Curriculum-based Federated Learning for Machine Fault Diagnosis with Noisy Labels

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

***Abstract*— Federated Learning (FL) has emerged as an effective machine-learning paradigm for collaborative machine fault diagnosis in a privacy-preserving scheme. However, due to the perception limitation and different annotation criteria of annotators, the data in clients may have noisy labels with varied noise levels, leading to degraded FL performances. Most existing methods in FL for tackling the label noise issue, assume that there is label noise in all clients and treat all clients with the same denoising training. However, these methods may result in sub-optimization and even training instability of local models, so that they cannot perform well on heterogeneous label noise across clients in FL. To address this issue, we propose a curriculum-based federated learning method (called FedCNL) to combat the heterogeneous label noise in FL settings. Firstly, our proposed FedCNL exploits a noise modeling module to adaptively estimate the clean clients and noisy clients, and identify the clean samples and noisy samples in noisy clients in an unsupervised manner. Then, a multi-stage curriculum learning is designed by regarding the noise level as learning complexity, where the model learns from clean to noisy samples, gradually improving the performance of the global model. Moreover, a mixed loss correction method is explored in the curriculum stage to maximize the utilization of data with noisy labels. Experiments performed on fault datasets in non-IID settings indicate that our proposed method addresses the label noise issue for machine fault diagnosis in heterogeneous FL with favorable effectiveness, achieving state-of-the-art performances.**

***Index Terms*— Curriculum learning, Fault diagnosis, Federated learning, Label noise, Loss correction**

I. Introduction

FEDERATED learning (FL) [1], [2] has received massive attention in machine fault diagnosis [3], [4] for its decentralized machine-learning paradigm in a privacy-preserving scheme. FL provides an effective solution to data island problem for fault diagnosis [5] in the industry by cooperatively training the local diagnosis models, while keeping the local data safely stored in local clients. This new decentralized learning paradigm in machine fault diagnosis is of great significance as it promotes the cooperative fault diagnosis between companies, further improves the classification accuracy of fault diagnosis model, and ensures the safe and economic operation of mechanical systems.

The data of machinery in different companies is almost non-identically and independently distributed (non-IID), so that FL applied to machine fault diagnosis in practice faces the challenge of data heterogeneity [4]–[7]. Furthermore, the complex operating conditions of machines lead to discrepancies in the perception of fault types and annotation criteria for annotators in different companies. This inevitably causes noisy labels during the annotation process in some clients, with the noise level being heterogeneous across clients. Noisy labels can lead to performance degradation of deep model, as the models are easily overfitted to the noisy labels [8], [9]. In FL, the clients with label noise not only affect their own local model training but also degrade the performances of the other clients participating in the FL, resulting in a significant performance degradation of FL [10], [11]. Consequently, in addition to the data heterogeneity issue, the label noise issue is also an unignorable problem for machine fault diagnosis in FL.

Although the label noise issue has been studied extensively in centralized learning [12]–[16] over the past years, it is still challenging to tackle the label noise issue for machine fault diagnosis in FL, due to the limited sizes of local fault datasets and the data privacy constraints. To alleviate the label noise issue in FL, some methods have been proposed, which can be mainly divided into two complementary perspectives. The first one assumes that label noise exists in all clients and adopts the same training strategy in all clients to combat the label noise [17], [18]. However, these methods cannot deal with the heterogeneous label noise across clients, because the training strategy for label noise cannot be adapted well to data with low level or no noisy labels and may result in sub-optimization and even training instability of local model. Another focuses on selecting relatively clean clients and heavily relies on the

This work was supported in part by the National Natural Science Foundation of China under Grant U23A20620 and in part by the National Research Foundation, Singapore under its AI Singapore Programme (AISG2-RP-2021-027). (Corresponding author: Ruqiang Yan and Ruibing Jin.)

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

Ruqiang Yan is with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China, and the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, 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)

predictions of the global model federated by these clean clients [19]–[21]. Nevertheless, these methods cannot make full use of the data with noisy labels and limit the performance of the global model especially in FL settings with high data heterogeneity due to the low confidence of the global model federated by limited amount of training data in clean clients.

To address the challenge of heterogeneous label noise across clients without falling into sub-optimization and to maximize the utilization of all local data, we propose a **C**urriculum-based **Fed**erated label **N**oise **L**earning method (called FedCNL) for federated machine fault diagnosis with noisy labels. The proposed FedCNL method is inspired by curriculum learning [13], [22], which formalizes the learning strategies of human beings and advocates learning starting from easy samples and gradually progressing to complex samples and tasks. The noisy labeled samples are more challenging to learn than the clean samples. Considering the different levels of label noise across clients, our proposed FedCNL regards the noise level as learning complexity and develops a multi-stage curriculum learning strategy by ranking both client and sample learning complexities. This strategy allows the model to begin learning from relatively clean samples and gradually progress to noisy samples, thereby making full use of local data and gradually improving the performance of the global model.

Specifically, the proposed FedCNL firstly exploits a noise modeling module in the pre-processing stage to adaptively identify the clean clients and noisy clients, and subsequently, identify the clean samples and noisy labeled samples in noisy clients using posterior probabilities in an unsupervised manner. Then, a multi-stage curriculum learning strategy is developed by ranking both client and sample learning complexities, which enables the federated training to start from the clean clients to noisy clients and then from the clean samples in noisy clients to the noisy samples in noisy clients, to gradually enhance the performance of the global model for heterogeneous label noise. Finally, a mixed loss correction method is explored in noisy clients to bootstrap the training loss using the label noise posterior probabilities for the curriculum-based FL to maximize the utilization of local data with noisy labels.

To the best of our knowledge, this work is the first to investigate the label noise issue for machine fault diagnosis in federated learning. The main contributions of this work can be summarized as follows:

1) A noise modeling module is exploited for heterogeneous label noise across clients to identify the clients and the samples that are likely to be noisy in an unsupervised manner for curriculum learning.
2) A multi-stage curriculum learning strategy is developed by ranking both client and sample learning complexities to tackle the label noise issue in federated machine fault diagnosis.
3) A mixed loss correction method is explored in the curriculum to bootstrap the training of noisy clients, fully utilizing noisy data and improving the model robustness to label noise.
4) Comprehensive experiments are conducted on machine fault datasets under different FL settings. The results indicate that our proposed FedCNL outperforms the competing methods with strong robustness to heterogeneous label noise and data in FL for machine fault diagnosis.

The remainder of the paper is organized as follows. Section II introduces the related works. In Section III, the problem setting of our work is provided. In Section IV, the methodology of the proposed FedCNL is presented. Section V provides the experimental setup and results in FL. Finally, a conclusion is provided in Section VI.

II. Related Works

A. Label Noise Learning

Label noise learning has been widely studied in centralized learning [12]–[16], [23], [24]. Existing studies can be roughly divided into two complementary perspectives: no-selection methods [14], [16], [23] and selection-based methods [12], [13], [15], [24]. For the former, Zhang et al. [23] proposed mixup data augmentation to improve the robustness against label noise without explicitly modeling it. Reed et al. [14] proposed a mechanism to deal with noisy labels by adding a perceptual term to the standard cross-entropy loss with bootstrapping. For the latter, Guo et al. [13] proposed a CurriculumNet for weakly supervised learning from large-scale web images. Han et al. [15] proposed Co-teaching, which trains two deep neural networks simultaneously and selects data of possibly clean labels for cross-training. Cheng et al. [24] proposed a semi-supervised learning strategy based on time series to simultaneously train two DNNs with cross-training for sample selection and label correction for fault diagnosis.

