FedRare: Federated Learning with Intra- and Inter-Client Contrast for Effective Rare Disease Classification

Nannan Wu, Li Yu, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan

Abstract—Federated learning (FL), enabling different medical institutions or clients to train a model collaboratively without data privacy leakage, has drawn great attention in medical imaging communities recently. Though inter-client data heterogeneity has been thoroughly studied, the class imbalance problem due to the existence of rare diseases still is under-explored. In this paper, we propose a novel FL framework FedRare for medical image classification especially on dealing with data heterogeneity with the existence of rare diseases. In FedRare, each client trains a model locally to extract highly-separable latent features for classification via intra-client supervised contrastive learning. Considering the limited data on rare diseases, we build positive sample queues for augmentation (i.e. data re-sampling). The server in FedRare would collect the latent features from clients and automatically select the most reliable latent features as guidance sent back to clients. Then, each client is jointly trained by an inter-client contrastive loss to align its latent features to the federated latent features of full classes. In this way, the parameter/feature variances across clients are effectively minimized, leading to better convergence and performance improvements. Experimental results on the publicly-available dataset for skin lesion diagnosis demonstrate FedRare’s superior performance. Under the 10-client federated setting where four clients have no rare disease samples, FedRare achieves an average increase of 9.60% and 5.90% in balanced accuracy compared to the baseline framework FedAvg and the state-of-the-art approach FedIRM respectively. Considering the board existence of rare diseases in clinical scenarios, we believe FedRare would benefit future FL framework design for medical image classification. The source code of this paper is publicly available at https://github.com/wnn2000/FedRare.

Index Terms—Federated learning, contrastive learning, skin lesion, rare disease diagnosis

I. INTRODUCTION

Deep learning is a data-driven technology, requiring a collection of multi-source data to extract domain-invariant features for better generalizability. However, due to growing privacy concerns, especially for medical imaging data, fusing multi-source data for training is infeasible, making deep learning less feasible for large-scale deployment in clinical scenarios.

Recently, federated learning (FL) [1]–[7], allowing different parties to train a deep learning model collaboratively without sharing their private data, has received explosive popularity. A classical algorithm of FL is FedAvg [1] whose process can be briefly described as follows. Firstly, each client trains a model, with a shared architecture among clients, based on its own data. Then, the parameters of the locally-updated model are sent to a central server. After collecting the model parameters from clients, the federated model is updated through parameter aggregation, and its parameters are then sent back to each client to start a new training round until convergence. As only the parameters, instead of raw data, of clients are utilized for joint training, data privacy is well preserved.

Despite the success of FL in medical imaging [20], [21], [31]–[38], it still suffers from data heterogeneity, i.e. non-independent and identically distributed (Non-IID) data [39]–[42] among sources/clients which can be further divided into the feature-distribution skew and the label-distribution skew [8]. Data heterogeneity on the feature-distribution skew, usually referring to image variations, mainly is due to different data acquisition protocols. To address this in FL, Peng et al. [30] adopted adversarial domain adaptation for feature-level alignment to alleviate variations among EEG signals. Yan et al. [10] proposed to minimize inter-client variations by image-level alignment through Generative Adversarial Networks (GAN) [22]–[24]. Liu et al. [11] proposed to reduce the discrepancy of clients’ images via frequency space interpolation. Li et al. [17] minimized the feature shift among clients by keeping batch normalization locally. Jiang et al. [12] proposed a revised softmax function to restrict the updates of specific weights in local training.

For the label-distribution skew where clients may only own a subset of the class set [9], FedProx [13] penalized the $\ell_2$ distance between global and local models’ parameters to avoid huge bias between models. Karimireddy et al. [14] aimed at correcting the model’s drift during local training with variance reduction. Contrastive learning [25]–[29] has also been used against this problem. Li et al. [15] used contrastive learning to maintain parameter consistency among clients’ locally-updated models. Mu et al. [16] proposed a prototypical contrastive loss to guide the training of local models. Considering the limitation of softmax, Li et al. [9] proposed a revised softmax function to restrict the updates of specific weights in local training.

All the above approaches target typical data heterogeneity on label distributions as illustrated in Fig. 1. Though the class distributions within each client vary significantly, the total available data amount of each class is similar. In FL, each client is likely to learn robust features for one or more classes.
without interactions among clients and without addressing the heterogeneous label distribution problem. However, in clinical scenarios, not all classes (e.g., diseases) are equally important and have similar occurrence frequencies. On the one hand, there can exist rare diseases with fewer samples compared to others. For instance, in the widely-used skin lesion dataset, HAM10000 [33], only around 2.57% of the images in the officially-provided training set belong to dermatofibroma (DF) and vascular (VASC) lesions (2 out of 7 classes). On the other hand, medical institutions can vary significantly in expertise and scale, resulting in heterogeneous label distributions where some clients have no rare disease samples (i.e., missing classes) as shown in Fig. 1. As is proven that heterogeneous data could cause drifts of models’ parameters [12], [13], [17], the existence of rare diseases would make it more challenging for FL in clinical scenarios, which has not been explored.

