Abstract—Federated learning (FL), training deep models from decentralized data without privacy leakage, has drawn great attention recently. Two common issues in FL, namely data heterogeneity from the local perspective and class imbalance from the global perspective have limited FL's performance. These two coupling problems are under-explored, and existing few studies may not be sufficiently realistic to model data distributions in practical scenarios (e.g., medical scenarios). One common observation is that the overall class distribution across clients is imbalanced (e.g., common vs. rare diseases) and data tend to be agglomerated to those more advanced clients (i.e., the data agglomeration effect), which cannot be modeled by existing settings. Inspired by real medical imaging datasets, we identify and formulate a new and more realistic data distribution denoted as \( L^2 \) distribution where global class distribution is highly imbalanced and data distributions across clients are imbalanced but forming a certain degree of data agglomeration. To pursue effective FL under this distribution, we propose a novel privacy-preserving framework named FedIIC that calibrates deep models to alleviate bias caused by imbalanced training. To calibrate the feature extractor part, intra-client contrastive learning with a modified similarity measure and inter-client contrastive learning guided by shared global prototypes are introduced to produce a uniform embedding distribution of all classes across clients. To calibrate the classification heads, a softmax cross entropy loss with difficulty-aware logit adjustment is constructed to ensure balanced decision boundaries of all classes. Experimental results on publicly-available datasets demonstrate the superior performance of FedIIC in dealing with both the proposed realistic modeling and the existing modeling of the two coupling problems.

Index Terms—Federated Learning, Class Imbalance, Contrastive Learning, Data Agglomeration

I. INTRODUCTION

Despite the remarkable success of deep learning (DL) in computer vision [1], natural language processing [2], etc., it has reached the consensus that DL is a data-driven technology and collecting a large amount of training data is often the first and the most significant step. However, it’s infeasible in practice to fuse multiple data sources for large-scale data collection due to privacy concerns, especially in medical or financial scenarios. Fortunately, federated learning (FL), allowing multiple clients to jointly train a unified deep learning model without data sharing [3], [4], is recently proposed to address this issue. In vanilla FL, each participating client first trains a model locally based on its data following standard DL and then collaborates with others through parameter/gradient aggregation for global/federated model update [5]. In this way, only clients’ model parameters/gradients, instead of raw training data, are shared, which is privacy-preserving.

N. Wu, L. Yu, X. Yang, and Z. Yan are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China (e-mail: z.yan@connect.ust.hk).
K.-T. Cheng is with the School of Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong.
Corresponding author: Zengqiang Yan
Given that data is individually collected by each client in FL, there widely exists data heterogeneity [9] in both feature space (e.g., varying intensity, contrast, etc.) [10]–[15] and class space (e.g., varying class distributions) [7], [16]–[18] across different clients. In this paper, we focus on the latter one and analyze different data heterogeneity settings from the following two perspectives.

Global: Global data, i.e., the aggregation of all local data, is FL usually is assumed to be balanced like CIFAR-10 [19], CIFAR-100 [19], Tiny-Imagenet [20], etc. However, in practice, it is infeasible to ensure uniform global distributions, which is under-explored. Few most-rent works propose to address data imbalance by formulating global data into a long-tailed class distribution with varying class probabilities [8].

Local: In addition to global class imbalance, local class and data distributions across clients can also be imbalanced. To simulate this, each class follows the same Dirichlet distribution (e.g., $\text{Dir}(\alpha = 1.0)$), where $\alpha$ defines the concentration of each sampled distribution) for data partitioning to clients [7], [8], [15], [21]. By repeating the above data allocation process for each class independently (i.e., data partitioning of classes can be different but share similar concentration degrees), local class and data distributions become non-IID (i.e., non-independent identically distribution). This data distribution implicitly assumes all classes and clients are equally important.

Based on the above global and local distributions, we reformulate existing data heterogeneity settings as follows:

1) $U$ distribution. Global: uniform class distribution. Local: non-IID class and data distributions following a Dirichlet distribution as illustrated in the second row of Fig. [1].

2) $L$ distribution. Global: long-tailed class distribution. Local: non-IID local class and data distributions following a Dirichlet distribution (i.e., the same as $U$ distribution) as illustrated in the third row of Fig. [1].

Unfortunately, both $U$ and $L$ distributions may not always be realistic due to the following observations in clinical or financial scenarios:

1) Global: Following a uniform or a standard long-tailed distribution can hardly model complicated class distributions. In clinical scenarios, diseases can differ dramatically in morbidity (e.g., common vs. rare diseases), resulting in highly-imbalanced class distributions as illustrated in the first row (i.e., the global part) of Fig. [1].

2) Local: Classes and clients are not equally important. One common observation in clinical or financial scenarios is the data agglomeration phenomenon that the more advanced a medical/financial institution is, the more patients/customers it owns of a specific class. Furthermore, rare classes can be more concentrated than common classes, as the identification of rare classes highly depends on clients’ expertise. Consequently, common classes can distribute to nearly all clients while rare classes only exist in few top institutions as illustrated in the first row (i.e., the local part) of Fig. [1].

In this paper, we, for the first time, identify and formally formulate the new imbalanced distribution in FL as $L^2$ distribution: non-uniformly sampled long-tailed global class distribution and non-IID local class and data distributions coupled with data agglomeration as illustrated in the fourth row of Fig. [1]. To pursue effective FL under this distribution, we propose a novel FL framework FedIIC to construct a balanced and unified objective for all clients to simultaneously address data and class imbalance. Specifically, we first introduce supervised contrastive learning (SCL) [22] with an adaptive temperature re-weighting the cosine similarity according to the data amount of each two contrast categories for balanced learning of each local client. Then, a set of class-wise global embeddings are constructed by aggregating and fine-tuning the knowledge of all clients and sent to each client to assist SCL between each sample embedding and global embeddings. In this way, each client calibrates its own embedding space through the knowledge from others for balanced learning. Furthermore, we propose a difficulty-aware logit adjustment method to calibrate the linear classifier of each local model. It should be noted that the establishment of global embeddings in FedIIC does not require clients to upload extra information like previous studies [8], [16], [23], which helps protect data privacy and reduce communication overhead.

The main contributions can be summarized as follows:

1) We identify the special data distribution in medical or financial scenarios as $L^2$ distribution. Compared to both $U$ and $L$ distributions, $L^2$ distribution discussed in this paper is more realistic.

2) We present a novel privacy-preserving framework FedIIC for imbalanced federated learning which is effective for not only the $L^2$ distribution but also the $L$ distribution [8].

3) Superior performance against the state-of-the-art approaches on federated learning with imbalanced data for medical and natural image classification.