Although there are massive studies on label noise learning, they are proposed under the centralized learning. Direct application of these centralized learning methods in FL is challenging and results in performance degradations in FL settings [20], [25]. This is primarily due to the data scarcity and data privacy issues associated with local data used for machine fault diagnosis in FL.

B. Federated Learning with Label Noise

Recently, several FL methods have focused on addressing label noise issues to enhance the robustness of FL against noisy labels. These approaches can be categorized into two complementary perspectives: no-client selection methods [17], [18], [26] and client selection based methods [20], [21], [25]. For the former, Fang et al. [17] proposed a robust heterogeneous federated learning (RHFL) method to handle the label noise and model heterogeneity issues simultaneously. Yang et al. [18] proposed a label correction method during the training with interchanging class-wise centroids. FedFixer [26] presented a dual model structure to cross-train the global model and local model with clean samples only. However, these methods adopt the same noise training strategy in all clients, which easily results in sub-optimization of local models. Alternatively, FedCorr [20] selected clean clients via the LID scores and used the global model federated by the clean clients to correct the noisy labels in noisy clients. Finally,

the global model was trained in the conventional FL stage with all clients. FedNoRo [21] selected clean clients with the per-class average loss values of clients and applied a knowledge distillation loss to combat label noise in noisy clients. Nevertheless, these methods heavily rely on the global model federated by clean clients and may easily fail in extreme non-IID settings, since the limited amount of data in clean clients leads to a poor global model that cannot guide the noise training effectively.

In contrast, our proposed FedCNL selects both the clients and the samples to design the multi-stage curriculum learning strategy to gradually learn from the clean samples to noisy samples for bootstrapping the noise training, maximizing the utilization of local data and gradually improving the performance of the global model for heterogeneous label noise and data.

III. Problem Setting

In this paper, the heterogeneous label noise issue in federated learning for machine fault diagnosis is investigated and the main assumptions for federated machine fault diagnosis are presented as follows:

1) Multiple clients with similar machines participate in federated learning to obtain a shared global model.
2) All clients have a union set of fault labels and adopt the same fault diagnosis model architecture.
3) Each client has its local data varied in category and number, and with different levels of noisy labels.
4) The local data of each client is private and cannot be communicated with each other.

The K clients with dataset $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^{K}$, where $\mathcal{D}_k = \left\{(x_{k,i}, \widehat{y}_{k,i})_{i=1}^{|\mathcal{D}_k|}\right\}$ is the local private dataset of client k, participate in FL system for machine fault diagnosis. Here, $x_{k,i}$ is the i-th sample in client k and $\widehat{y}_{k,i}$ indicates the label of sample $x_{k,i}$, which may be the correct label $y_{k,i}$ or the niosy one $\widehat{y}_{k,i} \neq y_{k,i}$, $|\mathcal{D}_k|$ denotes the number of data samples in client k. We define θ_k as the local model parameters of client k and θ as the global model parameters, $f(\cdot)$ representing the fault diagnosis model function, and $\mathcal{A}$ indicating a set of all K clients. In FL, the objective is to optimize the global model θ across all clients as follows,

$$\min_{\theta} L(\theta) = \sum_{k \in \mathcal{A}_t} \frac{|\mathcal{D}_k|}{\sum_{i \in \mathcal{A}_t} |\mathcal{D}_i|} L_k(\theta), \tag{1}$$

where $\mathcal{A}_t \in \mathcal{A}$ indicates the subset of selected clients, $L_k(\theta)$ is the local training loss of client k on dataset $\mathcal{D}_k$, formulated as,

$$L_k(\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{i=1}^{|\mathcal{D}_k|} L_{ce}^{i}(\widehat{y}_{k,i}, f(x_{k,i}; \theta)), \tag{2}$$

where L_{ce} is the cross entropy loss.

A standard FL algorithm FedAvg [1], solves the global optimization problem through iteratively communication between local training and global aggregation as follows,

$$\text{Local}: \theta_k = \arg\min_{\theta_k} L_k(\theta_k), \text{ initialized with } \theta, \tag{3}$$

$$\text{Global}: \theta \leftarrow \sum_{k \in \mathcal{A}_t} \frac{|\mathcal{D}_k|}{\sum_{i \in \mathcal{A}_t} |\mathcal{D}_i|} \theta_k, \tag{4}$$

where the local training is performed with multiple epochs of stochastic gradient descent to minimize the local loss L_k for the local model θ_k. The global aggregation is realized by taking weighted average over local model parameters.

However, the FedAvg algorithm can easily lead to a poor global model with degraded performance in the presence of label noise in clients. Thus, we propose FedCNL to address the label noise issue in FL for machine fault diagnosis.

IV. Methodology

Our proposed FedCNL designs a multi-stage curriculum learning strategy to start the federated training from clean samples gradually to noisy samples, improving the performance of the global model stage by stage. Considering the local data is stored in clients and cannot be accessing to each other, our proposed FedCNL firstly exploits a noise modeling module in the pre-processing stage to identify the clients and the samples in clients that are likely to be noisy during the federated training. Subsequently, a multi-stage curriculum learning is designed by ranking both client and sample learning complexities to gradually learn from clean to noisy samples. The framework of our proposed FedCNL is illustrated in Fig. 1 and the corresponding training process of our proposed FedCNL is elaborated in Algorithm 1.

A. Pre-processing Stage for Label Noise

The pre-processing stage is conducted for early federated training to model the different levels of label noise. During this stage, the label noise levels across clients and within clients are both modeled to identify the clients and the samples that are likely to be noisy. Previous studies [8], [9], [12] have indicated that samples with noisy labels exhibit higher training loss compared to clean samples in the early stage of model training. Thus, our proposed FedCNL models on the per-sample loss to identify whether a sample is clean or noisy. Accordingly, the local training loss of a client in the early stage is modeled to identify the clients.

1) Federated Training: In the pre-processing stage, a global model is trained based on FedAvg for several warm-up rounds (T_0) for noise modeling. To enhance the robustness of the global model to label noise, a popular data augmentation technique Mixup [23] is applied in the federated training, which has exhibited strong robustness to label noise. It constructs synthetic training samples $(\widetilde{x}, \widetilde{y})$ on convex combinations of sample pairs (x_i, y_i) and (x_j, y_j):

$$\begin{aligned} \widetilde{x} &= \lambda x_i + (1-\lambda) x_j, \\ \widetilde{y} &= \lambda y_i + (1-\lambda) y_j, \\ L &= \lambda L_{ce}^{i} + (1-\lambda) L_{ce}^{j}, \end{aligned} \tag{5}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ is randomly sampled from the beta distribution and $\alpha \in (0, \infty)$. Mixup extends the training distribution by random interpolations to enhance the sample diversity. For label noise, Mixup improves the robustness to label noise by combining clean and noisy samples, computing