In this paper, we propose a new FL framework FedRare for medical image classification which well takes in account data heterogeneity with rare diseases. The key idea is to reduce parameter/latent feature drifts among clients with and without missing classes (i.e., rare diseases). For all clients, we first apply supervised contrastive learning [13], [19] to learn compact representations of each class. Considering the limited amount of rare diseases (i.e., class imbalance), we construct a positive sample queue to increase their importance via re-sampling. Next, we align the latent feature distributions of each class across clients by inter-client contrastive learning. Through this, the latent feature distributions of the classes across clients are regularized to be consistent regardless of the existence of rare diseases, leading to better convergence of FL. Experimental results on skin lesion classification demonstrate the superior performance of FedRare on learning with rare diseases against the state-of-the-art FL approaches. The contributions of this paper can be summarized as follows.

1) The first FL framework aiming at medical image classification with specific focus on data heterogeneity with rare diseases. Compared to typical data heterogeneity, heterogeneous label distributions with rare diseases is more challenging, while common, in clinical applications.

2) Intra-client contrast combined with a positive sample queue for better representation learning in the presence of class imbalance (i.e., rare diseases). It is validated to be effective for increasing data diversity in federated learning.

3) Inter-client contrast for latent feature alignment among clients with and without missing classes (i.e., rare diseases).

The paper is organized as follows. The challenge of medical image classification with missing classes in FL is analyzed in Section II. Section III presents details of the proposed FedRare framework. In Section IV we evaluate the effectiveness of the proposed framework through extensive comparison experiments. Section V provides a discussion and Section VI concludes the paper.

II. PROBLEM ANALYSIS

Given a disease classification task, FL aims to minimize the following objective function:

$$\min_{\theta_g} \left[ F(\theta_g) := \sum_{(x, y) \in D_F} \ell(x, y; \theta_g) \right],$$

where $\theta_g$ represents the federated (i.e., global) model’s parameters, $\ell$ is a generic loss function to measure the distance between ground-truth labels and the model’s outputs (e.g., cross-entropy loss), and $D_P$ is a dataset consisting of samples (denoted as $(x, y)$) which is expected to reflect the real distribution of diseases in the population. As data is distributed across clients without sharing in FL, the federated model is optimized indirectly through the optimization of each client’s local model according to

$$\theta_{g}^{t+1} = \theta_{g}^{t} - \eta \frac{\partial F_{g}^{t}}{\partial \theta_{g}^{t}},$$

where $K$ is the number of clients, $w_{i}$ is the aggregation weight of the $i$-th client satisfying $\sum_{i=1}^{K} w_{i} = 1$ and $\theta_{i}$ represents its model parameters at time (i.e., training round) $t + 1$ updated by

$$\theta_{i}^{t+1} = \theta_{i}^{t} - \eta \frac{\partial F_{i}^{t}}{\partial \theta_{i}^{t}},$$

where $\eta$ is the learning rate and $F_{i}^{t} = \sum_{(x, y) \in D_{i}} \ell(x, y; \theta_{i}^{t})$ is the objective function of client $i$ trained on its local dataset $D_{i}$. With homogeneous data, FL optimization is written as

$$\forall i, j \in [1, K], F_{i}^{t} \approx F_{j}^{t} \Leftrightarrow \theta_{i}^{t+1} \approx \theta_{j}^{t+1} \approx \theta_{g}^{t+1} \approx \theta_{i}^{t+1}.$$
Fig. 2: Exemplar latent feature distributions under different cases. The latent features are illustrated through normalized 2D vectors inspired by [9]. From left to right: the latent feature distribution under homogeneous data $L_{\text{homo}}$, the latent feature distributions under heterogeneous data with missing classes (i.e. three out of five classes) $L_{\text{homo}}^{\ast}$ and rare diseases (i.e. full classes with highly-imbalanced data amount) $L_{\text{rare}}$ respectively, the ideally highly-separable latent feature distribution with rare diseases $L_{\text{rare}}^{\ast}$, and the latent feature alignment process from $L_{\text{homo}}$ to $L_{\text{rare}}$. Here, the “pull force” is to minimize the distance between the same classes across clients while the “push force” is to maximize the distance among different classes across clients.}

approaching to centralized learning. Under data heterogeneity, the above FL optimization is rewritten as

\[ \forall i, j \in [1, K], D_i \neq D_j \Leftrightarrow F_i^{t} \neq F_j^{t} \Leftrightarrow \theta_i^{t+1} \neq \theta_j^{t+1}, \]  

and

\[ \langle F_i^{t}, F_j^{t} \rangle \uparrow \Leftrightarrow \langle \theta_i^{t+1}, \theta_j^{t+1} \rangle \uparrow \Leftrightarrow \langle \theta_g^{t+1}, \theta_g^{t+1} \rangle \uparrow, \]  

where $\langle \cdot \rangle$ denotes a distance measure and $\theta_g^{t+1}$ represents the optimal parameters of the federated model (i.e. with homogeneous data).