II. RELATED WORK

A. Federated Learning

Federated Learning (FL) [5] has drawn great attention due to its privacy-preserving nature [3], [4]. In FL, one major concern is data heterogeneity across clients affecting federated convergence [24], [25]. To address this, FedProx [24] introduced a loss term measuring the distance between the global model and each local model to avoid the local training process being dominant by local data. SCAFFOLD [25] introduced a control variate to correct the update direction of local models. MOON [7] employed contrastive learning (CL) in the model level to constrain the updated model to be close to the global model while away from the previous local model. Mu et al. [16] created a set of prototypes for each class which was used to guide the update for heterogeneous clients. Zhang et al. [17] re-formulated this problem from the perspective of class imbalance and established uniform optimization objectives for all clients via logit calibration. Li et al. [18] focused on an extreme non-IID case that each client could only own partial categories of data, and presented a restricted softmax function. Except for the heterogeneity of class prior probabilities, prior studies also focused on the heterogeneity of cross-client training images and minimized...
image heterogeneity through either feature-level alignment or image-level alignment. Feature-level alignment among cross-client imaging data can be implemented by generative adversarial networks (GAN) \cite{26} and local batch normalization \cite{27}. As for image-level alignment, typical solutions are based on synthesis \cite{28} in the image space and augmentation in the frequency domain \cite{29, 30}.

B. Imbalanced Learning

Given the prevalence of imbalance in real-world datasets, imbalanced learning has been extensively studied in computer vision. Conventional solutions to deal with this problem mainly focus on re-balancing the imbalanced data via re-sampling \cite{31–33} and re-weighting \cite{34–36} to increase the sampling frequency and loss weights of minority classes respectively, which improves the performance of minority classes at the cost of degrading majority classes \cite{37}. In addition to re-balancing, some studies suggested setting different margins for different classes in cross-entropy (CE) losses \cite{38, 39}. Margin is a monotonically decreasing function of class prior probabilities, enforcing the network to emphasize more on the tail classes. Another type of solution toward imbalanced learning is decoupling \cite{37, 40}. For instance, Kang et al. \cite{37} proposed to decouple an integral learning process into two separate stages: representation learning and classifier learning. BBN \cite{40} used a bilateral-branch network to focus on the majority classes and minority classes respectively. Both of these trained relatively balanced classifiers with clever sampling methods instead of exploring how to obtain better representations to boost performance even further. Fortunately, thanks to the superior performance of supervised contrastive learning (SCL) \cite{22} in representation learning, it has been proven to be effective to handle class-imbalance \cite{41–45}. Hybrid-SC \cite{42} introduced a two-branch network and performed feature learning and classifier learning successively. PaCo \cite{43} introduced a set of learnable class-wise centers to make SCL more friendly to imbalanced learning. KCL \cite{44} sampled a uniform number of positives for each class to avoid SCL being biased to head classes. The two most-recent approaches TSC \cite{44} and BCL \cite{45} inspire us the most, both are re-defined as follows:

C. Federated Learning with Imbalanced Data

Class imbalance is common yet under-explored in real-world FL applications. Wang et al. \cite{46} presented a ratio loss (i.e., a weighted form of CE loss) depending on a balanced auxiliary dataset for the server to calculate weights, which was infeasible in most cases. Yang et al. \cite{47} presented an algorithm based on a multi-arm bandit to select clients for imbalance minimization. Jiang et al. \cite{48} expanded the problem to the paradigm of semi-supervised learning and addressed it via a dual bank learning scheme. CRoFF \cite{8} and CLIMB \cite{49} are the most-recent approaches. Inspired by decoupling \cite{37}, CRoFF retrained a new classifier with balanced synthetic features in the server to improve the overall performance of classification. However, feature synthesis required clients to upload the gradients of the classifiers during training, which may cause privacy leakage under well-designed attacks. CLIMB assigned larger weights to clients more likely to own minority classes via a meta-algorithm. However, it may fail to deal with more general data distributions, as the basic unit of weighting is a client instead of a class or a sample.

III. Methodology

A. Background and Definitions

Given \( K \) participants in FL, we assume each participant owns a private dataset \( D_k = \{ (x_i, y_i) \}_{i=1}^{N_k}, k \in \{1, 2, ..., K\} \), where \( N_k \) is the data amount of \( D_k \), and denote each image-label pair as \( (x_i \in \mathbb{R}^d, y_i \in \mathbb{Y} = \{1, 2, ..., L\}) \). In addition, we denote the global dataset as \( D_g = \{ (x_i, y_i) \}_{i=1}^{\sum_{k=1}^{K} N_k} \) for distribution analysis which is the fusion of all local datasets though it is inseparable due to data privacy in FL.

**Definition 1 (Global and Local Class Imbalance).** Following \cite{46}, we define two types of class imbalance in FL. For both the global dataset and each local dataset whose class prior probability is denoted as \( p(y) = [p^1, p^2, ..., p^L], \sum p^i = 1 \), it is regarded as imbalanced if \( p(y) \) is not uniform and the degree of imbalance is expressed as \( \Gamma = \frac{\max(p(y))}{\min(p(y))} \).

**Definition 2 (Data Concentration)** Assume all \( K \) clients follow a descending order in data amount, i.e. \( N_1 \leq N_2 \leq ... \leq N_K \). For any category/class \( c \in \mathbb{Y} \), assuming samples are distributed to \( K \) clients with probability \( p^c = [p^c_1, p^c_2, ..., p^c_K] \), data distribution is concentrated if \( p^c_1 \leq p^c_2 \leq ... \leq p^c_K \).

**Definition 3 (Degree of Data Concentration)** If samples of any class \( c \) are concentrated in just few clients (i.e., \( p^c_i \in p^c \) differs dramatically), the probability distribution \( [p^c_1, p^c_2, ..., p^c_N] \) is of low entropy. Following this, the degree of data concentration is defined as \( \text{Conc}^c = 1/\text{entropy}(\{p^c_1, p^c_2, ..., p^c_N\}) \).

Base on the above definitions, \( U \), \( L \), and \( L^2 \) distributions are re-defined as follows:

**U distribution:** With local imbalance, but without global imbalance and data concentration \cite{7}:

**L distribution:** With both global and local imbalance, but without data concentration \cite{8}:

**L^2 distribution:** With both global and local imbalance and data concentration. Degrees of data concentration can vary across classes.

From the perspective of data heterogeneity, \( L^2 \) distribution can be viewed as a specialized form of \( L \) distribution which is more realistic, especially in clinical scenarios.

