed method can be combined with homomorphic encryption [29] methods to ensure that the information in the uploading and downloading phases does not leak any local data privacy. *To the best of our knowledge*, this work is the first to investigate the data-free knowledge distillation for machine fault diagnosis in federated learning.

The main contributions can be summarized as follows:

1) Instead of using a proxy dataset, a compact and convolutional generator is trained in the server to learn the data distributions from the local models in a data-free distillation scheme, and then produce the pseudo features in a global view to convey the distribution knowledge.
2) FedAlign addresses data heterogeneity issue by aligning models from both local training and global model aggregation perspectives for federated machine fault diagnosis, where the generator is utilized as a bridge for producing the pseudo features to augment the local data for local model alignment and distillate knowledge to global model for global model alignment.
3) Comprehensive experiments and analyses on machine fault diagnosis in FL settings have been performed. The results show that our proposed FedAlign achieves significant performance gains for machine fault diagnosis in non-iid FL settings.

The paper is organized as follows. In Section II, the related works are introduced. Then, in Section III, the data-free knowledge distillation methodology in the proposed FedAlign is presented. FL experimental setup and results are provided in Section IV. At last, the conclusion is provided in Section V.

II. Related Works

A. Fault Diagnosis in Federated Learning

Federated learning (FL) [7], which emerges as a decentralized machine-learning paradigm preserving data privacy among different parties, has attracted increasing attention and provided a solution to the data island problem in machine fault diagnosis. Zhang et al. [13] were the first to apply federated learning to machine fault diagnosis and proposed a method based on the FedAvg algorithm. Ma et al. [14] proposed an asynchronous update paradigm of FL to realize the real-time update of clients' network parameters with the sequential Kalman filter and the extended Kalman filter for fault diagnosis. Du et al. [30] utilized the forgetting Kalman filter (FKF) combined with cubic exponential smoothing (CES) to improve the efficiency of federated learning. Zhang et al. [31] proposed a federated transfer learning method using prior distributions to reduce the domain discrepancy between clients for industrial machinery fault diagnostics. Lu et al. [32] proposed a federated learning method for class-imbalanced fault diagnosis on a decentralized wind turbine with the balanced auxiliary dataset in the server. Jiang et al. [8] designed a multi-scale residual attention network model in federated learning for fault detection of wind turbines. Mehta et al. [33] proposed a duplet classifier to develop the federated learning framework for the diagnosis of mixed faults.

Although the above approaches have been proposed for federated machine fault diagnosis, most of them are the applications of FedAvg and consider the relatively ideal FL scenarios, neglecting the diverse non-iid scenarios in the real industry. Different from the existing FL methods, our proposed FedAlign aims to address the data heterogeneity challenge in FL and solve the model drift and knowledge forgetting issues of FedAvg when learning in non-iid FL settings.

B. Knowledge Distillation in Federated Learning

Knowledge distillation [34] is introduced to learn a student model by distilling knowledge from several powerful teacher models. In FL, the knowledge distillation based FL methods [27], [28], [35] usually required a proxy dataset to deliver the ensemble knowledge from local models. FedMD [28] learned the knowledge through averaged logits on a public dataset to train local models on both labeled public and local datasets. FedDF [27] utilized an unlabeled dataset to distill ensemble knowledge for model fusion. MHAT [35] distilled its server model by training it on a labeled proxy dataset with the soft-labeled version. With the help of the proxy dataset, knowledge distillation methods alleviate the model drift issue caused by heterogeneity, enriching the model with knowledge from local models. However, this proxy dataset requires careful consideration and prior knowledge of local private data, which is usually infeasible for most applications. Recently, FedGen [25] utilized a data-free method to distill knowledge to regulate the local model training. However, it only focuses on local model training and ignores the knowledge forgetting problem in the global model.

In comparison, our proposed FedAlign does not require the public proxy dataset and it can learn the global knowledge from the local models in a data-free manner. Moreover, our proposed FedAlign distills knowledge to both the local models and the global model, together improving the quality of the global model for machine fault diagnosis.

III. Methodology

A. Problem Setting

In the paper, the federated learning problem in machine fault diagnosis is investigated and the main assumptions for the problem are presented as follows:

1) Multiple clients with similar machines participate in the collaborative training of the federated learning system to obtain a robust global model.
2) There are several common fault issues between clients. A union set of fault labels is applied to these clients. The fault diagnosis model adopted by clients and the server is the same.
3) Each client has its own local data, which is insufficient and imbalanced in categories among different clients.
4) The local data of each client is private and cannot be communicated with others.

For machine fault diagnosis in FL scheme with K clients, each client owns its local private dataset $\mathcal{D}_k = \left\{(x_{k,j}, y_{k,j})_{j=1}^{|\mathcal{D}_k|}\right\}$, $y \in \{1, 2, \cdots, C\}$, where x and y denote the input data samples and the corresponding class labels, respectively. $|\mathcal{D}_k|$ denotes the number of samples and C is the class number. The goal of our paper is to learn a global model over the whole training data $\mathcal{D} \triangleq \bigcup \{\mathcal{D}_k\}_{k=1}^{K}$ in the server without accessing local data. Usually, in FL, clients and the server share the same model architecture. The fault diagnosis model parameterized by $\theta := \left[\theta^f, \theta^p\right]$ consists of two modules, i.e. feature extractor f parameterized by θ^f and the predictor g parameterized by θ^p. The model in the k-th client is defined as $F(\theta_k) = g\left(f\left(x; \theta_k^f\right); \theta_k^p\right)$ and the global model in the server is $F(\theta) = g\left(f\left(x; \theta^f\right); \theta^p\right)$. The basic FL algorithm FedAvg [7] performs a weighted average for local model aggregation to update the global model in the server and this process can be presented as follows,

$$\theta = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} \theta_k. \tag{1}$$

In the training process, for each communication round t, the server sends the global model parameters θ to a set of clients $\mathcal{A}$. The client trains and updates its model parameters θ_k. Then, all clients send their local models $\{\theta_k\}_{k \in \mathcal{A}}$ back to the server to update the global model parameters. However, this simple model aggregation as Eq. (1), may lead to model drift and performance degradation in data heterogeneity scenarios [23].

To address this issue, we propose a FedAlign method to align both global and local model via a data-free manner, improving the quality of the aggregated model. The framework of our proposed FedAlign is shown in Fig. 2, and its training process is elaborated in Algorithm 1.