According to Eq. [6], the variations between $\theta_i$ and $\theta_j$ can be effectively minimized by alleviating the drifts of their local objectives. The key step is to leverage the features from the full-class clients (e.g. client $j$) to align the features of those clients with missing classes (e.g. client $i$), which can be written as:

\[ F_i = \sum_{(x,y) \in D_i} \ell(x,y; \theta_i) + \ell_{\text{LFA}}(x, f_j; \theta_i), \]  

where $f_j$ is latent features from client $j$ and $\ell_{\text{LFA}}$ denotes a loss function for latent feature alignment (LFA).

The latent feature distributions under different cases are illustrated in Fig. 2. Compared to the distribution $L_{\text{homo}}$ under homogeneous data, there exist large variances in $L_{\text{homo}}$ when some classes (i.e. rare diseases) are missing. Fortunately, the highly-separable distribution $L_{\text{homo}}$ makes it less challenging for the alignment from $L_{\text{homo}}$ to $L_{\text{homo}}$. With the existence of rare diseases, the latent feature distribution with full classes $L_{\text{rare}}$ would be further distorted compared to $L_{\text{homo}}$ as shown in Fig. 2 making the alignment from $L_{\text{homo}}$ to $L_{\text{rare}}$ more challenging. It is mainly due to the limited amount of training data corresponding to those rare diseases. Therefore, making $L_{\text{rare}}$ more separable is crucial before latent feature alignment from $L_{\text{homo}}$ to $L_{\text{rare}}$. Ideally, $L_{\text{rare}}^{\ast}$ after improvement is close to $L_{\text{homo}}$, based on which the latent feature alignment of $L_{\text{homo}}$ can be effectively performed through intra- and inter-class penalties as illustrated in Fig. 2.

### III. METHODOLOGY

The key idea behind FedRare is to align the latent feature distributions of the clients without rare diseases with those of the full-class clients under the federated setting. In the following, we describe in detail the design of FedRare as illustrated in Fig. 3.

#### A. Overview

Like classical FL frameworks, in FedRare, each client trains a local model with its private data and interacts with a credible server without leaking data privacy. Following the analysis in Section II for each client in FedRare, we first employ supervised contrastive learning to produce highly-separable latent features while minimizing classification losses. Specifically, given the limited data amount of the rare diseases, we build a queue to improve their importance via re-sampling. After local training, the class-wise latent features from the full-class clients are sent to the server for confidence-aware aggregation. Then, the federated latent features are sent back to all the clients as guidance for supervised contrastive learning. Through intra- and inter-client contrastive learning, the latent features of clients are well aligned regardless of rare diseases.

#### B. Networks Structure

The architecture of the local network structure, as illustrate in Fig. 3 is composed of three components, namely the feature extractor, the projection head, and the classification head. The feature extractor $f_e(\cdot)$ is a convolutional neural network (CNN) that encodes an input image $x \in \mathbb{R}^{H \times W}$ to a vector $v \in \mathbb{R}^P$, where $H \times W$ represents the size of $x$. The projection head $f_p(\cdot)$, a multi-layer perceptron (MLP) network followed by a $\ell_2$ normalized layer, is to project $v \in \mathbb{R}^P$ to a normalized vector space $z \in \mathbb{R}^Q$ according to [25]. The classification head $f_c(\cdot)$ is a linear classification network combined with a softmax function for prediction $p \in \mathbb{R}^K$. 

```plaintext
Exemplar latent feature distribution $L_{\text{homo}}$ (with full classes and no rare diseases) Latent feature distribution $L_{\text{homo}}^{\ast}$ (without rare diseases) Latent feature distribution $L_{\text{rare}}$ (after inter-class separation) Latent feature alignment (from $L_{\text{homo}}$ to $L_{\text{rare}}$)
```

Normalized Feature Region Ideal Feature Region Class Average Features Class Guidance Features Pull Force Push Force
C. Local Training

1) Intra-client contrastive learning: To produce highly-separable latent features for classification, we adopted supervised contrastive learning, which has been validated to be effective to extract compact latent features [18], [19], [44]. The supervised contrastive loss can be written as follows:

\[ \ell_{\text{intra}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} \mathbb{1}_{y_i = y_j} \log \frac{e^{(z_i \cdot z_j)/\tau}}{\sum_{k=1}^{N} \mathbb{1}_{i \neq k} e^{(z_i \cdot z_k)/\tau}}, \]  

(8)

where \(N\) is the number of samples in each mini-batch, \(\tau\) is the temperature parameter of contrastive learning, \(z\) is the normalized latent features generated by the projection head, \(\cdot\) is the inner-product operation, and \(\mathbb{1}\) is the indicator function.