B. Overview

For robust federated learning with \( L^2 \) distribution, we propose FedILC as illustrated in Fig. 2. The key idea is to calibrate both the feature extractor and the classification head which may be badly affected by data imbalance. To calibrate the feature extractor, we introduce 1) intra-client contrastive learning with a modified similarity measure to alleviate local class imbalance and 2) inter-client contrastive learning with shared global prototypes to produce a uniform embedding space for all classes across clients. To calibrate the classification head, we propose a difficulty-aware logit
adjustment algorithm to produce balanced decision boundaries for all classes. Details are presented in the following.

C. Intra-Client Contrastive Learning

The goal in FL is to obtain a well-performed global classification model $F \leftarrow f(g(X))$, where $g(\cdot)$ and $f(\cdot)$ represent the feature extractor and the linear classifier respectively. One straightforward solution is to pursue well-performing models during the local training phase before aggregation. Unfortunately, as local data is often limited and imbalanced, simply imposing a cross entropy loss may be sub-optimal. Therefore, we introduce supervised contrastive learning (SCL) to assist local training, which is proven to be effective for representation learning \cite{41}. The basic loss function of SCL can be formulated as

$$L_{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log \left( \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \right), \quad (1)$$

where $I$ denotes the index set of the multiviewed batch, $|\cdot|$ represents the number of elements in a set, $A(i) = I \setminus \{i\}$, $P(i) = \{s \in A(i) \mid y_s = y_i\}$, $\tau$ represents the temperature, and $z$ denotes the $l_2$-normalized embedding of a sample. Note that in this paper, we use a 2-layer MLP $h(\cdot)$ to obtain $z$ before it is normalized as \cite{51}, i.e., $z = \frac{h_i(z)}{||h_i(z)||_2}$. The key idea of SCL is to keep the embeddings of the same class closer (measured by the cosine similarity) while pushing the embeddings of different classes further away, producing a compact embedding space for the downstream task (e.g., image classification).

However, SCL cannot perfectly address class imbalance from the perspective of optimization. Assuming there are both majority classes and minority classes in the dataset, the fastest paths to minimize $L_{SCL}$ are:

1) Increasing the intra-class similarity of a majority class;
2) Decreasing the inter-class similarity of any two different majority classes.

Therefore, SCL tends to focus more on the majority classes like traditional training methods (e.g. the CE loss). To overcome this problem, instead of either choosing a specific/fixed number of positive samples of each class for contrastive learning \cite{41} (i.e., balancing the numerator in Eq. \ref{eq:1} or adjusting the ordinary sum to a class-weighted sum in the denominator \cite{45} in Eq. \ref{eq:1}), we modify the calculation of similarity in SCL with a zoom factor $P$. Specifically, the similarity between two embeddings $z_i$ and $z_j$ is calculated as

$$\frac{z_i \cdot z_j}{(P\tau)^{\tau'}} = \frac{z_i \cdot z_j}{(p^t/p^i)^{\tau'}}. \quad (2)$$

where $p_i$ is the prior probability of class $i$ in the local dataset and $t$ is a parameter set as 0.5 by default. Through $P$, sample pairs of the minority classes are more emphasized than those of the majority classes, leading to better balance. Denoting a new dynamic temperature factor $\tau'$ as $P\tau$, the loss function of intra-client contrastive learning is defined as

$$L_{Intra} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log \left( \frac{\exp(z_i \cdot z_j / (P\tau)^{\tau'})}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / (P\tau)^{\tau'})} \right). \quad (3)$$

D. Inter-Client Contrastive Learning

The ideal embedding space is balanced where each class occupies a space uniformly as illustrated in Fig. \ref{fig:cib} (a). However, due to class imbalance or class missing, the occupation of classes becomes non-uniform, where the majority classes occupy larger space than the minority classes as shown in Fig. \ref{fig:cib} (b) and (c) \cite{52}. In FL, the degrees of local imbalance $\Gamma$
of different clients can vary dramatically and may even be infinity due to class missing. As a result, each client may own a non-uniform and different embedding space occupation across clients as illustrated in Fig. [3] which in turn produces parameter difference and affects the convergence of FL.

To address this, the idea is to create a uniform embedding space occupation for all clients. Assuming a set of class-wise prototypes \( V = \{v^1, v^2, ..., v^L\} \), we can adopt inter-client contrastive learning similar to PCL in FedProc [16] to calibrate the embedding space trained by the embedding space occupation for all clients. Assuming a set of class-wise prototypes \( V = \{v^1, v^2, ..., v^L\} \), we can adopt inter-client contrastive learning similar to PCL in FedProc [16] to calibrate the embedding space trained by

\[
L_{Inter} = \sum_{i \in I} \frac{-1}{|P(i)|} \log \frac{\exp(z_i \cdot v^{y_i} / \tau)}{\sum_{j=1}^{L} \exp(z_i \cdot v^j / \tau)}, \tag{4}
\]

where \( y_i \) is the label of sample \( i \). When minimizing \( L_{Inter} \), the embedding of each sample will get close to the prototype of the same class while away from the prototypes of different classes. As prototypes are shared for all clients, clients tend to share a similar embedding space occupation.

To this end, how to produce high-quality prototypes is the key to inter-client contrastive learning. In previous studies, one common method to generate prototypes or global features in FL is uploading and aggregating extra information. For example, FedProc [16] and FedIRM [23] uploaded features to the server directly and CReFF [8] uploaded gradients to calculate features indirectly. However, it may cause privacy leakage under well-designed attacks and will introduce extra communication costs.

In FedIIC, we propose a new method to generate global prototypes without uploading extra information. Considering that the essence of linear classification is to calculate similarity based on vector inner product, the weights of a well-trained linear classifier are nearly co-linear with the features vectors of different classes [42], [43], [54]. Therefore, the weights of a linear classifier denoted as \( W = [w^1, w^2, ..., w^L] \), can represent the class features to some extent and be used as prototypes. Specifically, given a global model \([f_g(\cdot), g_g(\cdot), h_g(\cdot)]\) (i.e., \( f_g(\cdot), g_g(\cdot), \) and \( h_g(\cdot) \) are the feature extractor, the linear classifier, and the projection head for contrastive learning respectively) after model aggregation in the server, the weights of \( g_g(\cdot) \) are fed to \( h_g(\cdot) \) to calculate the initial prototypes \( \tilde{V} = \{\tilde{v}^1, \tilde{v}^2, ..., \tilde{v}^L\} \). Considering that the ideal embeddings of different classes in SCL should be uniformly distributed on a hypersphere [44], we fine-tune \( \tilde{V} \) via gradient descent by

\[
\tilde{V} \leftarrow \tilde{V} - \nabla \sum_{i \in I} \max_{j \in Y, j \neq y} \left( \frac{\tilde{v}^i \cdot \tilde{v}^j}{\|\tilde{v}^i\|_2 \|\tilde{v}^j\|_2} \right). \tag{5}
\]

According to Eq. [5] the cosine similarity of any \( (\tilde{v}^i, \tilde{v}^j) \) pair in \( \tilde{V} \) is minimized to be equal, resulting in a uniform embedding distribution of all classes. Finally, the class-wise prototypes \( V \) are defined as the element-wise \( l_2 \)-normalization of \( \tilde{V} \).