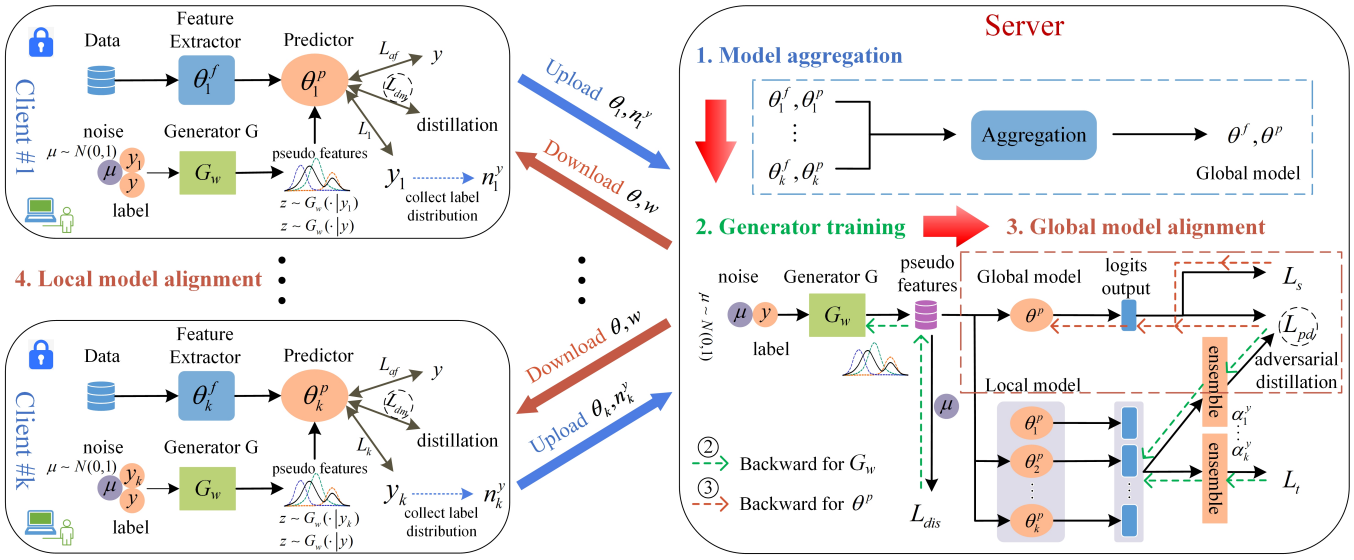

Fig. 2. The framework of our proposed FedAlign method. First, the clients initialize the local models with received global model parameters and train the local models. After that, they upload the local models and local label distributions to the server. Then, the server aggregates the local models in the communication round to update the global model. Next, it trains the generator G with label distributions to generate the pseudo features in a data-free distillation scheme, where the generator is simultaneously adversarially trained with the global model through L_{pd} loss for distilling knowledge to the global model. After that, the global model in the server can be refined with the pseudo features produced by the generator G for global model alignment. Finally, the server broadcasts the trained generator G and the global model to clients. In clients, the generator G can produce the pseudo features to augment and match the local features for local model alignment. With our proposed FedAlign, the global model and the local models are both aligned to improve the performance of the federated global model.

B. Data-free Knowledge Distillation in FedAlign

In our proposed FedAlign, a conditionally convolutional generator G is designed and trained in the server for data-free knowledge distillation. Concretely, the generator G is trained to estimate the data feature distribution in a global view from local models with the local label distributions in a data-free manner. Then, the generator G is able to produce the pseudo features which is consistent with the global data features to convey the knowledge. As shown in Fig. 2, the trained generator not only produces pseudo features in the server to distill knowledge to accurately refine the global model predictor, but also is broadcasted to the clients to produce pseudo features as data augmentation to regulate the local model training. The pseudo features in both server and the clients alleviate the data heterogeneity issues, such as data class imbalance and data scarcity issues, for machine fault diagnosis in FL, guiding the federated training to a robust global model.

1) Learning a Generator to Estimate the Data Distribution in a Global View: Traditional generative adversarial nets (GANs) [36] need raw data for training to generate complex and indistinguishable samples like the raw ones. By comparison, our proposed generator is trained with label distributions in a data-free distillation [37], [38] scheme, without accessing to the raw data. Specifically, the generator in our proposed FedAlign is trained in a data-free manner to learn from the local models (teacher model) and then produce the pseudo data to convey the knowledge learned from the local models to other student model. In a way, the local models can be viewed as a fixed discriminator in traditional GANs. Inspired by the generative models [39], [40], we designed a compact generator G with convolutional architecture in our proposed FedAlign for learning the data distributions in a data-free manner and producing the pseudo data in a global view for data-free knowledge distillation.

Different from images, time series data usually lacks the inherent relationship and is even affected by random noise, which heavily affects the raw data generation. To address this issue, instead of producing raw data, our proposed FedAlign aims to generate pseudo features, which enables the generator G to be compact and useful for generating abstract features.

Fig. 3 shows the architecture of generator G. Firstly, the input is projected to a latent space, where a high dimensional feature vector is produced. Then, two consecutive and convolutional layers are used to convert this high dimensional feature representation into a feature vector. With this architecture, the generator G is trained to produce highly qualified pseudo features for knowledge distillation.

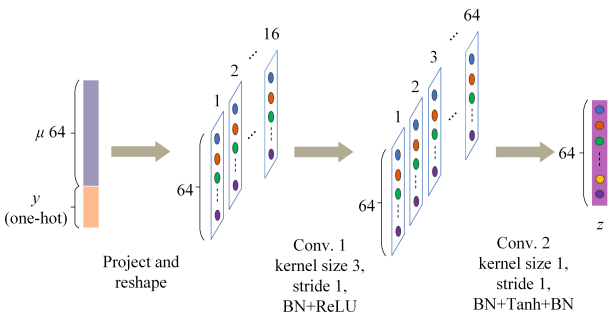

Fig. 3. The architecture of generator G.

In our FedAlign, this proposed generator G is trained in

the server. To capture the data distributions in clients, our generator is used to produce pseudo features as follows,

$$z = G(\mu, y; w), \tag{2}$$

where $\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ represents a noise vector, y indicates the one-hot class label vector arbitrarily sampled from the statistical label distribution $\hat{P}(y)$, w indicates the parameters of generator G, and z is the generated pseudo features. $\hat{P}(y)$ is obtained by counting the local label distributions from clients. The pseudo features $z \sim G_w(\cdot\,|y)$ generating process is similar to the reparameterization of label y [41].

Algorithm 1 FedAlign learning

INPUT: Communication round T; Client number K; Datasets of clients $\{\mathcal{D}_k\}_{k\in\{1,\cdots,K\}}$; Global model parameters θ; Local model parameters $\{\theta_k\}_{k=1}^{K}$; Generator parameters w; Local steps, generator steps and global steps I_l, I_d, I_g; Label distribution sampling n_k^y;
OUTPUT: Global model θ.