The intra-client contrastive loss in Eq. (8) attempts to maximize the similarity of features of samples belonging to the same classes (i.e. reducing the intra-class distance) while minimizing the feature similarity among samples from different classes (i.e. increasing the inter-class distance). However, the success of supervised contrastive learning largely relies on the availability of sufficient training data, which is impractical in dealing with rare diseases. With the existence of rare diseases, there can exist only one sample per rare disease in a mini-batch, making contrastive learning useless. To address this, we construct positive sample queues to dynamically increase the number of samples for rare diseases. During training, samples of rare diseases are automatically added to the queues. Given a mini-batch containing only one sample of a certain disease, if the queue is not empty, a sample of the same disease is dequeued and concatenated to the mini-batch for training. In practice, the maximum length of each queue is set as \(L\) to save memory. In this way, the positive sample queues work as re-sampling to enrich the training samples of rare diseases.

2) Inter-client contrastive learning: Following the analysis in Section II, we aim to align the objective of each client via latent feature alignment. For this, a novel supervised contrastive loss is defined as follows:

\[ \ell_{\text{inter}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=0}^{C-1} \mathbb{1}_{y_i = j} \log \frac{e^{(z_i \cdot \hat{z}_j)/\tau}}{\sum_{k=0}^{C-1} e^{(z_i \cdot \hat{z}_k)/\tau}}., \]  

(9)

where \(C\) is the number of total classes, \(\hat{z}_j\) is the federated feature of the \(j\)-th class sent by the server for local latent feature alignment of samples belonging to class \(j\). Compared to \(\ell_{\text{intra}}\), \(\ell_{\text{inter}}\) works in a similar way but utilize the federated
Algorithm 1: FedRare

**input**: $D_i$: local dataset of client $i$

**parameter**: $R$: maximum training rounds
$E$: number of local epochs
$K$: number of clients
$K_e$: number of excellent clients

**output**: $\theta_g$: parameters of the federated model

Server executes:

initialize $\theta_g, \bar{z}, conf$

for $t = 1 : R$
do

$W = \text{softmax}(\frac{conf}{\sum_{j} dim = 0})$
$\bar{z} \leftarrow \text{sum}(W \cdot \bar{z}, dim = 0)$

for $i = 1 : K$
in parallel:do

send $\theta_g$ to Client $i$
send $\bar{z}$ to Client $i$
$\theta_i, \bar{z}, \text{conf}^i \leftarrow \text{LocalTraining}(\theta_i, D_i, \bar{z}, R)$
end

$\theta_g \leftarrow \sum_{i=1}^{K} \sum_{j=1}^{\lfloor D_i \rfloor} \theta_i$ 
to Client $i$

$\bar{z}, conf \leftarrow [\bar{z}^1, \bar{z}^2, ..., \bar{z}^K], [\text{conf}^1, \text{conf}^2, ..., \text{conf}^K]$end

return $\theta_g$

LocalTraining($\theta_i, D_i, \bar{z}, R$):

initialize $Z, P, Y \leftarrow [, [, ]$for $t = 1 : E$
do

for each batch $\{ (x_j, y_j) \}$ in $D_i$ do

dequeue()
$z_j \leftarrow f_{\theta}(f_{\theta}(x_j))$
$p_j \leftarrow f_{\theta}(f_{\theta}(x_j))$
$\ell_{cc} \leftarrow \text{CrossEntropyLoss}(p_j, y_j)$
$\ell_{intra} \leftarrow \text{IntraContrastiveLoss}(z_j, y_j)$
$\ell_{inter} \leftarrow \text{InterContrastiveLoss}(z_j, \bar{z}, y_j)$
$\ell = \ell_{cc} + \alpha \ell_{intra} + \beta(R) \ell_{inter}$
enqueue($\langle x_j, y_j \rangle$)
$\theta_i \leftarrow \theta_i - \eta \frac{\partial \ell}{\partial \theta_i}$
$Z.append(z_j), P.append(p_j), Y.append(y_j)$
end

$\bar{z} \leftarrow \text{CalAverageRepresentation}(Z, Y, P)$
$\text{conf}^i \leftarrow \text{CalConfidenceScore}(Y, P)$
return $\theta_i, \bar{z}, \text{conf}^i$


where $\alpha$ and $\beta(t)$ are trade-off parameters. $\beta(t)$ determines when to employ $\ell_{inter}$, defined as

$$\beta(t) = \begin{cases} 
0, & t \leq R_{inter} \\
\beta_{max}, & t > R_{inter}
\end{cases},$$

where $t$ is the current training round and $R_{inter}$ is the pre-defined timing to introduce $\ell_{inter}$ for training. It is because using unreliable latent features from full-class clients for guidance can be counter-productive, especially in early training rounds.