### E. Difficulty-Aware Logit Adjustment

In addition to the calibration of \( g(\cdot) \) through inter-client contrastive learning, we further adopt Logit Adjustment (LA) to calibrate the linear classifier \( f(\cdot) \). Before moving on to LA, we first revisit the multiclass classification problem with \( X \sim \mathbb{R}^d \) and \( Y \sim \mathbb{R}^L \). Given a sample set \( (x, y) \), for unknown probability \( p(y|x) \) over \( X \times Y \), a multiclass classification task is to learn a function \( F : X \rightarrow \mathbb{R}^L \) by minimizing the misclassification error \( p_{x,y}(y \neq \arg \max_{y' \in Y} F_{y'}(x)) \). Typically, one minimizes a surrogate loss \( L : Y \times \mathbb{R}^L \rightarrow \mathbb{R} \), such as the softmax cross-entropy loss:

\[
L_{CE}(y, F(x)) = -\log \frac{e^{F_{y}(x)}}{\sum_{y' \in Y} e^{F_{y'}(x)}} = \log[1 + \sum_{y' \neq y} e^{F_{y'}(x)-F_{y}(x)}]. \tag{6}
\]

Under data imbalance where \( p(y) \) is highly imbalanced, \( L_{CE} \) tends to focus on the majority classes as a trivial predictor which classifies all samples to the majority classes will attain a low misclassification error. Consequently, classes can have imbalanced decision boundaries. To address this, one solution is to add a per-class margin into the softmax cross-entropy, denoted as Logit Adjustment (LA) defined as

\[
L_{LA}(y, F(x)) = \log[1 + \sum_{y' \neq y} e^{\delta F_{y'}(x)-F_{y}(x)}]. \tag{7}
\]

In quantity-based approaches, \( \delta_q \) is calculated based on the prior class probabilities, e.g., \( \delta \propto p(y)^{-1/4} \) [38], \( \delta \propto \log(p(y_{\text{true}})/p(y)) \) [39], and \( \delta \propto [p(y)]^{-1/4} \) [17]. In this way, minority classes (i.e., with relatively smaller \( p(y) \)) are upweighted to learn more balanced decision boundaries.

In FedIIC, the per-class margin is calculated based on not only the prior class probabilities but also difficulties, denoted as Difficulty-Aware Logit Adjustment (DALA) inspired by [55]. For instance, in clinical scenarios, some disease types/classes may have large intra-class variations and are difficult to diagnose even with a large data amount. Therefore, the above quantity-based margin-setting approaches may not be appropriate, and the variations in classification difficulty among classes should be included for margin adjustment. In FedIIC, we utilize the loss of each class which directly reflects the difficulty in optimization for margin setting. Specifically, we set \( \delta \propto \log(p(y_{\text{true}})/p(y)/I_{ce}(y))^2 \), where \( I_{ce}(y) \) is the average CE loss of all samples belonging to class \( y \) in any round and \( q \) is a hyper-parameter set as 0.25 by default. \( I_{ce}(y) \) is calculated as follows. At any round \( r \), the total sample number of class \( y \), denoted as \( N^y_r, \) belonging to clients of communication is calculated. After receiving the global model from the server and before local training, each client \( i \) uploads \( I_{ce}^i(y) \), the total loss of class \( y, \) to the server. Finally, \( I_{ce}(y) \) is calculated as \( \frac{1}{T} \sum_{r=1}^{T} I_{ce}^i(y). \) In Section V.D, we prove that this process has no risk of privacy leakage in data distribution. Based on this form of \( \delta, \) it will assign a larger margin due to either low prior probability or high training loss for a specific class \( y, \) By replacing \( \delta \) in Eq. 7 with \( \log(p(y_{\text{true}})/p(y)/I_{ce}(y))^2 \) = \( \log(p(y)/I_{ce}(y)^2) \) - \( \log(p(y)/I_{ce}(y)^2) \), the softmax cross entropy loss
TABLE I: Data distributions of five data sources from three datasets (i.e., the imbalanced distribution in the first row of Fig. 1).

| Source       | Category | MEL | NV  | BCC | AK  | BKL | DF  | VASC | Total |
|--------------|----------|-----|-----|-----|-----|-----|-----|------|-------|
| PH2          |          | 40  | 160 | 0   | 0   | 0   | 0   | 0    | 200  |
| ViDIR (Legacy) | Atlas   | 67  | 350 | 5   | 0   | 10  | 4   | 3    | 439  |
|              |          | 268 | 575 | 42  | 0   | 69  | 20  | 29   | 1003 |
|              | Rosendahl| 342 | 803 | 296 | 295 | 490 | 30  | 3    | 2259 |
|              | ViDIR (Current) | 680 | 1832| 211 | 32  | 475 | 51  | 82   | 3363 |
| test         |          | 500 | 500 | 500 | 500 | 154 | 165 | 2819 |

with difficulty-aware logit adjustment $L_{DALA}$ is written as

$$L_{DALA}(y, F(x)) = \log(1 + \sum_{y' \neq y} e^{\log(\frac{p(y')}{|c_e(y')|^q}) - \log(\frac{p(y)}{|c_e(y)|^q}) e^{F(y')(x) - F_y(x)})$$

$$= \log(1 + \sum_{y' \neq y} e^{[F(y')(x)+\log(\frac{p(y')}{|c_e(y')|^q})] - [F_y(x)+\log(\frac{p(y)}{|c_e(y)|^q})]})$$

and DALA is written as

$$F_y(x) \leftarrow F_y(x) + \log(\frac{p(y)}{|c_e(y)|^q}) \tag{9}.$$  

Note that the calculation of $L_{DALA}$ does not rely on the multi-viewed batch like $L_{Intra}$ and $L_{Inter}$. For a fair comparison with other methods based on CE loss, only the first view of the multi-viewed batch is used to calculate $L_{DALA}$.