Initialize model parameters θ and w;
for t=1 **to** T **do**
  $\mathcal{A} \leftarrow$ (a set of all K clients);
  // Clients execute:
  for client $k \in \mathcal{A}$ in parallel **do**
    $\theta_k \leftarrow \theta$;
    for i=1 **to** I_l **do**
      $w \leftarrow w$
      Update label distribution sampling n_k^y;
    $\theta_k \leftarrow \mathrm{Update}\,(\theta_k; \mathcal{D}_k, G_w)$ by Eq.(13);
    end for
    Client sends θ_k, n_k^y back to server.
  end for
  // Server execute:
  Update global model $\theta = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|}\theta_k$, and compute $\hat{P}(y)$, α_k^y based on n_k^y according to Eq. (15) and Eq. (16), respectively;
  for i=1 **to** I_d **do**
    Sample a batch of (μ, y) from $\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and $y \sim \hat{p}(y)$;
    Train the generator and update w according to Eq. (7);
  end for
  for j=1 **to** I_g **do**
    $(z, y) \leftarrow G_w(z\,|y)$ generate pseudo features;
    Refine the global predictor according to Eq. (9);
  end for
  Send θ and w to clients;
end for

To estimate the global data distribution by using the label distributions of clients, our generator G learns from the ensemble knowledge of local models to generate pseudo features that are similar to the data features from a global view. This process can be formulated as follows.

$$\min_{w} \mathbb{E}_{y\sim\hat{p}(y)}\mathbb{E}_{z\sim G_w(z|y)}\left[L_t\right], \tag{3}$$

where L_t is the semantic loss and can be computed in Eq. 4.

$$L_t = \sum_{k\in\mathrm{A}} \alpha_k^y L_{ce}(\sigma(g(z;\theta_k^p)), y), \tag{4}$$

where L_{ce} indicates the cross-entropy loss, θ_k^p is the parameters of the model predictor in client k, $g(\cdot)$ represents the logits output of a predictor, and σ is the softmax function. In addition, α_k^y is the weight for different local models, which can be computed through the local label distributions and more details are provided in Sec. III-C. By minimizing L_t, the local data distribution in the client is combined and our generator can learn the data distribution from a global view.

For better performances, based on adversarial distillation [38], we use the adversarial distillation to further train our generator. During this process, our generator produces the pseudo features, which are forwarded to the global model and the local models. The optimize object is defined as follows.

$$\begin{aligned} &\max_{w}\min_{\theta^p} \mathbb{E}_{y\sim\hat{p}(y)}\mathbb{E}_{z\sim G_w(z|y)}\left[L_{pd}\right], \\ &\text{where } L_{pd} = D_{KL}\left(\sigma\left(g(z;\theta^p)\right)\middle\|\sigma\left(\sum_{k\in\mathcal{A}}\alpha_k^y g(z;\theta_k^p)\right)\right) \end{aligned} \tag{5}$$

where L_{pd} is the prediction discrepancy loss between the global model predictor θ^p and the local model predictors θ_k^p, D_{KL} denotes the Kullback-Leibler divergence. By maximizing L_{pd}, the generator can explore pseudo features in latent space whose distribution is matched with the global feature distribution for the global model.

Apart from it, we propose a diversity loss L_{dis} to increase the diversity of pseudo features and avoid the model collapse according to [37], [42]. This loss is formulated as follows,

$$L_{dis} = e^{\frac{1}{b*b}\sum_{i,j\in\{1,\cdots,b\}}\left(-\|z_i - z_j\|_1 * \|\mu_i - \mu_j\|_1\right)}, \tag{6}$$

where b indicates the number of training samples and z_i represents the pseudo feature generated from μ_i.

In each training round, the total optimization objective of training the generator G can be presented as,

$$\min_{w} \mathbb{E}_{y\sim\hat{p}(y)}\mathbb{E}_{z\sim G_w(z|y)}\left[L_t - L_{pd} + L_{dis}\right]. \tag{7}$$

2) Global Model Alignment: In federated learning, global model may suffer from knowledge forgetting issue in model aggregation. To mitigate this issue, we use the pseudo features from our generator G to convey knowledge to align the global model, further alleviating the model drift and knowledge forgetting issues.

The alignment of the global model consists of two parts: pseudo feature classification and knowledge distillation. In the first part, the semantic loss L_s is used to optimize the global model predictor with the pseudo features, which is defined as below.

$$L_s = L_{ce}(\sigma(g(z;\theta^p)), y). \tag{8}$$

For the knowledge distillation, a prediction discrepancy loss L_{pd} is applied to distill knowledge to global model for matching the feature distributions. The overall objective of refining the global model predictor in the server is formulated as,

$$\min_{w} \mathbb{E}_{y\sim\hat{p}(y)}\mathbb{E}_{z\sim G_w(z|y)}\left[L_s + L_{pd}\right]. \tag{9}$$

By minimizing L_{pd} and L_s, the global model predictor can be refined by the pseudo features with its labels, enforcing the global model to align the local models and induce ideal predictions from a global view. The knowledge forgetting issue in the global model is solved.

3) Local Model Alignment: In federated learning, a drifted model in clients may be trained due to the data heterogeneity. For clearly presenting our proposed method, we briefly present the general model updating in clients.

In each training round, the server broadcasts the global model to each client. The local model is trained with the local data and the optimization in the k-th client is described as,

$$L_k = \frac{1}{|\mathcal{D}_k|} \sum\nolimits_{x \in \mathcal{D}_k} \left[L_{ce} \left(\sigma \left(g \left(f \left(x; \theta_k^f \right); \theta_k^p \right) \right), y_k \right) \right], \tag{10}$$

where L_k indicates the empirical risk [25]. $f(\cdot)$ represents the feature extractor function and θ_k^f is the parameters of the local feature extractor. y_k denotes the local data labels.

However, during this process, the non-iid of local data may cause model drifted problem. To address this issue, our proposed FedAlign additionally broadcasts the trained generator G to clients to convey the knowledge for local model alignment. Our local model alignment is composed of two parts: data augmentation with pseudo features and distribution alignment. In the first part, the trained generator G is used to generate the pseudo features $z \sim G_w(\cdot|y)$ for local models to balance the feature distributions of local data. And the objective of a local model to yield ideal predictions for the augmented features is presented as,

$$\begin{aligned} &\min_{\theta_k} \mathbb{E}_{y \sim \hat{p}(y)} \mathbb{E}_{z \sim G_w(z|y)} [L_{af}], \\ &\text{where } L_{af} = L_{ce}(\sigma(g(z; \theta_k^p)), y), \end{aligned} \tag{11}$$

where L_{af} is the augmented feature loss. In this way, each client can observe beyond its local data and regulate its predictions to approach the global one.