D. Server Aggregation

1) Latent feature aggregation: To better model the realistic disease distributions, in FedRare, all the clients owning rare diseases are required to upload their average latent features of each class calculated by

$$\bar{z}^j_c = \frac{\sum_{i=1}^{\lfloor D_i \rfloor} 1_{y_i = c} \cdot 1_{\text{argmax}(p_i) = c} z_j^i}{\sum_{i=1}^{\lfloor D_i \rfloor} 1_{y_i = c} \cdot 1_{\text{argmax}(p_i) = c}},$$

where $\lfloor D_i \rfloor$ represents the total number of samples in client $j$. According to Eq. (13) only those correctly-classified samples would be counted and the mean latent features of them is calculated for uploading to prevent data privacy leakage and interference from outliers.

In terms of aggregation, not all clients are equally important due to various data quality and quantity. Therefore, in addition to $\bar{z}^j_c$, client $j$ will upload a confidence score $\text{conf}_c^j$ for each class $c$ defined as

$$\text{conf}_c^j = \frac{\sum_{i=1}^{\lfloor D_i \rfloor} 1_{y_i = c} \cdot 1_{\text{argmax}(p_i) = c} d^j_{i,c}}{\sum_{i=1}^{\lfloor D_i \rfloor} 1_{y_i = c} \cdot 1_{\text{argmax}(p_i) = c}},$$

where $d^j_{i,c}$ is defined as

$$d^j_{i,c} = \frac{1}{C-1} \sum_{k=0, k \neq c}^{C-1} (p_{i,c} - p_{i,k}),$$

to reflect the average margin to other classes. $\text{conf}_c^j$ is to measure how separable the latent features of class $c$ is learned by client $j$.

After collecting the average representations and the corresponding confidence scores from all clients, the federated latent features are defined as

$$\hat{z}_c = \text{norm}(\sum_{j=1}^{K} W^j_\theta z^j_c),$$

where $\text{norm}(\cdot)$ is $\ell_2$ normalization and $W^j_\theta$ is the weight of the average representation corresponding to class $c$ of client $j$ defined as

$$W^j_\theta = \text{softmax}(\frac{\text{conf}_c^j}{T} | j) = \frac{\exp(\text{conf}_c^j/T)}{\sum_{i=1}^{C} \exp(\text{conf}_c^i/T)},$$

where $T$ is a hyperparameter that will be discussed in Section V. According to Eq. (16) only the most confident latent features of each class among the clients are selected as guidance for the above local training.

latent features as guidance. In Section III-D, we describe in detail the construction of $\hat{z}_c$ in FedRare.

3) Overall learning strategy: As for the classification task, a weighted cross-entropy loss is adopted to deal with the class-imbalance problem, defined as:

$$\ell_{ce} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=0}^{C-1} w^c_{i,c} y_{i,c} \log(p_{i,c}),$$

where $w^c_{i,c}$ is the weight of class $c$, and $y_{i,c}$ and $p_{i,c}$ are the ground-true value and the predicted value of the $i$-th sample belong to class $c$. Then, the total loss is:

$$\ell = \ell_{ce} + \alpha \ell_{intra} + \beta(t) \ell_{inter},$$

$$\ell_{inter} = \sum_{j=1}^{K} \sum_{i=1}^{\lfloor D_i \rfloor} 1_{y_i = c} \cdot 1_{\text{argmax}(p_i) = c} d^j_{i,c},$$

where $\alpha$ and $\beta(t)$ are trade-off parameters. $\beta(t)$ determines when to employ $\ell_{inter}$, defined as

$$\beta(t) = \begin{cases} 
0, & t \leq R_{inter} \\
\beta_{max}, & t > R_{inter}
\end{cases},$$

where $t$ is the current training round and $R_{inter}$ is the pre-defined timing to introduce $\ell_{inter}$ for training. It is because using unreliable latent features from full-class clients for guidance can be counter-productive, especially in early training rounds.
TABLE I: Data split under the 10-client setting.