Finally, the overall loss function is written as

$$\mathcal{L} = L_{DALA} + k_1 L_{Intra} + k_2 L_{Inter} \tag{10}.$$  

where $k_1$ and $k_2$ are trade-off hyper-parameters. After minimizing $\mathcal{L}$ during the local training phase of each client, the global model is updated by FedAvg [5].

IV. EVALUATION

A. Dataset

1) Real Multi-Source Dermoscopic Image Datasets (denoted as Real): Five data sources from three datasets, including PH2 [56], Atlas [57], and HAM10000 [6], are adopted for realistic evaluation where each source is considered as an individual client in FL. Details of different sources are summarized in Table I. For evaluation, we construct a separate test set by selecting data from the training set of ISIC 2019 and ensure that the test set has no overlap with the above five data sources. Note that the extra class named SCC in ISIC 2019 is considered a sub-class of AK when constructing the test set. Among the seven classes, DF and VASC are defined as minor and others are major.

2) Intracranial Hemorrhage Classification (denoted as ICH): The RNSA ICH dataset [58], containing five ICH subtypes, is adopted for realistic evaluation. Data is pre-processed in the same way as [23], [48], and in total 67969 images with only one single hemorrhage type are selected. Among the five classes, Epidural, with the least data amount (i.e., 2.20%) is defined as minor and others are major. Following [23], [48], data is split according to 7:1:2 for training, validation, and testing respectively.

3) Skin Lesion Classification (denoted as ISIC): The training data of ISIC 2019 [6], [59], containing eight classes of 25331 images, for skin lesion classification is used for realistic evaluation. Among the eight classes, Dermatofibroma (DF) and Vascular Lesion (VASC), accounting for about 0.94% and 0.99% of the dataset, are defined as the minority classes and others are the majority classes. Following [23], [48], we split the dataset by 7:1:2 for training, validation, and testing respectively.

B. Evaluation Metrics

As the test sets are imbalanced, balanced accuracy (BACC), i.e. the official metric of the ISIC 2019 Challenge, is adopted for evaluation, defined as

$$BACC = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}, \tag{11}$$

where $TP_c$ and $FN_c$ represent true positives and false negatives of class $c$ respectively. BACC treats all classes equally instead of being biased toward the majority classes. For the evaluation of Real, the average and the standard deviation of BACC of the last five rounds are reported for comparison.

C. Data Partition

Following [7], [8], [16], [21], we use a Dirichlet distribution $Dir(\alpha)$ to divide the training data for each client, where the only hyper-parameter is $\alpha$. With a larger $\alpha$, the sampled probability distribution $p \sim Dir(\alpha)$ is more likely to be even. As analyzed in Section III in $L^2$ distribution, the degree of data concentration of the minority classes is greater than that of the majority classes, i.e. $Conc^{\text{minor}} \gg Conc^{\text{major}}$. To simulate this, we use two Dirichlet distributions with a large $\alpha_1$ and a small $\alpha_2$ for data partitioning of the minority and the majority classes respectively. To simulate data agglomeration in $L^2$ distribution, for each class, the sampled probabilities produced by $Dir(\alpha)$ are sorted first in descending order before assigning samples to the corresponding clients. In this way, samples of classes are relatively more concentrated to some clients as illustrated in Fig. 1. Specifically, we set $\alpha_1 = 50,$
\( \alpha_2 = 0.1, \) and \( K = 20 \) for ICH and \( \alpha_1 = 50, \alpha_2 = 0.5, \) and \( K = 10 \) for ISIC, resulting in the generated data distributions as shown in Fig. 4.

D. Experimental Details

EfficientNet-B0 [61], pre-trained by ImageNet [20], is adopted as the backbone for classification trained by an Adam optimizer with betas as 0.9 and 0.999, weight decay as 5e-4, a constant learning rate as 1e-4 for Real and 3e-4 for both ICH and ISIC, and a batch size of 32. For ICH, the multiviewed batch for contrastive learning is generated by following [23], [48]. For both Real and ISIC, the multiviewed batch is generated by 1) RandAug [62] and 2) SimAugment [51]. The hyper-parameters \( k_1 \) and \( k_2 \) in Eq. 10 are set as 2.0 for ICH and ISIC and are set as 1.2 and 1.0 respectively for Real. For federated training, the local training epoch is set as 1 and the complete framework is trained for 200 rounds for ICH and ISIC and 30 rounds for Real. At each training round, we evaluate the global model on the validation set and save the best model for testing for ICH and ISIC. At each round, all clients (i.e., 100%) are included for model aggregation.

E. Quantitative Comparison

Nine approaches are included for comprehensive comparison and evaluation, including FedAvg [5], FedProx [24] addressing data heterogeneity, MOON [7] and FedProc [16] utilizing contrastive learning in FL, [16], FedFocal [60] utilizing focal loss for balancing, FedRS [18] addressing the class-balancing problem, FedLC [17] addressing the non-IID issue by logits calibration, and CLIMB [49] and CREff [8] addressing the global imbalance in FL.

Quantitative results under the Real, ICH, and ISIC settings are summarized in Table I. Among comparison approaches, FedIIC achieves the best overall performance. Furthermore, separate evaluation on the classification of the minority and the majority classes is conducted to analyze how approaches perform with varying data amounts of different classes. Under both the Reals and ICH settings, FedIIC performs the best on both the minority and the majority classes, while achieving the best and the second-best results on the majority and the minority of ISIC respectively. Though FedProx achieves the second-best results on the majority classes of both ICH and ISIC, it is at the cost of the training process being dominated by the majority classes, resulting in poor performance for the minority classes. Comparatively, CREff achieves the best overall performance among comparison approaches with more balanced results on both the minority and the majority classes.

F. Ablation Study

To validate the effectiveness of each component in FedIIC, a series of ablation studies are conducted on Real, ICH, and ISIC following the same experimental details described in Section IV. Quantitative results are summarized in Table III. Under severe global imbalance, FedAvg is struggling. With the introduction of DALA, the performance of FedAvg is effectively improved (i.e., an average increase of 9.45%, 9.63%, and 3.09% for Real, ICH, and ISIC respectively). It is consistent with the analysis in [8], [37] that the bias caused by class imbalance mainly exists in classifiers, and calibrating the classifiers is the most efficient way for balanced learning. Based on this, introducing either intra- or inter-client contrastive learning is helpful, indicating their benefits for better representation learning under data imbalance. It should be noted that separately introducing intra- and inter-client contrastive learning to Real is less effective compared to both ICH and ISIC, mainly due to highly limited training data in total. Comparatively, jointly adopting both intra- and inter-client contrastive learning is beneficial across all settings. By combining all the components, FedIIC achieves the best overall performance, outperforming FedAvg with large margins.