The pseudo features convey the global distribution knowledge to the local model. To align the distribution of local data feature space with that of the pseudo feature space from our generator G, a conditional KL-divergence is leveraged to match the two distributions,

$$\begin{aligned} &\min_{\theta_k} \mathbb{E}_{z \sim G_w(z|y_k)} [L_{dm}], \\ &\text{where } L_{dm} = D_{KL}(\sigma(g(\tilde{x}; \theta_k^p)) \| \sigma(g(z; \theta_k^p))) \end{aligned} \tag{12}$$

where L_{dm} is the distribution matching loss which performs distribution matching over the latent space, $\tilde{x}$ is the local data features from the local model. The overall objective of the local model training in clients is formulated as,

$$\min_{\theta_k} \; L_k + \mathbb{E}_{z \sim G_w(z|y_k)} [\beta L_{dm}] + \mathbb{E}_{y \sim \hat{p}(y)} \mathbb{E}_{z \sim G_w(z|y)} [\lambda L_{af}], \tag{13}$$

where β, λ are the hyperparameters.

C. Collecting Local Label Distributions in Clients

In the learning process of our generator, a direct way on weight α_k^y assignment in Eq. 4 is to give the same weight on each client, i.e., $\alpha_k^y = 1/|\mathcal{A}|$. However, due to data heterogeneity, class label distributions vary among different clients. To fully utilize the label distribution of local data, our proposed FedAlign dynamically assigns the weight α_k^y according to the class occupation in each client.

Firstly, the class label distribution of the local data is collected and uploaded to the server for summary as,

$$\sum\nolimits_k \mathbb{E}_{(x_k, y_k) \sim \mathcal{D}_k} [I(y_k = y)] = \sum\nolimits_k n_k^y, \tag{14}$$

where $I(\cdot)$ is an indicator function, n_k^y indicates the sample number of class y in client k. The statistical label distribution $\hat{P}(y)$ is also obtained in this process, so that the generator can produce pseudo features of all classes existing in clients,

$$\hat{p}(y) \propto \sum\nolimits_k n_k^y. \tag{15}$$

Then, the ensemble weights for knowledge distillation from local models can be further calculated as,

$$\alpha_k^y = n_k^y \Big/ \sum\nolimits_{i \in \mathcal{A}} n_i^y. \tag{16}$$

In this way, the distribution of local data is fully exploited, which minimizes the effects of local label distribution shifts and helps our generator to estimate the global data distribution.

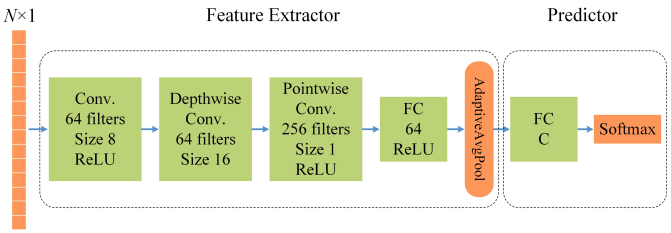

Fig. 4. The architecture of fault diagnosis model.

D. Fault Diagnosis Model in FedAlign

The fault diagnosis model architecture in our FedAlign is shared by clients and the server. Our proposed model architecture consists of two modules: the feature extractor and the predictor. The architecture is shown in Fig. 4. To better extract fault features, large kernels are used in the first few layers. The separable convolution [43] including depth-wise and poin-twise convolution, is used to extract robust features while reducing the model parameters. Our proposed model is lightweight, while remaining good performances.

IV. Experiment Study

A. Dataset Description

Experiments are conducted on two fault datasets: the CWRU dataset [44] and the DDS dataset. The DDS dataset is the planetary gearbox dataset from the Spectra Quests Drivetrain Dynamics Simulator (DDS) test rig. Fig. 5 shows the test rig of DDS.

The CWRU dataset obtains data by acceleration transducers from the drive-end bearings at a sampling frequency of 12 kHz. It includes fault state on the rolling element, the inner raceway, and the outer raceway. Fault diameters of 0.007, 0.014, and 0.021 are set for each fault state. Therefore, the

CWRU dataset obtains 10 class health states. The bearing data is acquired under four operational conditions with rotating speed changing between 1730 and 1797 rpm based on the load 0, 1, 2, and 3 hp. The CWRU dataset in the paper chooses 50 samples for each health state under each load for training and the same number of samples for test. It indicates that the CWRU dataset obtains 2000 samples of ten health states under four domains for training and 2000 samples of the same states for testing. The data length of each sample is 1024.

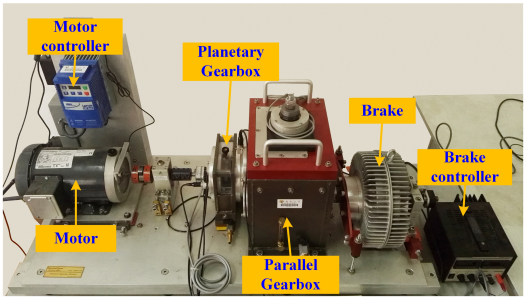

Fig. 5. The test rig of DDS.

TABLE I
PLANETARY GEARBOX CONDITION DESCRIPTIONS

Component	Type	Description
Bearing	Ball	Crack occurs in the ball
	Combo	Crack occurs in the both inner and outer ring
	Inner	Crack occurs in the inner ring
	Outer	Crack occurs in the outer ring
Gear	Chipped	Crack occurs in the feet
	Miss	One of the feet is missing
	Root	Crack occurs in the root of the feet
	Surface	Wear occurs in the surface

The DDS dataset obtains the data acquired by the 608A11 vibrating sensors placed on the planetary gearbox at a sampling frequency of 5120 Hz. The data is collected under various speed-load conditions, where the data from the speed-load conditions of 30Hz_2, 30Hz_4, 20Hz_0, and 40Hz_0 are chosen for experiments. The bearing-gearbox fault state descriptions are listed in Table I. Therefore, DDS dataset has 9 class health states including 8 fault states and 1 healthy state. The DDS dataset consists of 200 samples for each health state under each working condition for training and 200 samples of each state for testing. It means that the DDS dataset has 7200 samples as training data and 7200 samples as testing data. The data length of each sample is 1024.

B. Experimental Settings

We simulate different fault diagnosis tasks to investigate the performance of our proposed FedAlign, where different data distributions of local data are set in different clients. The general FL scenarios are presented as follows.

(1) IID: Independent and identically distributed (IID) scenario is the basic scenario, where the local data of clients are identically distributed. Concretely, the local data of clients includes all machine health states under all working conditions with the same distribution. 10 clients are set for experiments.