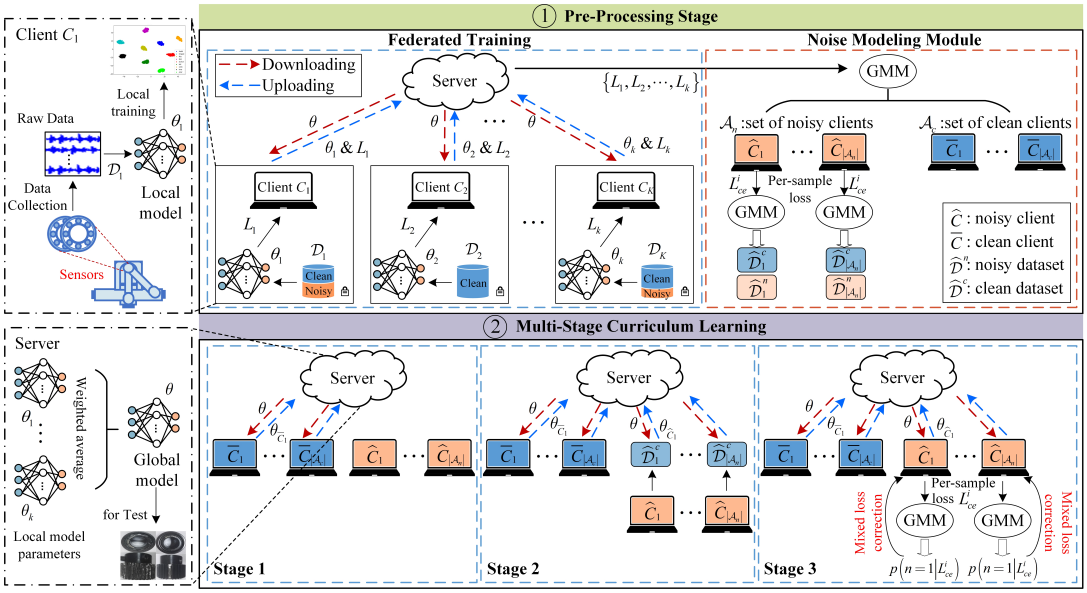

Fig. 1. The framework of our proposed FedCNL method.

Algorithm 1 FedCNL learning

INPUT: Communication rounds T_0, T_1, T_2, T_3; Number of Clients K; Datasets of clients $\{\mathcal{D}_k\}_{k=1}^K$; Global model parameters θ; Local model parameters $\{\theta_k\}_{k=1}^K$.

Initialize model parameters θ;
// Pre-processing:
for t=1 **to** T_0 **do**
  $\mathcal{A} \leftarrow$ (a set of all K clients);
  for client $k \in \mathcal{A}$ in parallel **do**
    $\theta_k \leftarrow \theta$;
    $\theta_k \leftarrow$ Update(θ_k;$\mathcal{D}_k$, Mixup) by Eq.(2) and (5);
    Upload θ_k and loss value L_k to the server;
  Update global model θ by Eq. (4);
  Compute a GMM based on $\{L_k\}_{k=1}^K$ to divide the clients into clean set $\mathcal{A}_c$ and noisy set $\mathcal{A}_n$;
  for noisy client $k \in \mathcal{A}_n$ in parallel **do**
    Compute a GMM based on $\left\{L_{ce}^i\right\}_{i=1}^{|\mathcal{D}_k|}$ to divide the local dataset into clean subset $\widehat{\mathcal{D}}_k^c$ and noisy subset $\widehat{\mathcal{D}}_k^n$;
// Multi-Stage Curriculum Learning:
for t=1 **to** T_1 **do**
  Update θ by federated training among clients in $\mathcal{A}_c$;
for t=1 **to** T_2 **do**
  Update θ by federated training among clients in $\mathcal{A}_c$ and clients in $\mathcal{A}_n$ which only provide its clean subset $\widehat{\mathcal{D}}_k^c$;
for t=1 **to** T_3 **do**
  for clean client $k \in \mathcal{A}_c$ in parallel **do**
    Update θ_k by minimizing loss function in Eq.(10) with $\omega = 0$;
  for clean client $k \in \mathcal{A}_n$ in parallel **do**
    Compute a GMM to obtain the posterior probability $p\left(n=1\left|L_{ce}^i\right.\right)$ for each sample;
    Update θ_k by minimizing loss function in Eq.(10) with $\omega = p\left(n=1\left|L_{ce}^i\right.\right)$;
  Update global model θ by Eq. (4);
OUTPUT: Global model θ.

a more representative loss to reduce the negative impact caused by noisy labels.

In the federated training at the pre-processing stage, the local training loss value L_k on the mixup augmentation of dataset $\mathcal{D}_k$ in the k-th client is also uploaded to the server along with the local model θ_k for modeling the label noise level across clients. It should be noted that our proposed FedCNL is privacy-preserving, since only the additional client-wise training loss value is uploaded to the server in comparison to the usual FL, and the loss value is a single scalar that cannot recover the local data.

2) Noise Modeling Module: To adaptively model the label noise, Gaussian Mixture Models (GMMs) are utilized in our proposed FedCNL for noise modeling in an unsupervised manner. The probability density function (pdf) of a GMM model on the loss value l with N components is formulated as:

$$p(l) = \sum_{n=1}^{N} \gamma_n p(l|n), \tag{6}$$

where γ_n are the mixing coefficients of each individual pdf $p(l|n)$, $N = 2$ indicates that a two-component GMM is fitted to model the distribution of the loss values $l \sim \mathcal{N}(\mu_n, \sum_n)$. And the Expectation Maximization (EM) algorithm [27] is applied to fit the GMM to the distributions. Then, the probability of a client(sample) being clean or noisy can be obtained through the posterior probability:

$$p(n|l_i) = \frac{p(n)\,p(l_i|n)}{p(l_i)}, \tag{7}$$

where n is the Gaussian component and we use $n = 0$ (1) to denote the clean (noisy) components for intuitive distinction. The clean component corresponds to the Gaussian distribution with the small μ_n and the noisy component corresponds to the Gaussian distribution with the large μ_n.

Our proposed FedCNL employs GMMs to model the label noise levels across clients and within clients for curriculum de-

sign. At the last federated training round in the pre-processing stage, a GMM is computed on the client-wise training loss values $\{L_k\}_{k=1}^{K}$ in the server to identify the clean clients and noisy clients. With the GMM, the clients can be divided into two subsets: clean clients $\mathcal{A}_c$ and noisy clients $\mathcal{A}_n$. After identifying the noisy clients, each noisy client $k \in \mathcal{A}_n$ locally computes the per-sample cross-entropy loss L_{ce}^{i} using the global model and estimates a GMM on the per-sample loss values $\left\{L_{ce}^{i}\right\}_{i=1}^{|\mathcal{D}_k|}$ of all local samples to identify the clean samples and the noisy samples. As a result, the local dataset $\mathcal{D}_k$ in noisy client k can be divided into two subsets: clean dataset $\widehat{\mathcal{D}}_k^c$ and noisy dataset $\widehat{\mathcal{D}}_k^n$. Then, a multi-stage curriculum learning strategy can be designed by ranking both client and sample learning complexities.

B. Multi-Stage Curriculum Learning

The design of the multi-stage curriculum relies on the intuition that the federated training is conducted sequentially from easy samples to complex ones. Thus, considering the learning complexities of both clients and samples, a three-stage curriculum learning strategy is developed.

In the first curriculum stage, only a set of clean clients $\mathcal{A}_c$ participate in the federated training with mixup augmentation for T_1 rounds until the global model training converges. In the second stage, the noisy clients with clean datasets are incorporated into the federated training for T_2 rounds, providing more labeled training samples for the global model training. Through the first two-stage curriculum learning, the federated global model achieves high confidence in fault classification. Consequently, in the third stage, the noisy clients add the noisy datasets into the federated training, involving all clients and all local data in training for T_3 rounds. To maximize the use of noisy data and further enhance the robustness of the global model to heterogeneous label noise and data across clients, a mixed loss correction method and a local proximal regularization term are explored in this stage. The relevant training details for this third curriculum stage are introduced as follows.