| Client ID | Disease Categories |
|-----------|--------------------|
|           | MEL    | NV | BCC | AKIEC | BKL | DF | VASC |
| 0         | 80     | 468| 37  | 23    | 78  | 8  | 11   |
| 1         | 80     | 468| 37  | 23    | 77  | 7  | 10   |
| 2         | 80     | 468| 37  | 23    | 77  | 7  | 10   |
| 3         | 80     | 467| 37  | 23    | 77  | 0  | 0    |
| 4         | 80     | 467| 37  | 23    | 77  | 0  | 0    |
| 5         | 80     | 467| 37  | 23    | 77  | 0  | 0    |
| 6         | 79     | 467| 37  | 23    | 77  | 0  | 0    |
| 7         | 79     | 467| 37  | 23    | 77  | 0  | 0    |
| 8         | 79     | 467| 37  | 23    | 77  | 0  | 0    |
| 9         | 79     | 467| 37  | 22    | 77  | 0  | 0    |
| Total     | 796    | 4673| 370 | 229   | 771 | 29 | 41   |

2) Model aggregation: For federated model update, the most widely-used FedAvg [1] is adopted, where the weight of each client for parameter update in Eq. 2 is defined as

\[ w_i = \frac{|D_i|}{\sum_{j=1}^{K} |D_j|}. \]  

The pseudocode of FedRare is presented in Algorithm 1.

IV. EVALUATION

A. Dataset

The publicly-available dataset HAM10000 [43], for the lesion diagnosis task of the ISIC 2018 challenge [45], was used for evaluation. The dataset consists of seven diseases, of which two diseases with the fewest samples (i.e. DF and VASC) were defined as the rare diseases in our experiments.

The dataset is officially divided into three parts, namely the training set (10015 images with the corresponding labels), the validation set (193 images with the corresponding labels), and the test set (1512 images without labels). As the labels of the test set are not accessible, the predictions of all learning frameworks were uploaded to the challenge website for evaluation. Considering the small scale of the official validation set, we divided the training set into the training set (7011 images) and the validation set (3004 images) via a 70%-30% split ratio and used the re-defined validation set for a fair evaluation.

To simulate the 10-client FL setting with the existence of rare diseases, we first divided the training set into 10 partitions and evenly distributed them to 10 clients. Then, the rare disease samples of the last six clients were removed. The data distributions across clients are illustrated in Table I.

B. Evaluation Metrics

Balanced accuracy (B-ACC), the officially-used metric of the ISIC 2018 challenge, was adopted for evaluation due to the extreme class-imbalance problem, defined as

\[ B - ACC = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}, \]  

where \( TP_i \) and \( FN_i \) are the true positives and the false negatives of class \( i \) respectively. Similarly, balanced precision (B-PER) and balanced F1 score (B-F1) were adopted for evaluation, defined as

\[ B - PER = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i}, \]  

and

\[ B - F1 = \frac{1}{C} \sum_{i=1}^{C} 2 \times \frac{P_i \times R_i}{P_i + R_i}. \]  

where \( P_i \) and \( R_i \) represent the precision and the recall of class \( i \) respectively.

C. Implementation Details

1) Data Augmentation: Following [47], all training images were augmented via the modified RandAugment, followed by random crop to 224×224 pixels. For validation and testing, the multi-crop operation was used to conduct multiple forward rounds on each input image to obtain the average results.

2) Network Architecture: EfficientNet-B0 [46] was selected as the base model due to its effectiveness and lightweight. To employ intra- and inter-client contrastive learning, a multi-layer perceptron (MLP) head was inserted into EfficientNet-B0, where the dimension of the last layer was set as \( Q \) (e.g. 256 in our experiments). More discussions about the hyper-parameter \( Q \) can be found in Section V.

3) Optimization: All the networks were trained using an Adam optimizer with an initial learning rate of 5e-4 and a batch size of 64. Following the practice in [47], the learning rate was decayed every 10 rounds by 0.1 after 30 training rounds. All methods were implemented within the PyTorch framework and trained on Nvidia GeForce Titan RTX 3090 GPUs for 70 rounds. The federated training was performed synchronously and the federated model parameters were updated every training round.

D. Learning Frameworks

The following learning frameworks have been re-implemented for comparison, including:

1) Local learning (LL): Each client trains a model locally with its private data. All the models will be evaluated on the validation set and the test set independently.

2) Centralized learning (CL): All the training data from different sources/clients are collected to train a global model regardless of data privacy.

3) FedAvg [1]: The most classical FL framework. Each client trains a local model on its data and sends the model parameters to a trustworthy server. The server collects the model parameters and updates a shared federated model via parameter aggregation. The parameters of the federated model are then sent back to each client for training till convergence. It can be regarded as the baseline FL framework for comparison.