(2) Non-IID-Class: In non-iid-class scenario, the local data of each client consists of healthy data and one fault state data both under all working conditions. In this scenario, the local models will be extremely biased due to the extremely heterogeneous local data. Thus, 9 clients are set for CWRU dataset and 8 clients are set for DDS dataset in this scenario.

(3) Non-IID-Domain: In non-iid-domain scenario, the local data of different clients have different working conditions. In this case, the distribution shift issue is simulated in FL settings. Specifically, 4 domain clients are implemented where the local data of each client is under one working condition with all health states.

(4) Non-IID-Random: To simulate more non-iid scenarios in the real industry, we use a Dirichlet distribution $\mathbf{Dir}(\varepsilon)$ used in [18], [25]–[27] to simulate the non-iid data distributions among clients. In Dirichlet distribution $\mathbf{Dir}(\varepsilon)$, a smaller ε indicates higher data heterogeneity. Dirichlet distribution makes the label radios biased for clients, so that the local data of different clients obtain a random number of data with different classes under different working conditions. This scenario is more realistic and complex. In experiments, we set different ε, i.e., $\varepsilon = 0.1$ and $\varepsilon = 0.3$ to control the heterogeneity of data during the implementation. 10 clients are set, which is the same as the IID scenario. To clearly show the Dirichlet distribution, the illustration of the sample number per class in each client is illustrabuted in Fig. 6. It can be seen from Fig. 6 that the data in one client is extremely class imbalanced and the data of different clients are heterogeneous from each other.

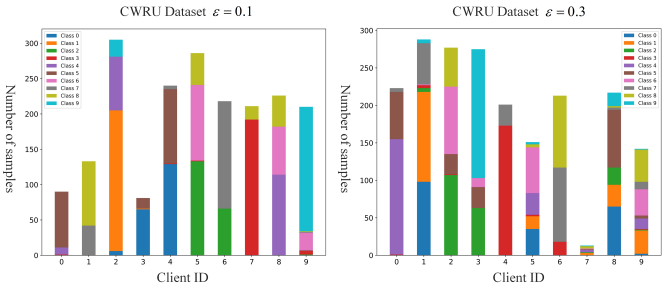

Fig. 6. Illustration of sample number per class in each client under Dirichlet distribution. Smaller ε indicates higher data heterogeneity.

For model testing, we all use the complete test dataset to evaluate the global model performance, which includes data from all classes under all four working conditions.

C. Compared Methods

Seven classic learning schemes are considered and re-implemented in the paper to show the superiority of our proposed FedAlign approach. The specific set of these learning schemes is presented as follows.

(1) Baseline: Each client trains its local model using the local data without communication with each other. In non-iid scenarios, the local data is insufficient and class-biased, so the fault model would be badly generalized on the test dataset with different distributions.

(2) Centralized: In this scheme, the local data of all clients are gathered to train a centralized model without considering data privacy.

(3) FedAvg: FedAvg is the vanilla FL algorithm proposed in [7] which is set as that of our FedAlign.

(4) FedS-S: FedS-S represents the algorithm proposed in [13] for machine fault diagnosis. We re-implement this method, and set its hyper-parameters as in [13].

(5) FedBalance: FedBalance indicates the algorithm like in [26] which uses a class-balanced loss in the local model training and combines the FedAvg algorithm for FL. Here, we adopt the Focus loss [19], [45] as the class-balanced loss with FedAvg.

(6) FedProx: FedProx is a federated optimization algorithm [24] with data heterogeneity. FedProx adds a local proximal term in the local training to restrict the local updates to be closer to the global model. Its experimental setting refers to that of FedAvg.

(7) FedGen: FedGen [25] is a data-free knowledge distillation method proposed for image tasks in FL. Its experimental setting is the same as that of our method for a fair comparison. The dimension of the hidden layer of its MLP generator is 256.

For fair comparisons, all approaches in the paper are performed in the same FL scenarios. Our experimental results are the average of five random runs. The detailed hyperparameters of our FedAlign method is listed in Table II. The loss weights β and λ in clients are initialized to be 1, and is decayed quadratically with weight 0.99. The input dimension of noise μ for our generator is 64. All approaches in the paper are programmed on a PC with Python 3.8.10 with torch 1.8.0, Cuda version 11.1 and executed on computer operating system Windows 10, Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz, 24.0GB RAM, and GPU NVIDIA GeForce RTX 3080, 10GB.

TABLE II
HYPERPARAMETERS OF OUR PROPOSED FEDALIGN

	Hyperparameter	value
CWRU	Communication rounds	100
DDS	Communication rounds	200
Clients	Learning rate	0.3
	Optimizer	sgd
	Local steps (I_l)	10
	Batch size	32
	Distillation batch size	32
Generator	Learning rate	0.03
	Optimizer	adam
	Generator steps (I_d)	50
	Batch size	64
Global	Learning rate	0.01
	Optimizer	sgd
	Global steps (I_g)	10
	Batch size	32

D. Comparison Experiments

We conduct comparison experiments and list the experimental results in Table III. It can be seen that our FedAlign achieves the best performance in all scenarios among all FL algorithms. The performance of the centralized learning is the upper limit result, where all clients' data are centralized together to form a complete dataset. The performance of Baseline is the lower limit result, which shows the diagnosis result of one client with insufficient data without federate learning. Compared with the Baseline, the methods with FL algorithms (i.e., FedAvg, FedProx, FedGen and our proposed FedAlign) achieve large performance gains and are more effective for machine fault diagnosis.

In the IID scenario, the Baseline performance in CWRU dataset is 97.83%. Compared with Baseline, FedAvg has a performance gain of +1.99%, FedBalance has a performance gain of +1.66%, FedProx has a performance gain of +1.97%, and FedGen has a performance gain of +2.01%. Our FedAlign achieves the best performance and obtains a performance gain of +2.05% for CWRU dataset compared with Baseline. For DDS dataset, the Baseline gets a 64.37% accuracy, which is surpassed by FedAvg, FedBalance, FedProx, FedGen, and FedAlign by at least 27.19%. This may be that the DDS dataset is more complex and requires more data to train. For the FedS-S method, it show unsatisfactory performances due to its simple model, which cannot extract robust features for DDS dataset.