Standard cross-entropy loss is ill-fitted to the label noise tasks, as it tends to easily fit the noisy labels [8], [9]. Meanwhile, the bootstrapping loss [14] is proposed to improve the standard cross-entropy loss for label noise by adding a perceptual term to the loss function which can help to correct the training objective:

$$L_B = -\sum_{i=1}^{M} \left((1-\omega)\,\widehat{y}_i + \omega z_i\right)^T \log\left(q_i\right), \tag{8}$$

where q_i is the softmax probabilities produced by the training model, z_i denotes the class prediction produced by the model, M is the number of training samples, ω indicates the weight for bootstrapping the model prediction z_i and it is fixed in [14].

Unfortunately, using a fixed weight value of ω for all samples cannot well prevent the model from fitting the noisy labels. Our proposed FedCNL improves the bootstrapping loss by using the posterior probability of a sample being noisy to weight each sample individually. Consequently, the value of weight ω is set to be $p\left(n=1\left|L_{ce}^{i}\right.\right)$ for each sample in each noisy client for dynamically bootstrapping. Combined with the mixup augmentation, a mixed loss correction method is explored to implement a robust per-sample loss correction:

$$\begin{aligned} L_n = -\lambda \left[\left((1-\omega_i)\,\widehat{y}_i + \omega_i z_i\right)^T \log\left(q\right)\right] - \\ (1-\lambda)\left[\left((1-\omega_j)\,\widehat{y}_j + \omega_j z_j\right)^T \log\left(q\right)\right], \end{aligned} \tag{9}$$

where L_n is the mixed bootstrapping training loss, $\omega_i = p\left(n=1\left|L_{ce}^{i}\right.\right)$ and $\omega_j = p\left(n=1\left|L_{ce}^{j}\right.\right)$ are inferred from the GMM model which controls the confidence in the labels, q indicates the softmax probabilities of the mixed samples. Note that in the third curriculum stage, the posterior probabilities estimated by the GMM models are updated in every federated round for each noisy client. Thus, the mixed loss correction method can lead to a robust global model by trusting in progressively better predictions during training, fully using all the local data.

To further improve the robustness of global model to data heterogeneity, a local proximal regularization term is finally applied in all clients to constrain the local models not to deviate from the global model. Hence, the overall objective of the local model for clients in the third curriculum stage is formulated as,

$$L_k^* = L_n + \beta\|\theta_k - \theta\|^2. \tag{10}$$

where β is the hyperparameter to control the overall effect of the local proximal regularization term. It should be noted that ω in L_n for noisy clients is set to be $p\left(n=1\left|L_{ce}^{i}\right.\right)$, while the ω in L_n for clean clients is set to be 0, which means the L_n is the cross-entropy loss on mixup augmentation for clean clients. We use a small β value in the early rounds of this stage and only focus on the mixed loss correction training as shown in Eq. (9). Then we increase the β value to constrain the local training for handling the heterogeneous data across clients, together improving the performance of the global model for fault diagnosis.

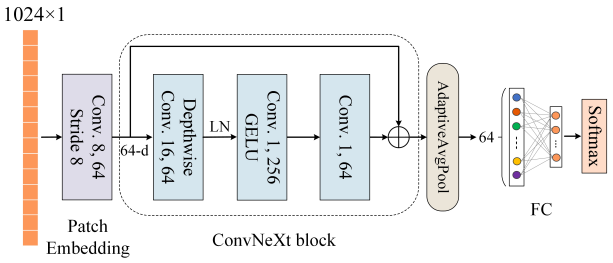

Fig. 2. Architecture of the compact ConvNet for fault diagnosis.

C. Fault Diagnosis Model in FedCNL

The CNN networks with residual structure have shown strong robustness to label noise in fault diagnosis [16], [28]. In our proposed FedCNL, a compact ConvNet with one ConvNeXt block [29] is implemented as fault diagnosis model for machine fault diagnosis, which is constructed with residual structure. The architecture of the compact ConvNet is shown

in Fig. 2. As illustrated in Fig. 2, the proposed architecture contains a patch embedding layer, a ConvNeXt block, an AdaptiveAvgPool layer, and a fully connected (FC) layer for classification. The ConvNeXt block consists of a depthwise convolution layer with layer normalization behind, a pointwise convolution layer activated by a Gaussian Error Linerar Unit (GELU), and another pointwise convolution layer connected with residual connection. Our proposed fault diagnosis model is compact and robust, and its architecture is shared by all clients and the server in our proposed FedCNL.

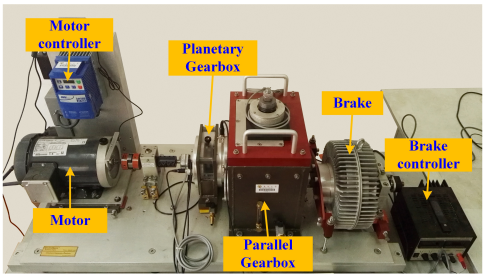

Fig. 3. The test rig of DDS.

TABLE I
PLANETARY GEARBOX CONDITION DESCRIPTIONS

Component	Type	Description
Bearing	Ball	Ball wear fault
	Combo	Combo wear fault in both inner and outer ring
	Inner	Inner race-wear fault
	Outer	Outer race-wear fault
Gear	Chipped	Broken teeth fault
	Miss	Missing teeth fault
	Root	Teeth root crack fault
	Surface	Surface wear fault

V. EXPERIMENT STUDY

A. Dataset Description

We conduct experiments on two fault datasets to investigate the effectiveness of our proposed FedCNL.

The PU dataset [30], [31], which is the bearing dataset from the Paderborn University Bearing Data Center, collects bearing data by a piezoelectric accelerometer on the top end of the rolling bearing module at a sampling frequency of 64 kHz. We select the vibration signals from 13 real damage bearings (KA04, KA15, KA16, KA22, KA30, KB23, KB24, KB27, KI14, KI16, KI17, KI18, and KI21) and 1 healthy bearing (K001) under the working condition N15_M07_F10 for experiments. Consequently, the PU dataset obtains 14 different class states. From each state, 500 samples are chosen for training and another 500 samples for testing.

The DDS dataset consists of vibration signals of planetary gearbox acquired by ourselves from the test rig of Spectra Quests Drivetrain Dynamics Simulator (DDS), as shown in Fig. 3. We acquires the bearing-gear fault data by the 608A11 vibrating sensors placed on the planetary gearbox at a sampling frequency of 5120 Hz under different operating conditions of 30Hz_4, 30Hz_2, 20Hz_0, and 40Hz_0 for experiments. Here 20Hz, 30Hz, and 40Hz indicate the motor rotation frequency, corresponding to the working speed of 1200 rpm, 1800 rpm, and 2400 rpm respectively, 0, 2 and 4 behind denote the corresponding load of 0 N.m, 3.66 N.m and 10.98 N.m. The planetary gearbox has 8 class faulty states listed in Table I and 1 healthy state. The DDS dataset includes 200 samples from each state under each operating condition for training and also 200 samples for testing. The data length of each sample in two datasets is both 1024.

B. Experimental Settings

We investigate the performance of our proposed FedCNL with different label noise levels under heterogeneous data partitions. The ways of data partition and label noise generation in FL scenarios are presented as follows.

1) Data Partition: Following the previous works [20], [21], Dirichlet distribution $\mathbf{Dir}(\alpha)$ is used to simulate the non-IID scenarios for heterogeneous data partitions, in which a smaller α indicates higher data heterogeneity. Through Dirichlet distribution, the local training data of different clients are varied in both class distribution and sample number. Here, we set $\alpha = 1$ to control the heterogeneity of data for non-IID data partition. As for the IID data partition, the whole training dataset is uniformly distributed across all clients. There are 10 clients setting for our FL experiments.