V. DISCUSSION

A. Extension to Natural Image Classification

1) Datasets and Evaluation Metric: Following previous studies, two widely-used datasets CIFAR-10 and CIFAR-100 for image recognition [19] are adopted for evaluation. The former contains 10 classes while the latter contains 100 classes, and both of them contain 50000 images for training and 10000 images for testing. Considering these two datasets are balanced, two new imbalanced datasets are sub-sampled from them to simulate global imbalance. Specifically, inspired by [38], we use an exponential decay function to determine the data amounts of different classes. To simulate the \( L^2 \) distribution as illustrated in Fig. 1, especially the gap between the majority classes (i.e., the first seven classes in CIFAR-10, and the first 70 classes in CIFAR-100) and the minority classes (i.e., the last three classes in CIFAR-10, and the last 30 classes in CIFAR-100) in the global distribution, data of classes is not evenly sampled in the decay function like [38]. The exact data amount of each class is determined by

\[
\text{num}(c) = \begin{cases} 
\text{num}_{\max} \times \Gamma \frac{L - L^c}{L - L^m} & \text{c is major} \\
\text{num}_{\max} \times \Gamma \frac{L^c - L^m}{L^c - L^m} & \text{c is minor}
\end{cases}
\]

(12)

where \( L \) is the total number of classes, \( \text{num}_{\max} \) is the max size of all classes (i.e., 5000 in CIFAR-10 and 500 in CIFAR-100) and \( \Gamma \) is the imbalance factor defined in Section I. For both datasets, Top-1 accuracy is used for evaluation due to the available balanced test set.

2) Experimental Details: Following [8], \( \Gamma \) is set as 10, 50, and 100 separately to create imbalanced datasets. As for data partitioning, we set \( \alpha_1 = 10, \alpha_2 = 0.2, \) and \( K = 100 \) for CIFAR-10 and \( \alpha_1 = 10, \alpha_2 = 0.5, \) and \( K = 50 \) for CIFAR-100. The generated data distributions of CIFAR-10 and CIFAR-100 are illustrated in Fig. 1(a)-(f). For both datasets, ResNet-8 [63] is adopted as the backbone for classification and trained by the SGD optimizer with momentum as 0.9, weight decay as 5e-4, a constant learning rate as 0.03, and a batch size of 64. In terms of both intra- and inter-client contrastive learning, the multiviewed batch is generated by 1) random crop and flip and 2) SimAugment [51]. The hyper-parameters \( k_1 \) and \( k_2 \) in Eq. 10 are set as 1.0 and 0.6 respectively. For federated training, the local training epoch is set as 1 and the complete framework is trained for 1500 rounds. At each round,
TABLE II: Quantitative results of different methods conducted under the Real, ICH, and ISIC settings. For Real, following [18], the average and the standard deviation of BACC (%) of the last 5 rounds are reported. For ICH and ISIC, BACC (%) of the test set is reported. In addition, BACC (%) of the minority classes, the majority classes, and all classes on the test sets are reported. Bold and underlined numbers represent the best and the second-best results respectively.

| Method   | Real Minority | Majority | Avg.   | ICH Minority | Majority | Avg.   | ISIC Minority | Majority | Avg.   |
|----------|---------------|----------|--------|--------------|----------|--------|---------------|----------|--------|
| FedAvg [5] | 47.99 ± 2.84 | 44.10 ± 0.72 | 45.21 ± 0.99 | 15.01 | 86.53 | 72.22 | 76.13 | 76.77 | 76.61 |
| FedProx [24] | 48.24 ± 2.86 | 44.15 ± 0.79 | 45.32 ± 1.08 | 35.78 | 87.66 | 77.29 | 78.63 | 79.25 | 79.10 |
| MOON [7] | 44.07 ± 2.16 | 42.90 ± 1.37 | 43.23 ± 1.43 | 6.07 | 86.47 | 70.39 | 73.18 | 74.58 | 74.23 |
| FedProc [16] | 25.33 ± 1.78 | 39.26 ± 0.89 | 35.28 ± 0.86 | 4.47 | 84.31 | 68.34 | 72.27 | 75.94 | 75.02 |
| FedFocal [60] | 48.44 ± 1.23 | 42.23 ± 1.08 | 44.00 ± 1.00 | 5.43 | 85.53 | 69.51 | 80.22 | 64.06 | 68.10 |
| FedRS [18] | 48.57 ± 0.69 | 43.90 ± 0.75 | 45.23 ± 0.93 | 11.18 | 86.65 | 71.56 | 75.00 | 76.91 | 76.43 |
| FedIC [17] | 51.51 ± 2.68 | 48.81 ± 0.60 | 46.73 ± 0.83 | 37.69 | 86.37 | 76.63 | 78.86 | 77.05 | 77.50 |
| CLIMB [49] | 47.09 ± 2.04 | 45.66 ± 0.87 | 46.07 ± 0.94 | 0.00 | 83.21 | 66.56 | 77.04 | 73.63 | 74.48 |
| CRReFF [8] | 68.53 ± 0.50 | 46.97 ± 0.26 | 53.13 ± 0.18 | 70.92 | 85.60 | 82.66 | 90.90 | 76.61 | 80.19 |
| FedIIC (Ours) | 71.41 ± 0.11 | 48.71 ± 0.60 | 55.20 ± 0.24 | 71.56 | 87.97 | 84.69 | 88.63 | 81.56 | 83.33 |

20% of clients are randomly selected for model aggregation and then performance evaluation.

3) Quantitative Comparison: Quantitative comparison results are summarized in TABLE IV. Across all settings, FedIIC achieves the best performance with a large margin compared to the second-best results. FedProx and MOON, though designed for addressing the non-IID problem, can hardly outperform FedAvg under global imbalance. The failure of FedProc demonstrates the necessity of designing appropriate prototypes in addition to contrastive learning. Similarly, FedFocal performing worse than FedAvg shows the limitation of focal loss on imbalanced learning. Given the class-missing problem in L2 distribution, FedRS achieves noticeable performance improvement compared to FedAvg by limiting the update of missing classes’ weights.

One interesting observation is that CLIMB fails to outperform FedAvg on both datasets. In L2 distribution, data imbalance among clients becomes more severe than both U and L distributions, making CLIMB struggle to find optimal clients’ weights for balanced learning. Among all comparison approaches, CRReFF achieves the best results on CIFAR-10 but encounters significant performance degradation on CIFAR-100. It is because that samples belonging to the minority classes can be extremely limited in L2 distribution, generating poor federated features and further misleading the retraining of the classifier in CRReFF.