In the non-iid-class scenario, the Baseline is extremely unpromising in both CWRU and DDS datasets with performances of 20.00% and 22.13%, respectively. This indicates that a single client with insufficient data cannot fulfill the requirement of fault diagnosis. Since the non-iid-class scenario is challenging, FedAvg only achieves 54.34% accuracy on CWRU dataset and 44.86% accuracy on DDS dataset. In comparison, our FedAlign achieves 84.24% on CWRU dataset and 57.94% accuracy on DDS dataset, which has a +29.90% gain on CWRU dataset and a +13.08% gain on DDS dataset, respectively compared with FedAvg. FedBalance that adopts the focus loss in FedAvg to address the class imbalance issues still does not perform as well as our proposed FedAlign in this non-iid-class scenario, for it can not balance the classes in different clients. FedProx can not outperform our proposed FedAlign as well. FedGen only uses the data-free knowledge distillation to regulate the local training, and achieves 77.37% accuracy on CWRU dataset, and 49.42% accuracy on DDS dataset, which is much lower than that of our method. These results show that our FedAlign is the most robust and effective FL method among all these compared approaches in the non-iid-class scenario.

In the non-iid-domain scenario, it can be observed that the Baseline achieves the low test accuracies in CWRU and DDS datasets with performances of 91.75% and 47.84%, respectively, due to the distribution shift of domains. The FL algorithms also improve the model performance in this scenario. The results of our proposed FedAlign are the closest to the upper limit results of the Centralized method. The test accuracy in non-iid-domain scenario is totally higher than that in non-iid-class scenario, which indicates that the class imbalance issue is more challenge than the domain distribution shift issue in FL for machine fault diagnosis. The benefits of our proposed FedAlign are also clearly shown in this non-iid scenario.

To further investigate the performance of our FedAlign, we conduct experiments under non-iid-random scenario, which

TABLE III
COMPARISON PERFORMANCES OF ALL APPROACHES

Dataset	Method	Test Accuracy (%)				
		iid	non-iid-class	non-iid-domain	non-iid-random $\varepsilon = 0.1$	non-iid-random $\varepsilon = 0.3$
CWRU	Centralized Learning	99.95±0.06				
	Baseline	97.83±0.26	20.00±0.00	91.75±0.80	29.81±0.24	45.28±0.34
	FedS-S	80.85±2.51	81.83±4.21	83.22±1.82	66.58±3.34	73.13±0.94
	FedAvg	99.82±0.08	54.34±6.08	99.90±0.05	82.62±1.53	94.90±1.32
	FedBalance	99.49±0.05	57.34±3.47	99.57±0.15	80.13±1.76	89.49±0.37
	FedProx	99.80±0.06	56.12±1.91	99.82±0.04	82.89±1.87	95.47±0.60
	FedGen	99.84±0.07	77.37±4.60	99.91±0.04	84.65±1.85	96.35±0.92
	FedAlign	**99.88±0.03**	**84.24±4.41**	**99.94±0.05**	**87.60±1.33**	**97.33±0.35**
DDS	Centralized Learning	97.02±0.39				
	Baseline	64.37±3.07	22.13±0.01	47.84±0.72	17.26±0.43	25.87±0.36
	FedS-S	25.43±0.69	31.96±1.60	30.10±0.88	18.97±0.15	23.47±1.26
	FedAvg	95.37±0.52	44.86±5.85	95.10±0.52	67.81±1.48	84.31±2.95
	FedBalance	91.56±1.01	51.80±0.98	90.09±0.71	54.75±1.89	71.39±1.32
	FedProx	95.84±0.37	44.98±3.18	93.01±1.09	64.21±0.17	83.53±1.14
	FedGen	96.03±0.60	49.42±3.27	95.74±0.47	69.83±1.58	86.48±1.63
	FedAlign	**96.87±0.55**	**57.94±2.98**	**96.24±0.39**	**72.16±1.58**	**87.82±0.95**

employs a Dirichlet distribution to control the degree of data heterogeneity. Our proposed FedAlign achieves the best accuracy among all approaches. This demonstrates that our FedAlign is effective for data heterogeneity in FL. Besides, the gain in performance of FedAlign is more notable in the extreme data heterogeneity scenario, and the gain in $\varepsilon = 0.1$ is larger than the gain in $\varepsilon = 0.3$ on both datasets. As the data heterogeneity degree decreases, the test accuracy of each method is increasing.

Through the comparison experiments, our proposed FedAlign outperforms other FL algorithms in all the setting scenarios. Our proposed FedAlign is able to align the global model to the local models and effectively solves the model drift and knowledge forgetting issues.

E. Study of Partial Client Participant

We explore different numbers of active clients participated in FL on the two dataset in non-iid scenario. There are 10 clients in total for our federated machine fault diagnosis. We vary the active ratios for clients ranging from 0.4 to 1 with a stride of 0.2 in Fig. 7. Fig. 7 shows the results of the impacts of partial clients participated in federated machine fault diagnosis. It can be seen that our proposed FedAlign still obtains the best performance under CWRU dataset with $\varepsilon = 0.1$, and under DDS dataset with $\varepsilon = 0.3$, regardless of the number of active clients. Besides, when all the clients participate in the federated machine fault diagnosis, the test accuracy achieves the highest in the above scenarios.

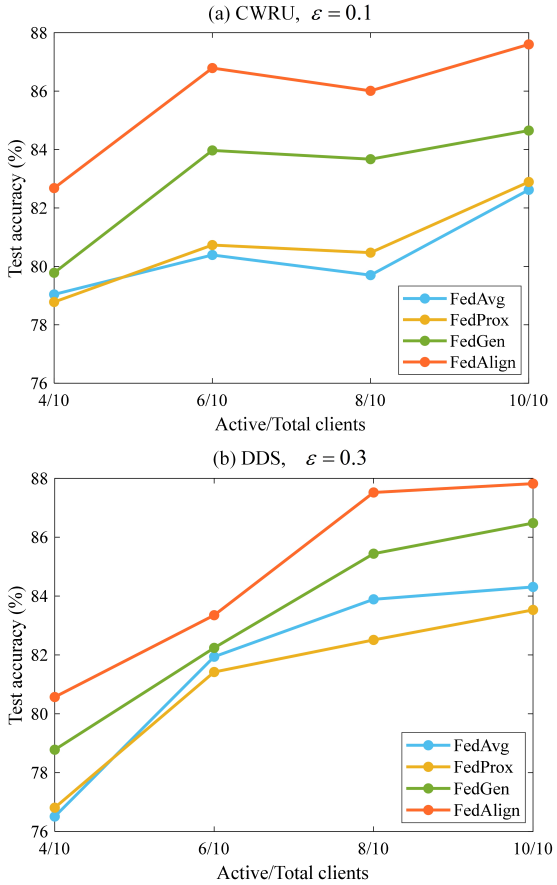

Fig. 7. The impacts of the partial client participant.

F. Ablation Study

In this subsection, analysis experiments are conducted to verify the effectiveness of our FedAlign.