2) Noise Generation: To simulate label noise for data in experiments, we use the noise model in [20] for synthetic label noise generation among 10 clients. Two parameters ρ and τ are utilized to control the heterogeneous label noise, where ρ represents the probability of a client to be noisy and τ indicates the lower limit of the noise level in a noisy client. In this way, the noise ratio of clients is controlled by ρ and the noise level in a noisy client is randomly sampled from the uniform distribution $U(\tau, 1)$. The two parameters enable the modeling of heterogeneous label noise across clients.

TABLE II
HYPERPARAMETERS OF OUR PROPOSED FEDCNL

Hyperparameter	value	Hyperparameter	value
Learning rate	0.05	Mixup	1
Optimizer	SGD	Warm-up rounds (T_0)	20
Local epoches	5	Stage 1 rounds (T_1)	60
Local batch size	16	Stage 2 rounds (T_2)	PU 25 / DDS 30
Client number	10	Stage 3 rounds (T_3)	PU 45 / DDS 50

C. Compared Methods

Six different learning methods are implemented to handle the heterogeneous label noise in our experiments for comparison. (1) Centralized learning: The local data of all clients are centralized for model training. The mixed loss correction method is also used in the model training. (2) FedAvg: FedAvg [1] is the vanilla FL algorithm and can be viewed as the baseline method. (3) FedProx: FedProx [7] is also a classic FL method, which adds a local proximal term in the local training to constraint the local updates to be closer to the global model. (4) FedAvg+Mixup: The FedAvg algorithm is combined with the mixup augmentation to improve the robustness of model.

TABLE III
COMPARISON PERFORMANCE OF ALL APPROACHES UNDER THE NON-IID SETTING WITH HETEROGENEOUS LABEL NOISE

Dataset	Method	Test Accuracy(%)						
		ρ=0.0	ρ=0.3		ρ=0.5		ρ=0.7	
		τ=0.0	τ=0.0	τ=0.5	τ=0.0	τ=0.5	τ=0.0	τ=0.5
PU	Centralized learning	98.55±0.25	98.53±0.23	98.61±0.11	98.06±0.51	97.17±0.14	97.88±0.21	95.48±0.55
	FedAvg [1]	98.50±0.09	94.92±0.63	93.33±0.99	91.48±0.60	87.12±0.96	90.88±0.62	87.30±0.97
	FedProx [7]	98.13±0.18	94.97±0.67	94.48±0.18	91.87±1.04	86.15±0.80	91.12±1.14	86.24±1.25
	FedAvg+Mixup	98.22±0.23	97.24±0.10	96.59±0.22	95.19±0.32	92.70±0.65	93.74±0.43	89.04±0.24
	RHFL [17]	98.44±0.12	96.52±0.59	96.16±0.47	94.06±0.46	91.22±0.73	93.79±0.38	91.01±0.74
	FedCorr [20]	97.11±0.94	96.51±2.60	97.49±0.17	93.85±2.53	88.07±1.59	93.71±0.52	91.07±0.95
	FedNoRo [21]	98.25±0.10	97.59±0.17	98.18±0.18	95.81±0.28	94.29±1.04	94.45±0.68	95.78±0.63
	FedCNL(ours)	**98.80±0.07**	**98.41±0.06**	**98.45±0.14**	**97.53±0.25**	**97.54±0.14**	**96.68±0.22**	**96.67±0.44**
DDS	Centralized learning	99.08±0.37	97.97±1.58	96.43±0.59	90.43±1.14	89.05±1.60	84.96±2.66	67.02±1.62
	FedAvg [1]	92.47±0.17	78.45±0.56	75.20±1.07	61.45±1.15	53.66±2.15	59.46±2.70	51.54±0.89
	FedProx [7]	88.19±0.66	75.87±1.37	72.00±0.97	62.77±1.70	52.70±1.04	60.83±1.46	52.34±0.52
	FedAvg+Mixup	87.63±0.99	82.58±0.72	79.79±0.78	69.15±0.89	63.78±1.80	65.83±1.07	56.86±1.17
	RHFL [17]	88.28±0.80	77.29±1.11	73.37±3.01	67.13±2.01	59.70±1.17	64.42±1.92	54.72±1.47
	FedCorr [20]	75.83±4.69	82.16±2.82	79.65±3.90	73.36±2.82	59.70±2.82	62.71±2.75	56.88±7.85
	FedNoRo [21]	74.98±2.65	85.60±0.46	86.23±0.83	79.49±0.79	71.18±0.80	70.61±0.63	60.14±3.84
	FedCNL(ours)	**96.23±0.82**	**93.79±0.46**	**94.58±0.47**	**85.17±0.50**	**79.17±2.02**	**79.15±1.27**	**65.78±1.11**

TABLE IV
COMPARISON PERFORMANCE OF ALL APPROACHES UNDER NON-IID SETTING OF $\alpha = 0.5$ WITH HETEROGENEOUS LABEL NOISE

Dataset	Method	Test Accuracy(%)						
		ρ=0.0	ρ=0.3		ρ=0.5		ρ=0.7	
		τ=0.0	τ=0.0	τ=0.5	τ=0.0	τ=0.5	τ=0.0	τ=0.5
PU	Centralized learning	98.63±0.25	98.58±0.35	97.91±0.34	98.20±0.28	97.11±0.32	98.03±0.18	96.42±0.38
	FedAvg [1]	98.36±0.21	94.15±0.55	92.19±0.62	91.27±0.98	86.42±0.89	90.50±0.78	86.05±0.64
	FedProx [7]	97.86±0.11	94.46±0.65	92.91±0.37	90.23±0.72	84.63±0.85	90.08±0.96	85.59±0.97
	FedAvg+Mixup	96.94±0.44	94.56±0.40	95.08±0.65	93.06±0.56	88.05±0.56	92.35±0.51	88.46±0.97
	RHFL [17]	98.18±0.17	95.54±0.34	95.05±0.37	92.96±0.44	89.65±0.52	92.40±0.99	89.63±0.98
	FedCorr [20]	97.07±0.57	96.51±0.74	96.10±0.60	90.92±1.78	86.80±1.66	90.71±1.42	83.73±2.19
	FedNoRo [21]	97.61±0.26	96.60±0.20	97.46±0.16	94.63±0.16	95.84±0.66	93.22±0.44	93.96±0.69
	FedCNL(ours)	**98.55±0.10**	**97.92±0.06**	**97.92±0.30**	**96.20±0.47**	**96.29±0.26**	**95.41±0.57**	**95.57±0.97**
DDS	Centralized learning	98.93±0.13	98.82±0.19	98.84±0.15	96.60±4.21	90.60±1.85	88.26±6.76	76.86±5.24
	FedAvg [1]	90.36±2.01	76.83±1.87	67.61±1.57	59.48±0.50	53.16±1.51	53.02±1.92	43.15±0.87
	FedProx [7]	86.47±1.29	74.26±1.67	66.69±1.79	58.49±1.76	56.07±1.89	51.04±0.85	41.50±3.21
	FedAvg+Mixup	84.47±1.19	74.09±1.44	67.58±1.27	67.46±0.53	59.39±1.62	61.61±1.02	50.85±0.75
	RHFL [17]	87.20±1.70	77.50±1.33	71.50±1.11	62.64±1.10	57.851.54	57.79±1.42	47.35±2.38
	FedCorr [20]	49.37±6.01	38.45±15.55	31.96±7.85	50.51±4.77	52.18±3.33	49.29±8.61	35.40±5.67
	FedNoRo [21]	84.36±3.11	81.02±0.69	77.73±1.77	72.98±1.52	63.64±0.87	63.48±1.01	62.92±1.10
	FedCNL(ours)	**93.51±0.50**	**88.09±0.45**	**80.64±1.00**	**79.76±0.42**	**74.45±0.60**	**76.83±1.14**	**73.36±2.82**

(5) RHFL: RHFL is the method proposed in [17]. Here, we use the symmetric cross entropy learning (SL) loss and the client confidence re-weighting scheme to reduce the negative effects caused by label noise. (6) FedCorr: FedCorr is a label correction method proposed in [20], which identifies noisy clients via LID scores and relabels the identified noisy samples with the labels predicted by the global model. (7) FedNoRo: FedNoRo is the method proposed in [21], which selects noisy clients based on per-class average loss values and applies knowledge distillation for noisy client training.