4) FedProx [13]: A FL framework for non-IID data. In the local training phase, a proximal term that measures the distance between the global model and each local model is penalized in addition to the classification loss.
### TABLE II: Quantitative results of different clients under local learning.

| Client ID | Test (%) | Validation (%) | Classes 0-6 | Classes 0-4 | Classes 5-6 |
|-----------|----------|----------------|-------------|-------------|-------------|
|           | B-ACC    | B-ACC B-PER B-F1 | B-ACC B-PER B-F1 | B-ACC B-PER B-F1 | B-ACC B-PER B-F1 |
| 0         | 61.2     | 62.1 51.7 54.8 | 62.2 57.8 58.6 | 61.9 36.6 45.3 |
| 1         | 61.6     | 64.1 47.8 53.2 | 67.7 54.5 58.6 | 55.3 31.1 39.6 |
| 2         | 55.6     | 62.3 48.3 52.8 | 66.0 53.4 57.5 | 52.9 35.7 41.2 |
| 3         | 58.4     | 65.2 52.6 56.9 | 67.9 59.4 62.0 | 58.3 35.7 44.2 |
| 4         | 49.1     | 50.2 N/A  N/A   | 70.2 58.4 62.9 | N/A  N/A  N/A  |
| 5         | 48.6     | 50.0 N/A  N/A   | 70.1 57.4 62.1 | N/A  N/A  N/A  |
| 6         | 47.4     | 48.3 N/A  N/A   | 67.6 57.6 60.7 | N/A  N/A  N/A  |
| 7         | 46.3     | 51.0 N/A  N/A   | 71.4 57.1 62.5 | N/A  N/A  N/A  |
| 8         | 48.6     | 48.1 N/A  N/A   | 67.4 56.5 60.5 | N/A  N/A  N/A  |
| 9         | 48.3     | 50.6 N/A  N/A   | 70.9 59.3 63.8 | N/A  N/A  N/A  |

**Avg.** 52.5 55.1 50.1 54.4 68.1 57.1 60.9 57.1 34.7 42.5

### TABLE III: Quantitative comparison results of different learning frameworks.

| Method  | Test (%) | Validation (%) |
|---------|----------|----------------|
|         | B-ACC    | B-ACC B-PER B-F1 | B-ACC B-PER B-F1 | B-ACC B-PER B-F1 |
| CL      | 79.2     | 81.5 65.3 70.7   | 82.8 64.4 70.5   | 78.2 67.6 71.5   |
| FedAvg [1] | 70.1    | 73.3 68.8 67.3   | 80.1 62.4 68.6   | 56.3 84.8 64.0   |
| FedProx [13] | 71.7   | 73.3 66.5 66.5   | 79.7 61.7 67.7   | 57.4 78.7 63.6   |
| MOON [15] | 73.1    | 73.6 62.4 65.4   | 79.2 60.7 66.8   | 59.8 66.8 62.0   |
| FedProc [16] | 66.8   | 68.5 63.6 62.5   | 77.0 62.1 67.5   | 47.1 67.3 49.9   |
| FedIRM [20] | 73.8   | 75.4 66.6 67.8   | 80.7 62.4 68.5   | 62.2 77.2 66.0   |
| FedRare  | 79.7     | 78.0 71.6 73.6   | 82.0 72.7 76.6   | 68.0 68.7 65.9   |

(i.e., $\ell_{FedProx} = \ell_{ce} + \beta \ell_{prox}$). For a fair comparison, we tune the hyper-parameter $\beta$ in FedProx from $\{0.001, 0.01, 0.1, 1\}$ and report the best results following the practice in [15].

5) MOON [15]: A FL framework for non-IID data. During local training, a contrastive loss is added to increase the similarity between the local model and the global model while decreasing the similarity compared to the previous local model (i.e., $\ell_{MOON} = \ell_{ce} + \beta \ell_{contrastive}$). We tune the hyper-parameter $\beta$ in MOON from $\{0.1, 1, 5, 10\}$ and report the best results.

6) FedProc [16]: A FL framework for non-IID data. FedProc builds global prototypes for all classes and uses contrastive learning for feature alignment. We carefully re-implement FedProc following the descriptions in [16].

7) FedIRM [20]: A FL framework for semi-supervised learning where the relationship across diseases is utilized to guide the training on the unlabeled data through inter-client supervision. For a fair comparison, the relationships learned from the full-class (i.e., with rare diseases) clients are utilized for the training of the clients without rare diseases.