TABLE III: Ablation study of components in FedIIC. Bold numbers indicate the best results.

| Method   | Real Minority | Majority | Avg. | ICH Minority | Majority | Avg. | ISIC Minority | Majority | Avg. |
|----------|---------------|----------|------|--------------|----------|------|---------------|----------|------|
| FedAvg   | ✓             | ✓        | Inter | 45.21 ± 0.99 | 72.22    | 76.61 | 15.01         | 86.53    | 72.22 |
| DALA     | ✓             | ✓        | Inter | 54.66 ± 0.30 | 81.85    | 79.70 | 45.76         | 77.04    | 68.40 |
| Intra    | ✓             | ✓        | Inter | 54.57 ± 0.28 | 84.49    | 82.40 | 54.62         | 84.10    | 82.10 |
| Inter    | ✓             | ✓        | Inter | 54.20 ± 0.24 | 84.69    | 83.33 | 55.20         | 84.69    | 83.33 |

B. Evaluation on L Distributions

In addition to the proposed L2 distribution, we further extend FedIIC to a more widely-studied L distribution and conduct comparison experiments by following the non-IID setting in [8] on both the CIFAR-10-LT (IF=10) [38] and ICH. Specifically, for CIFAR-10-LT, we set α = 0.5, K = 20, local_epoch = 10, and online_clients = 8 by following [8] and follow other settings described in Section IV.A. For ICH, we set α = 1.0 and follow other settings described in Section IV.C. CLIMB [49] and CRReFF [8] designed for FL with data imbalance are included for comparison.
TABLE IV: Quantitative results of different methods conducted on CIFAR-10 and CIFAR-100. Following [18], the average and the standard deviation of Top-1 accuracy (%) of the last 50 rounds are reported for comparison. Bold and underlined numbers represent the best and the second-best results respectively.

| Method   | CIFAR-10   | CIFAR-100  |
|----------|------------|------------|
|          | Γ=10       | Γ=50       | Γ=100      |
| FedAvg [5] | 79.92±1.80 | 68.34±1.54 | 62.94±1.26 |
| FedProx [24] | 80.23±1.69 | 68.42±1.45 | 63.27±1.40 |
| MOON [7]   | 80.00±1.61 | 69.41±2.33 | 64.59±1.98 |
| FedProc [16] | 71.39±3.13 | 57.99±1.19 | 57.21±1.27 |
| FedFocal [60] | 75.72±1.74 | 63.92±1.36 | 59.70±1.20 |
| FedRS [18]  | 81.07±1.44 | 70.35±1.38 | 65.32±1.46 |
| FedLC [17]  | 81.67±1.48 | 71.75±1.40 | 67.01±1.71 |
| CLIMB [49]  | 79.11±1.82 | 68.57±1.65 | 63.42±1.48 |
| CReFF [8]  | 82.59±1.01 | 75.94±1.58 | 73.01±1.49 |
| FedIIC (Ours) | 77.26±1.35 | 79.44±1.61 | 78.02±1.46 |

**TABLE V:** Quantitative results of different approaches on dealing with data imbalance under L distributions. For CIFAR-10-LT, the average and the standard deviation of Top-1 accuracy (%) of the last 10 rounds are reported. For ICH, BACC (%) of the test set is reported. Bold numbers indicate the best results.

| Method   | CIFAR-10-LT (IF=10) | ICH |
|----------|---------------------|-----|
|          | Rounds   | 40 | 200 | 400 | 200 |
| FedAvg [5] | 56.53±6.03 | 60.15±6.61 | 72.71 | 73.74 |
| CLIMB [49] | 59.01±4.33 | 64.86±5.96 | 71.73 | 71.73 |
| CReFF [8]  | 71.24±3.41 | 76.89±4.55 | 75.00 | 82.21 |
| FedIIC (Ours) | 77.26±1.35 | 79.44±1.61 | 78.02 | 84.23 |

**TABLE VI:** Ablation study of the hyper-parameter q in DALA on ISIC. BACC (%) of the test set is reported. Bold numbers indicate the best results.

| Rounds   | q  | 0  | 0.25 | 0.5 | 1  | 2  |
|----------|----|----|------|-----|----|----|
| 100      | 79.41 | 81.33 | 80.71 | 81.22 | 80.08 |
| 200      | 82.09 | 83.33 | 82.52 | 81.72 | 81.69 |

Quantitative results are summarized in Table. Among the three baseline approaches, CReFF achieves the best overall performance, which is consistent with the results in Tables and One interesting observation is that CLIMB outperforms FedAvg on CIFAR-10-LT under the L distribution. However, it still fails to surpass FedAvg under the ICH setting, indicating the limitation of client-wise re-weighting in CLIMB. Compared to the baseline approaches, FedIIC achieves the best overall performance across the two settings with a large margin, validating the extend-ability of FedIIC on handling other imbalanced data distributions in addition to L² distributions.

**C. Hyper-Parameter Discussion in DALA**

In FedIIC, one hyper-parameter is q in difficulty-aware logit adjustment (DALA), which determines the importance of class difficulties in logit adjustment. If q is large, logit adjustment is mainly dominated by class difficulties. Otherwise, DALA is dominated by the prior class probability when q approaches zero. To quantify the value of q in FedIIC, we conduct a series of ablation studies on the ISIC dataset with varying q as summarized in Table. Compared to the baseline logit adjustment ignoring class difficulties, increasing q is beneficial. But it does not mean that setting a larger q would necessarily lead to better results as stated in Table. With the increase of q, there exists noticeable performance degradation. It is because emphasizing too much on difficult classes may result in instabilities in training. According to the quantitative results in Table, setting q as 0.25 achieves the best performance on the ISIC dataset. For different tasks/datasets, q should be tuned accordingly based on data distributions and class difficulties.

**D. Communication Efficiency**

Communication efficiency is one of the most significant metrics to evaluate an FL algorithm as it measures how fast an algorithm converges. Following [7], the communication efficiency of each algorithm is defined as the number of training rounds when it achieves the same performance as FedAvg. Specifically, for ICH and ISIC, as the best model of FedAvg is selected based on the validation sets, the performance of each algorithm at round r is evaluated by the BACC score on the corresponding validation sets.

Quantitative results of communication efficiency are summarized in Table. Across all settings, FedIIC achieves the same performance as FedAvg with the fewest training rounds, proving that FedIIC is more communication-efficient than...
other FL frameworks. One inspiring and promising observation is that FedIIC reaches the performance of FedAvg with only 20 and 78 training rounds on the two real medical datasets, accelerating the convergence of FedAvg by 8.75x and 2.50x respectively. The above results further demonstrate the value of FedIIC in real FL applications.