The experimental results of all approaches are the average of five random runs, including the average accuracies and the standard deviations. For the four no-client selection methods, as FedAvg, FedProx, FedAvg+Mixup, and RHFL, the federated training is terminated when the global model reaches a local optimal. For client selection based methods, as FedCorr, FedNoRo and our proposed FedCNL, the federated training is terminated after the same number of communication rounds. The values of hyperparameters setting in our proposed FedCNL are presented in Table II. The hyperparameter β that controls the local proximal regularization term is initialized to be 0, and increased to be 0.5 for the last 30 communication rounds. Our works are all 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 @3.20GHz, 24.0GB RAM, and GPU NVIDIA GeForce RTX 3080, 10GB.

D. Experimental Results

1) Comparison with State-of-the-art Methods: The test accuracy results of our proposed FedCNL and all compared approaches on PU and DDS datasets under the non-IID setting are presented in Table III. Additionally, we also provide comparison results under a higher data heterogeneity scenario ($\alpha = 0.5$) in Table IV. It can be seen that our proposed FedCNL achieves the best performances across all noise settings on both datasets. The performance of the Centralized learning method can be regarded as upper limit of fault diagnosis performance.

Among federated methods, the two classic FL algorithms, FedAvg and FedProx, exhibit similar test accuracies, both of which are significantly affected by label noise. When combined with mixup augmentation, the test accuracy of FedAvg improves under most noise settings. Meanwhile, the RHFL method, which also applies the same training in all clients for label noise, achieves some performance gains compared with the performance of FedAvg. FedCorr and FedNoRo, both of which select clients for different training, also achieve performance gains. Notably, FedNoRo is more effective and robust under the higher data heterogeneity scenario ($\alpha = 0.5$), whereas FedCorr performs worse under this scenario. It may be because FedCorr fails to correctly filter out noisy clients in highly heterogeneous scenarios, leading to incorrect label corrections. Among the FL methods, our proposed FedCNL achieves the largest performance gains under different label noise settings in both non-IID scenarios. The performances of all approaches over different values for (ρ, τ) on DDS dataset under the IID setting are also evaluated in Table V. Our proposed FedCNL also outperforms the other FL methods under different label noise settings in the IID setting.

TABLE V
RESULTS ON DDS DATASET UNDER THE IID SETTING

Method	Test Accuracy(%)			
	ρ=0.3		ρ=0.5	
	τ=0.0	τ=0.5	τ=0.0	τ=0.5
Centralized Learning	98.82±0.31	98.72±0.49	98.87±0.27	96.66±0.85
FedAvg [1]	92.09±0.47	88.01±0.94	81.82±0.63	71.35±1.51
FedProx [7]	91.86±0.36	86.06±1.06	81.74±1.68	70.45±1.78
FedAvg+Mixup	95.94±0.14	93.58±0.40	87.97±1.34	80.58±1.95
RHFL [17]	94.18±0.22	91.80±0.83	86.38±1.32	76.93±5.60
FedCorr [20]	98.24±0.76	98.27±0.38	97.63±0.49	95.98±0.98
FedNoRo [21]	94.44±0.46	96.70±0.48	91.28±0.29	94.28±0.29
FedCNL(ours)	**98.28±0.19**	**98.29±0.12**	**97.71±0.14**	**97.30±0.31**

To evaluate our proposed FedCNL in more non-IID settings, we vary the non-IID settings via adjusting the parameter α in Dirichlet distribution, where a smaller α indicates the higher data heterogeneity. And we compare our proposed FedCNL with the FL methods at two noise levels. The experimental results under different non-IID settings are described in Table VI. It can be seen that our proposed FedCNL consistently outperforms the other FL methods, especially at high data heterogeneity level. For DDS dataset that contains the bearing-gear data under varied operating conditions, the performance gains of our proposed FedCNL in the non-IID settings are more obvious.

Our proposed FedCNL utilizes the client-wise training loss values for GMM modeling to identify the clean and noisy clients, while FedCorr computes the LID scores for modeling. To investigate the effectiveness of these two methods for client identification, we use the local training loss values instead of LID scores in FedCorr to ensure fair experiments in the non-IID setting. The comparison results are presented in Table VII. It can be seen that the client-wise training loss values are more effective than LID scores in identifying the noisy clients. This may be because the model training loss is directly obtained through the supervised training with the labels, while the LID scores are computed based on distances of k nearest neighbors.

TABLE VI
ACCURACIES OF DIFFERENT FL METHODS UNDER DIFFERENT NON-IID SETTINGS WITH HETEROGENEOUS LABEL NOISE

Dataset	Method	Test Accuracy(%)			
		α=0.1		α=6	
		ρ=0.3,τ=0.5	ρ=0.5,τ=0.5	ρ=0.3,τ=0.5	ρ=0.5,τ=0.5
PU	FedAvg+mixup	82.74±0.56	74.47±2.25	97.49±0.35	93.91±0.32
	RHFL [17]	86.30±0.95	76.54±2.01	95.95±0.56	91.54±1.24
	FedCorr [20]	79.30±2.45	63.82±5.52	98.37±0.18	98.06±0.21
	FedNoRo [21]	90.86±0.58	78.36±1.59	98.48±0.27	97.72±0.15
	FedCNL(ours)	**94.34±0.29**	**89.68±0.90**	**98.82±0.15**	**98.10±0.17**
DDS	FedAvg+mixup	55.70±0.82	44.22±0.79	85.94±0.38	73.79±1.32
	RHFL [17]	55.17±1.36	40.24±1.51	82.29±1.46	65.57±3.10
	FedCorr [20]	53.92±3.33	31.80±3.35	89.63±2.62	73.59±3.78
	FedNoRo [21]	55.74±2.43	46.29±2.58	91.07±0.72	79.19±1.50
	FedCNL(ours)	**73.13±3.59**	**54.57±0.76**	**96.16±0.43**	**92.08±0.91**

The model training losses can more directly reflect the label noise information, while the LID scores would be significantly influenced by the heterogeneous data distributions in clients. This is also consistent with the results in Table IV and VI that FedCorr works poorly in high data heterogeneity scenarios. Besides, our proposed FedCNL which designs a multi-stage curriculum learning strategy considering both client and sample learning complexities surpasses FedCorr(loss) a lot.