1) Impacts of generator architecture: The architecture of generator may affect its performances. We conduct experiments to compare the performances with a generator in two different architectures: the convolutional architecture and the two-layer MLP, on CWRU dataset. The experimental results are illustrated in Fig. 8. The two-layer MLP architecture follows the same structure defined in FedGen. It can be found that our proposed FedAlign is robust to the architecture of the generator, which achieves outstanding performance with both

architectures. Although FedGen also utilizes a generator in FL, it performs inferior to our FedAlign. This may be because that it does not have the knowledge to align the global model, limiting its performances.

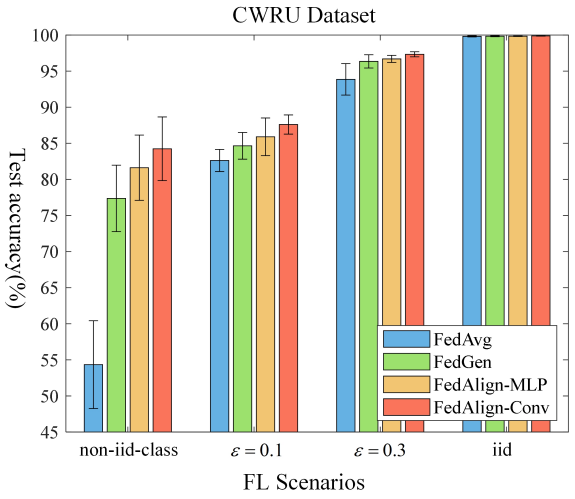

Fig. 8. The impacts of generator architecture.

2) Impacts of each component in global alignment: To investigate whether our FedAlign effectively solve this global model issue, we conduct experiments on the global alignment, and list the experimental results in Table IV. The experiments are performed on CWRU dataset in both non-iid-class and non-iid-random $\varepsilon = 0.1$ scenarios. In Table IV, -L_s and -L_{pd} represent removing pseudo feature classification and knowledge distillation, respectively. And -α_k^y indicates removing the dynamical weight assignment and using the same weight assignment.

TABLE IV
IMPACTS OF EACH COMPONENT IN THE SERVER

	Method	Test Accuracy(%)	
		non-iid-class	non-iid-random $\varepsilon = 0.1$
baseline	FedAlign	84.24±4.41	87.60±1.33
loss	-L_s	80.84±3.36	86.67±1.17
	-L_{pd}	79.37±8.78	85.45±2.07
weight	-α_k^y	81.83±3.94	85.52±2.00

It shows that removing any loss in the global alignment leads to worse performance. This demonstrates that our global alignment effectively solve the global issues. Although the semantic loss L_s indicates the supervised fine-tuning of the global predictor with pseudo features, the distribution of pseudo features may be different from the distribution of the global features, so the loss L_{pd} is required to distill knowledge to match the distributions and reduce the distribution discrepancy. Thus, removing one of them could lead to worse performance and the loss L_{pd} indicating knowledge distillation in global model affects the most. For the dynamical weight assignment, the test accuracy decreases by at least 2.08%, when we remove it. This shows that the dynamical weight assignment can help our generator to learn the distribution of local data. In summary, these experimental results shows that our proposed global alignment and dynamical weight assignment can effectively improve the performances of our FedAlign by aligning the global model to the local models.

V. CONCLUSION

This paper has proposed a novel approach FedAlign for machine fault diagnosis in FL, which effectively aligns a federated global model to local models via data-free knowledge distillation and solves the data heterogeneity issues in FL. The proposed FedAlign has trained a compact and convolutional generator in a data-free manner to learn the data distribution in a global view from local models. Then, this generator has been used to produce pseudo features to convey the global distribution knowledge for both global and local models alignment, alleviating the model drift and knowledge forgetting issues caused by data heterogeneity for machine fault diagnosis in FL and boosting the performance of the federated global model. The experiments performed on several different scenarios have verified that the proposed FedAlign approach facilitates FL in machine fault diagnosis with favorable effectiveness, and achieves significant performance gains in non-iid FL settings compared with state-of-the-art methods.

Although the generator in our proposed FedAlign is well trained in the data-free manner and used to convey knowledge for model alignment, it sometimes may face the unstable training issues. After applying regularization and batch normalization, our proposed generator becomes more stable in the training process. In addition, this paper also provides several directions for the future research. The distribution of mechanical data in the industry diverges significantly, and there are more heterogeneous scenarios. In order to further accommodate the data heterogeneity of clients, a more robust fault diagnosis model needs to be implemented. This work only generates pseudo features, and in the future, more sophisticated generator can be investigated to generate the pseudo raw data, balancing the data distributions of clients. Additionally, the data-free knowledge distillation method can also be extended to model heterogeneity scenarios by generating raw data for knowledge distillation.

REFERENCES

[1] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019.

[2] J. Chen, R. Huang, K. Zhao, W. Wang, L. Liu, and W. Li, "Multiscale convolutional neural network with feature alignment for bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[3] L. Zhang, H. Zhang, and G. Cai, "The multiclass fault diagnosis of wind turbine bearing based on multisource signal fusion and deep learning generative model," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[4] B. A. Tama, M. Vania, S. Lee, and S. Lim, "Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals," *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4667–4709, 2023.

[5] T. Zhang, J. Chen, F. Li, K. Zhang, H. Lv, S. He, and E. Xu, "Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions," *ISA transactions*, vol. 119, pp. 152–171, 2022.

[6] Y. Dong, Y. Li, H. Zheng, R. Wang, and M. Xu, "A new dynamic model and transfer learning based intelligent fault diagnosis framework for rolling element bearings race faults: Solving the small sample problem," *ISA transactions*, vol. 121, pp. 327–348, 2022.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[8] G. Jiang, W. Fan, W. Li, L. Wang, Q. He, P. Xie, and X. Li, "Deepfedwt: A federated deep learning framework for fault detection of wind turbines," *Measurement*, vol. 199, p. 111529, 2022.

[9] F. Deng, Z. Zeng, W. Mao, B. Wei, and Z. Li, "A novel transmission line defect detection method based on adaptive federated learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[10] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[11] J. C. Jiang, B. Kantarci, S. Oktug, and T. Soyata, "Federated learning in smart city sensing: Challenges and opportunities," *Sensors*, vol. 20, no. 21, p. 6230, 2020.

[12] D. C. Nguyen, M. Ding, Q.-V. Pham, P. N. Pathirana, L. B. Le, A. Seneviratne, J. Li, D. Niyato, and H. V. Poor, "Federated learning meets blockchain in edge computing: Opportunities and challenges," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 806–12 825, 2021.

[13] 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.

[14] X. Ma, C. Wen, and T. Wen, "An asynchronous and real-time update paradigm of federated learning for fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8531–8540, 2021.