E. Privacy

Compared to FedAvg, FedIIC requires each client to upload its data size $N^*$ and the CE loss $\tilde{L}_{CE}(y)$ at each round $r$. In this subsection, we will prove that FedIIC is privacy-preserving under the existing secure multi-party computation framework. This problem equals to “How can we compute the sum of a set of numbers without revealing any of them?” which can be solved by homomorphic encryption (HE) [49]. HE can be simplified as

$$D(E(a) \cdot E(b)) = a + b,$$  

where $a$ and $b$ are two numbers and $D$ and $E$ represent the decryption and encryption operations respectively. In FedIIC, each client transmits its encrypted number (e.g. $E(a)$) to the server and the server multiplies clients’ ciphertexts without decryption due to the lack of keys and sends the results (e.g. $E(a) \cdot E(b)$) to clients. Then, each client decrypts the downloaded results (e.g. $D(E(a) \cdot E(b)) = a + b$). In this way, either the server or the clients in FedIIC would have no access to clients’ privacy.

VI. CONCLUSION

In this paper, we, for the first time, identify and formulate a special and realistic imbalanced class distribution in federated learning, denoted as $L^2$ distribution, and propose a novel framework named FedIIC accordingly. The key idea behind FedIIC is to calibrate both the feature extractor and the classification head against divergence caused by learning from imbalanced data. Specifically, we calibrate the feature extractor by both intra- and inter-client contrastive learning and thus produce a uniform embedding space for all classes across clients. To calibrate the classification head, a difficulty-aware logit adjustment algorithm is deployed to produce balanced decision boundaries for all classes. Extensive experimental results demonstrate the effectiveness of FedIIC against the state-of-the-art federated imbalanced learning approaches. Furthermore, FedIIC is proven to be effective in dealing with data imbalance under other widely-studied distributions. We believe the newly-identified imbalanced distribution, together with FedIIC, would inspire future federated imbalanced learning work in realistic scenarios.

REFERENCES

[1] S. Minaei et al., “Image segmentation using deep learning: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 7, pp. 3523-3542, 2022.
[2] D. W. Otter, J. R. Medina, and J. K. Kalita, “A survey of the usages of deep learning for natural language processing,” IEEE Trans. Neural Networks Learn. Syst., vol. 32, no. 2, pp. 604-624, 2021.
[3] C. Zhang et al., “A survey on federated learning,” Knowl. Based Syst., vol. 216, p. 106775, 2021.
[4] W. Y. B. Lim et al., “Federated learning in mobile edge networks: A comprehensive survey,” IEEE Commun. Surv. Tutor., vol. 22, no. 3, pp. 2031-2063, 2020.
[5] B. McManah, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” 2016, arXiv:1602.05629.
[6] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” Sci. Data, vol. 5, no. 1, pp. 1-9, 2018.
[7] Q. Li, B. He, and D. Song, “Model-contrastive federated learning,” in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit., 2021, pp. 10713-10722.
[8] X. Shang, Y. Lu, G. Huang, and H. Wang, “Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features,” 2022, arXiv:2204.13399.
[9] P. Kairouz et al., “Advances and open problems in federated learning,” 2019, arXiv:1912.04977.
[10] Y. Kang et al., “Privacy-preserving federated adversarial domain adaptation over feature groups for interpretability,” IEEE Trans. Big Data, pp. 1-12, 2022.
[11] C. -M. Feng et al., “Specificity-preserving federated learning for MR image reconstruction,” IEEE Trans. on Med. Imaging, 2022.
[12] W. Huang, M. Ye, and B. Du, “Learn from others and be yourself in heterogeneous federated learning,” in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit, 2022, pp. 1033-1042.
[13] P. Guo, P. Wang, J. Zhou, S. Jiang, and V. M. Patel, “Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning,” in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit, 2021, pp. 2423-2432.
[14] B. Sun, H. Huo, Y. Yang, and B. Bai, “PartialFed: Cross-domain personalized federated learning via partial initialization,” in Proc. Adv. Neural Inf. Process. Syst., 2021, pp. 23309-23320.
[15] X. Li et al., “Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results,” Med. Imag. Anal., vol. 65, p. 101765, 2020.
[16] X. Mu et al., “FedProc: Prototypical contrastive federated learning on non-iid data,” 2021, arXiv:2109.12273.
[17] J. Zhang et al., “Federated learning with label distribution skew via logits calibration,” in Proc. Int. Conf. Mach. Learn., 2022, pp. 26311-26329.
[18] X. -C. Li and D. -C. Zhan, “FedRS: Federated learning with restricted softmax for label distribution non-iid data,” in Proc. ACM SigKDD Conf. Knowl. Discov. Data Mining, 2021, pp. 995-1005.
[19] A. Krizhevsky et al., “Learning multiple layers of features from tiny images,” Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.
[20] J. Deng et al., “Imagenet: A large-scale hierarchical image database,” in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit., 2009, pp. 248-255.
[21] M. Yurochkin, M. Agrawal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni, “Bayesian nonparametric federated learning of neural networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 7252-7261.
[22] P. Khosla et al., “Supervised contrastive learning,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 18661-18673.
[23] Q. Liu, H. Yang, Q. Dou, and P. -A. Heng, “Federated semi-supervised medical image classification via inter-client relation matching,” in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv., 2021, pp. 325-335.
[24] T. Li, K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” 2018, arXiv:1812.06127.
[25] S. P. Karimireddy et al., “SCAFFOLD: Stochastic controlled averaging for on-device federated learning,” 2019, arXiv:1910.06578.
[26] X. Peng, Z. Huang, Y. Zha, and K. Saenko, “Federated adversarial domain adaptation,” 2019, arXiv:1911.02054.
[27] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, “FedBN: Federated learning on non-iid features via local batch normalization,” 2021, arXiv:2102.07623.
[28] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, and K. -T. Cheng, “Variational-aware federated learning with multi-source decentralized medical image data,” IEEE J. Biomed. Health Inform., vol. 25, no. 7, pp. 2615-2628, 2020.
[29] Q. Liu, C. Chen, J. Qin, Q. Dou, and P. -A. Heng, “FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space,” in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit., 2021, pp. 1013-1023.
[30] M. Jiang, Z. Wang, and Q. Dou, “HarmoFL: Harmonizing local and global drifts in federated learning on heterogeneous medical images,” in Proc. AAAI Conf. Artif. Intell., 2022, pp. 1087-1095.