TABLE VII
COMPARISONS BETWEEN THE TRAINING LOSS VALUES AND THE LID SCORES

Dataset	Method	Test Accuracy(%)			
		ρ=0.3		ρ=0.5	
		τ=0.0	τ=0.5	τ=0.0	τ=0.5
PU	FedCorr(LID)	96.51±2.60	97.49±0.17	93.85±2.53	88.07±1.59
	FedCorr(loss)	98.13±0.31	98.27±0.17	94.76±0.81	89.49±0.45
	FedCNL(ours)	**98.41±0.06**	**98.45±0.14**	**97.53±0.25**	**97.54±0.14**
DDS	FedCorr(LID)	82.16±2.82	79.65±3.90	73.36±2.82	59.70±2.82
	FedCorr(loss)	82.71±4.46	81.18±3.11	78.45±2.06	62.94±2.71
	FedCNL(ours)	**93.79±0.46**	**94.58±0.47**	**85.17±0.50**	**79.17±2.02**

2) Communication Efficiency: To compare the communication efficiency of FL methods, we show the learning curves of different methods with certain communication rounds under noise level of $(\rho, \tau) = (0.5, 0.5)$ in the non-IID setting in Fig. 4, where our proposed FedCNL achieves distinct performance gain after 140 rounds on both datasets. FedAvg, FedAvg+Mixup, and RHFL methods, which use the same training for all clients, reach the local optimal after about 30, 50 and 40 communication rounds, respectively. The average training times of one random experiment to reach the local optimal in these three methods are 3.22 minutes (min), 5.95 min and 5.34 min, respectively. Then the global model becomes overfitted to the noisy labels, leading to a decline of the test accuracy. FedCorr, FedNoRo and our proposed FedCNL, which select clients first, need more communication rounds to obtain robust performances. The average training times of one random experiment in FedCorr, FedNoRo and our proposed FedCNL are 20.14 min, 26.94 min and 19.57 min, respectively. Our proposed FedCNL achieves the highest classification accuracy while taking the least training time among the three client selection based methods. Through the multi-stage curriculum learning, our proposed FedCNL can gradually improve the classification performance of the global

TABLE VIII
IMPACTS OF EACH COMPONENT OF THE MULTI-STAGE CURRICULUM LEARNING STRATEGY IN FEDCNL

non-IID setting	Method	Test Accuracy(%)							
		PU dataset				DDS dataset			
		ρ=0.3,τ=0.5	ρ=0.5,τ=0.0	ρ=0.5,τ=0.5	ρ=0.7,τ=0.5	ρ=0.3,τ=0.5	ρ=0.5,τ=0.0	ρ=0.5,τ=0.5	ρ=0.7,τ=0.5
$\alpha = 1$	**FedCNL(ours)**	**98.45±0.14**	**97.53±0.25**	**97.54±0.14**	**96.67±0.44**	**94.58±0.47**	**85.17±0.50**	**79.17±2.02**	**65.78±1.11**
	w/o mixup	97.14±1.14	92.38±0.90	94.00±1.34	92.96±0.58	89.28±1.12	79.87±0.57	70.23±0.89	60.94±0.69
	w/o curriculum learning	97.18±0.14	96.53±0.28	95.28±0.47	92.88±0.92	87.52±1.47	79.61±1.21	71.82±1.50	56.60±5.09
	w/o stage 2 + stage 3	97.54±0.13	95.03±0.13	89.52±0.40	92.24±1.08	85.18±1.10	75.16±1.86	66.91±1.98	55.98±2.33
	w/o stage 3	97.90±0.12	96.05±0.34	94.14±0.57	95.10±0.44	86.39±0.83	76.71±0.66	70.39±2.27	59.31±1.05
	w/o loss correction	97.90±0.19	95.73±0.31	94.66±0.47	90.97±0.78	91.55±0.90	80.99±1.01	72.27±0.89	62.64±1.70
	w/o local proximal	97.57±0.31	96.59±0.57	95.92±0.48	95.77±0.38	86.40±0.90	79.33±1.29	71.44±1.11	60.50±1.84
$\alpha = 0.5$	**FedCNL(ours)**	**97.92±0.30**	**96.20±0.47**	**96.29±0.26**	**95.57±0.97**	**80.64±1.00**	**79.76±0.42**	**74.45±0.60**	**73.36±2.82**
	w/o mixup	95.63±0.50	91.75±0.66	93.57±0.51	93.48±0.52	80.35±1.15	73.50±1.17	65.62±1.70	57.64±4.01
	w/o curriculum learning	96.34±0.15	95.40±0.40	92.80±0.96	92.99±0.43	74.35±1.51	78.08±1.51	64.57±1.44	47.84±6.90
	w/o stage 2 + stage 3	95.68±0.53	94.34±0.56	91.16±1.56	88.28±1.17	69.15±1.31	72.31±1.61	55.70±1.48	57.01±2.46
	w/o stage 3	96.19±0.12	95.03±0.23	92.67±0.68	90.55±0.74	70.41±0.73	73.57±0.80	62.10±0.96	58.41±2.33
	w/o loss correction	96.70±0.21	94.56±0.69	93.37±0.20	91.35±0.69	80.27±1.24	72.12±1.53	67.61±0.84	60.55±2.32
	w/o local proximal	95.81±0.47	93.44±0.68	93.05±0.36	92.02±0.63	72.43±0.73	72.53±0.36	67.78±1.81	67.60±2.87

model under heterogeneous label noise across clients and obtain high efficiency.

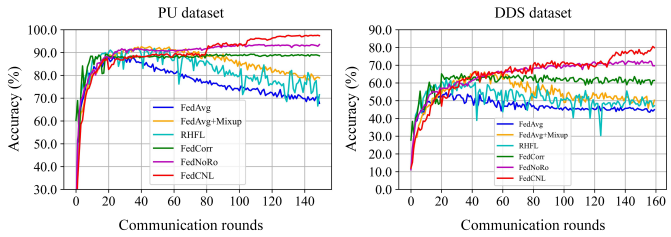

Fig. 4. The learning curves of different methods with certain communication rounds.

3) Ablation Study: **Impacts of each component in FedCNL:** We evaluate six different learning strategies in non-iid settings by removing each component of the multi-stage curriculum learning strategy in the proposed FedCNL, which are described as follows. w/o mixup: The mixup augmentation is not applied in FedCNL. w/o curriculum learning: The global model is directly trained with stage 3 using all clients with all local data. w/o stage 2 + stage 3: The global model is only trained in stage 1 using only the clean clients. w/o stage 3: The global model is trained with the first two-stage curriculum learning. w/o loss correction: The bootstrapping loss for loss correction is removed in stage 3. w/o local proximal: The local proximal regularization term is removed in stage 3. The experimental results are presented in Table VIII.

It is evident that removing any component in the multi-stage curriculum learning strategy of our proposed FedCNL would lead to degraded performances. Our proposed FedCNL with a three-stage curriculum learning significantly outperforms the method w/o curriculum learning that directly trains with all clients, especially under higher levels of label noise. Training our proposed FedCNL only with clean clients (w/o stage 2 + stage 3) has the worst performance under higher levels of label noise, due to the small sizes of datasets in clean clients. The proposed FedCNL w/o stage 3 curriculum, which adds the clean datasets in noisy clients into federated training, improves the test accuracy. However, it suffers from accuracy drops since it discards the noisy data in noisy clients. The proposed loss correction and the local proximal regularization in stage 3 also contribute to improving the accuracy of global model for fault diagnosis, both of which lead to performance degradation when removed individually.