[15] J. Lin, J. Ma, and J. Zhu, "Hierarchical federated learning for power transformer fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[16] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[17] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[18] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[19] X. Zhao, J. Yao, W. Deng, M. Jia, and Z. Liu, "Normalized conditional variational auto-encoder with adaptive focal loss for imbalanced fault diagnosis of bearing-rotor system," *Mechanical Systems and Signal Processing*, vol. 170, p. 108826, 2022.

[20] T. Zhang, J. Chen, S. Liu, and Z. Liu, "Domain discrepancy-guided contrastive feature learning for few-shot industrial fault diagnosis under variable working conditions," *IEEE Transactions on Industrial Informatics*, 2023.

[21] Z. Chen, J. Wu, C. Deng, C. Wang, and Y. Wang, "Residual deep subdomain adaptation network: A new method for intelligent fault diagnosis of bearings across multiple domains," *Mechanism and Machine Theory*, vol. 169, p. 104635, 2022.

[22] Y. Chen, D. Zhang, and R. Yan, "Domain adaptation networks with parameter-free adaptively rectified linear units for fault diagnosis under variable operating conditions," *IEEE Transactions on Neural Networks and Learning Systems*, 2023, doi: 10.1109/TNNLS.2023.3298648.

[23] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[24] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[25] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 878–12 889.

[26] H.-Y. Chen and W.-L. Chao, "On bridging generic and personalized federated learning for image classification," *arXiv preprint arXiv:2107.00778*, 2021.

[27] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2351–2363, 2020.

[28] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.

[29] Y. Aono, T. Hayashi, L. Wang, S. Moriai *et al.*, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, 2017.

[30] J. Du, N. Qin, D. Huang, Y. Zhang, and X. Jia, "An efficient federated learning framework for machinery fault diagnosis with improved model aggregation and local model training," *IEEE Transactions on Neural Networks and Learning Systems*, 2023, doi: 10.1109/TNNLS.2023.3238724.

[31] W. Zhang and X. Li, "Data privacy preserving federated transfer learning in machinery fault diagnostics using prior distributions," *Structural Health Monitoring*, vol. 21, no. 4, pp. 1329–1344, 2022.

[32] S. Lu, Z. Gao, Q. Xu, C. Jiang, A. Zhang, and X. Wang, "Class-imbalance privacy-preserving federated learning for decentralized fault diagnosis with biometric authentication," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 9101–9111, 2022.

[33] M. Mehta, S. Chen, H. Tang, and C. Shao, "A federated learning approach to mixed fault diagnosis in rotating machinery," *Journal of Manufacturing Systems*, 2023, doi: 10.1016/j.jmsy.2023.05.012.

[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[35] L. Hu, H. Yan, L. Li, Z. Pan, X. Liu, and Z. Zhang, "Mhat: An efficient model-heterogenous aggregation training scheme for federated learning," *Information Sciences*, vol. 560, pp. 493–503, 2021.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[37] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, "Data-free learning of student networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3514–3522.

[38] G. Fang, J. Song, C. Shen, X. Wang, D. Chen, and M. Song, "Data-free adversarial distillation," *arXiv preprint arXiv:1912.11006*, 2019.

[39] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[40] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[41] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[42] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1429–1437.

[43] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[44] "Case western reserve university bearing data center," http://csegroups.case.edu/bearingdatacenter/pages/welcomecase-western-reserve-university-bearing-data-center-website.

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

Wenjun Sun received her M.S. degree in instrument science and technology at the School of Instrument Science and Engineering at Southeast University, Nanjing, China, in 2017, and is currently working toward the Ph.D. degree in instrument science and technology at the School of Instrument Science and Engineering, Southeast University.

Her current research is focused on deep learning-based mechanical fault diagnosis and federated learning.

Ruqiang Yan (M07, SM11) received the M.S. degree in precision instrument and machinery from the University of Science and Technology of China, Hefei, China, in 2002, and the Ph.D. degree in mechanical engineering from the University of Massachusetts at Amherst, MA, USA, in 2007.

He was a Guest Researcher at the National Institute of Standards and Technology (NIST) in 2006-2008 and a Professor with the School of Instrument Science and Engineering, Southeast University, Nanjing, China from 2009 to 2018. He joined the School of Mechanical Engineering, Xian Jiaotong University, Xian, China, in 2018. His research interests include data analytics, machine learning, and energy-efficient sensing and sensor networks for the condition monitoring and health diagnosis of large-scale, complex, dynamical systems.

Dr. Yan is a Fellow of ASME (2019) and IEEE (2022). His honors and awards include the IEEE Instrumentation and Measurement Society Technical Award in 2019, the New Century Excellent Talents in University Award from the Ministry of Education in China in 2009, and multiple best paper awards. He is also the Associate Editor-in-Chief of the IEEE Transactions on Instrumentation and Measurement and an Associate Editor of the IEEE Systems Journal and the IEEE Sensors Journal.

Ruibing Jin Ruibing Jin received the B.Eng. degree from the University of Electronic Science and Technology of China in 2014 and the M.Eng and the Ph.D. degrees from Nanyang Technological University, Singapore in 2016 and 2020, respectively. He is a Scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. He won the First Place Winner in the CVPR 2021 UG2+ Challenge. His research interests include computer vision, machine learning, time series and related applications.

Rui Zhao Rui Zhao is Vice President, Head of data and quant research at Pluang Tech, Singapore. He received the B.Eng. degree in measurement and control from Southeast University, Nanjing, China, in 2012, and the Ph.D. degree in machine learning from Nanyang Technological University, Singapore, in 2017. His current research interests include machine learning and its applications in text mining, machine health monitoring and quantitative trading.

Zhenghua Chen Zhenghua Chen received the B.Eng. degree in mechatronics engineering from University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2011, and Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2017. Currently, he is a Scientist and Lab Head at Institute for Infocomm Research, and an Early Career Investigator at Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore. He has won several competitive awards, such as First Place Winner for CVPR 2021 UG2+ Challenge, A*STAR Career Development Award, First Runner-Up Award for Grand Challenge at IEEE VCIP 2020, Best Paper Award at IEEE ICIEA 2022 and IEEE SmartCity 2022, etc. He serves as Associate Editor for IEEE Transactions on Industrial Informatics, IEEE Transactions on Instrumentation and Measurement, IEEE Transactions on Industrial Cyber-Physical Systems, IEEE Sensors Journal, Springer Discover Artificial Intelligence, and Elsevier Neurocomputing. He is currently the Chair of IEEE Sensors Council Singapore Chapter and IEEE Senior Member. His research interests include data-efficient and model-efficient learning with related applications in smart city and smart manufacturing.