In stage 3, the value of β is dynamically adjusted to make the model adapt to the noisy data first and subsequently combat the data heterogeneity. The learning curve of our proposed FedCNL on DDS dataset in stage 3 under the non-IID setting is presented in Fig. 5 to intuitively show the impacts of β value before and after adjustment. It can be seen that adding β value which controls the local proximal regularization on the basis of loss correction can further improve the performance of the global model for fault diagnosis in heterogeneous FL.

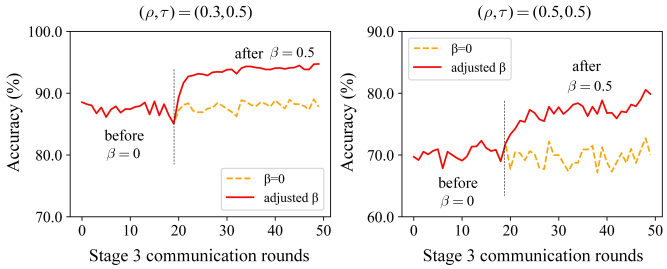

Fig. 5. The impacts of β before and after adjustment on DDS dataset.

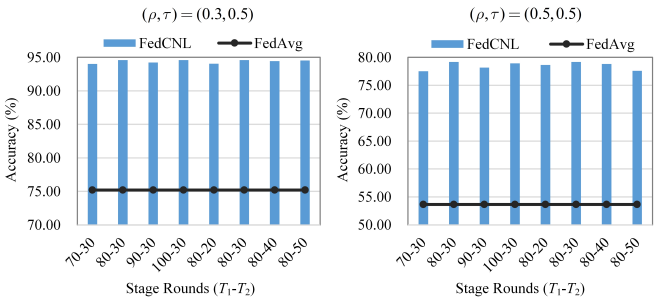

Fig. 6. The impacts of stage rounds on DDS dataset with different noise levels.

Impacts of Stage Rounds: We conduct experiments to

investigate the impacts of communication rounds on global model convergence in each curriculum learning stage of our proposed FedCNL, using DDS dataset under noise levels of $(\rho, \tau) = (0.3, 0.5)$ and $(\rho, \tau) = (0.5, 0.5)$ in the non-IID setting. We change the stage rounds T_1 and T_2 to investigate the robustness of our proposed FedCNL during stage transitions. Specifically, we adjusted T_1 from 70 to 100 with a stride of 10 while keeping the other stages constant to study the effects of the stage rounds on transition from stage 1 to stage 2. Similarly, the value of T_2 is also varied from 20 to 50 with a stride of 10 to study the effects of the stage rounds on transition from stage 2 to stage 3. The results are shown in Fig. 6. It can be seen that adjustments in stage rounds during transitions from near convergence to over-convergence have minimal impact on the performance of the global model. Our proposed FedCNL method demonstrates robustness against fluctuations in stage rounds, consistently outperforming the baseline method FedAvg in combating label noise.

VI. Conclusion

This paper has proposed a novel approach FedCNL for machine fault diagnosis in FL with label noise, where a multi-stage curriculum learning strategy is designed for tackling the heterogeneous label noise across clients. The proposed FedCNL have exploited a noise modeling module via GMM models to adaptively estimate the noisy clients and identify the clean samples and noisy label samples in noisy clients in an unsupervised manner. Then, the multi-stage curriculum learning strategy has been designed by regarding the noise level as learning complexity, which enables the model to learn from clean to noisy samples, gradually improving the performance of the global model. Moreover, a mixed loss correction method has been explored in the curriculum learning stage to maximize the utilization of data with noisy labels. The machine fault diagnosis experiments performed on two fault datasets with non-IID settings under different label noise levels have verified the effectiveness of our proposed FedCNL method for machine fault diagnosis in FL with heterogeneous label noise. Our proposed FedCNL has outperformed the state-of-the-art methods for federated learning with label noise with significant performance gains.

Although our proposed FedCNL method has effectively addressed label noise issues, it is evident that the classification performances on DDS dataset are worse than that on PU dataset due to its variable operating conditions within and across clients. In fact, the variable operating conditions in practical engineering further increase the difficulty of federated learning for machine fault diagnosis. In the future work, we will explore the domain generalization and domain adaptation methods for federated machine fault diagnosis, and further address the issue of variable operating conditions in FL for practical engineering.

References

[1] 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.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[9] H. Wang and Y.-F. Li, "Iterative error self-correction for robust fault diagnosis of mechanical equipment with noisy label," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[10] Y. Chen, X. Yang, X. Qin, H. Yu, B. Chen, and Z. Shen, "Focus: Dealing with label quality disparity in federated learning," *arXiv preprint arXiv:2001.11359*, 2020.

[11] S. Ke, C. Huang, and X. Liu, "Quantifying the impact of label noise on federated learning," *arXiv preprint arXiv:2211.07816*, 2022.

[12] E. Arazo, D. Ortego, P. Albert, N. OConnor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," in *International conference on machine learning*. PMLR, 2019, pp. 312–321.

[13] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang, "Curriculumnet: Weakly supervised learning from large-scale web images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 135–150.

[14] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596*, 2014.

[15] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[16] K. Zhang, B. Tang, L. Deng, Q. Tan, and H. Yu, "A fault diagnosis method for wind turbines gearbox based on adaptive loss weighted meta-resnet under noisy labels," *Mechanical Systems and Signal Processing*, vol. 161, p. 107963, 2021.

[17] X. Fang and M. Ye, "Robust federated learning with noisy and heterogeneous clients," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 072–10 081.

[18] S. Yang, H. Park, J. Byun, and C. Kim, "Robust federated learning with noisy labels," *IEEE Intelligent Systems*, vol. 37, no. 2, pp. 35–43, 2022.

[19] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.

[20] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "Fedcorr: Multi-stage federated learning for label noise correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 184–10 193.

[21] N. Wu, L. Yu, X. Jiang, K.-T. Cheng, and Z. Yan, "Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity," *arXiv preprint arXiv:2305.05230*, 2023.

[22] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[24] C. Cheng, X. Liu, B. Zhou, and Y. Yuan, "Intelligent fault diagnosis with noisy labels via semi-supervised learning on industrial time series," *IEEE Transactions on Industrial Informatics*, 2023.

[25] Z. Wang, T. Zhou, G. Long, B. Han, and J. Jiang, "Fednoil: A simple two-level sampling method for federated learning with noisy labels," *arXiv preprint arXiv:2205.10110*, 2022.

[26] X. Ji, Z. Zhu, W. Xi, O. Gadyatskaya, Z. Song, Y. Cai, and Y. Liu, "Fedfixer: Mitigating heterogeneous label noise in federated learning,"

in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 830–12 838.
[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
[28] H. Ruan, Y. Wang, Y. Qin, and B. Tang, "An enhanced intelligent fault diagnosis method to combat label noise," in *2021 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*. IEEE, 2021, pp. 1–6.
[29] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.
[30] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *PHM Society European Conference*, vol. 3, no. 1, 2016.
[31] "C. lessmeier kat-datacenter, chair of design and drive technology. paderborn, germany: Paderborn university, 2019. [online]. available:," https://mb.uni-paderborn.de/kat/forschung/datacenter/bearingdatacenter.

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 (F22) 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, AI, 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). His honors and awards include the IEEE Instrumentation and Measurement Society Technical Award in 2019 and Outstanding Service Award in 2022, and multiple best paper awards. Dr. Yan serves as the Editor-in-Chief of the IEEE Transactions on Instrumentation and Measurement. He is also an Editorial Board Member of Chinese Journal of Mechanical Engineering.

Ruibing Jin received the B.Eng. degree from the University of Electronic Science and Technology of China, Chengdu, 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 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 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.