E. Quantitative Results

Quantitative results of clients trained by LL are summarized in Table II. Limited by the training data of each client, the locally-learned models suffer from low generalizability, resulting in poor classification performance on the shared validation set. In addition, the classification performance of LL on the rare diseases (i.e., classes 5-6 in Table II) is much worse than that of other diseases, demonstrating the difficulties in correctly identifying the rare diseases. Based on the quantitative results of clients 0-3 and clients 4-9 on the classification of classes 0-4, the existence of rare diseases would also affect the classification performance of the common diseases. Quantitative results of LL across clients validate the great value of FL in learning from decentralized multi-client data for performance improvement.

Comparison results of different learning frameworks are stated in Table III. FedProc suffers the most from the existence of rare diseases, leading to the worst performance on both the validation set and the test set. It is mainly because the class-imbalance problem bought by the rare diseases can mislead the construction of the global prototypes in FedProc,
Fig. 5: The distributions of training data in the latent feature space corresponding to different learning frameworks, including local learning (client 0), FedAvg [1], FedProc [16], FedProx [13], MOON [15], FedIRM [20], the proposed FedRare, and centralized learning (CL). Classes are assigned with different colors for better visualization.

which would significantly affect the following training process. Comparatively, though FedAvg performs better than FedProc, the variances in the feature distributions of different classes among clients would produce severe parameter variances in aggregation, resulting in poor convergence. FedProx and MOON align the locally-learned models of clients to the federated model in the output space, which can improve the classification performance compared to FedAvg. However, the parameter variances in aggregation can hardly be alleviated. Among these approaches, FedIRM achieves the best classification performance. As FedIRM utilizes the feature distributions from the full-class clients to guide the training of those clients without rare diseases, the parameter gaps can be further reduced, producing consistent model parameters across clients.

Compared to FedIRM, the intra-client contrastive learning in FedRare, together with the positive sample queues for rare diseases, helps learn better latent features for classification. As the learned latent features are more separable, feature alignment across clients with and without rare diseases becomes easier. As a result, inter-client contrastive learning makes the model parameters more consistent, leading to better aggregation and convergence. Based on the quantitative results in Table III, FedRare outperforms other FL frameworks on both the validation set and the test set. Compared to CL, though the performance of FedRare is slightly worse, the classification results on the official test set are even better, demonstrating the high generalizability of FedRare.

In terms of the classification results on the common diseases (i.e. classes 0-4), the performance of FedRare is close to that of CL, validating the effectiveness of intra-client contrastive learning. For the rare diseases, though FedRare dramatically improves the classification performance of FL, there still exists a considerable performance gap compared to CL. In FedRare, as only the averaged latent features of clients, instead of the raw latent features of samples, are used for alignment, the diversity of training samples belonging to the rare diseases may not be sufficient compared to CL.

The training B-ACC curves of different learning frameworks are shown in Fig. 4. CL owns the best B-ACC performance but its training process is unstable especially at the early training stage, due to the limited data amount of the rare diseases. Comparatively, the training curves of FL frameworks are smoother, of which FedRare achieves the best B-ACC performance. The improvements are mainly brought by the intra-client contrastive learning in FedRare, making different classes more separable at the early training stage.

F. Qualitative Results

We perform PCA and t-SNE [48] to visualize the latent feature distributions of different learning frameworks as shown in Fig. 5. In general, the distributions are imbalanced and dominated by several major classes. For LL, according to the feature distributions, it can hardly distinguish different classes, especially for the rare diseases (i.e. classes 5-6). The poor generalizability of LL is mainly due to the limited
training data (i.e. 10% of the training set). Given more data, all other learning frameworks can produce more separable feature distributions, which is consistent with the quantitative results in Table III. Compared to FedAvg, introducing the interactions among clients in various ways is useful, resulting in better feature distributions of FedProc, FedProx, MOON, and FedIRM, especially for the class 6. Meanwhile, the feature distributions of those common diseases/classes can hardly be improved, showing the limitations of existing FL frameworks for data heterogeneity with rare diseases.

FedRare improves the feature distributions of both the common diseases and the rare diseases, compared to other FL frameworks. Thanks to inter-client contrastive learning, clients without rare diseases can make the common diseases more separable by referring to the federated features of the rare diseases, which in turn improves the classification performance of the common diseases. For rare diseases, both the intra-client contrastive learning and the positive sample queues can increase data diversity to fully utilize the distributed and limited samples. It should be noted that the feature distributions of the rare diseases in FedRare are not as compact as those of CL. In FedRare, as the samples across clients can not be directly shared due to data privacy, only the federated latent features of rare diseases are shared, which may limit the data diversity compared to CL.

V. DISCUSSION
In this section, we analyze the value of each component and hyper-parameter in FedRare through ablation studies.

A. Evaluation of Each Component
Quantitative comparison results of FedRare under different component combinations are given in Table IV. Compared to the baseline framework FedAvg, using $\ell_{\text{intra}}$ can better separate the latent features of classes, making it less challenging for aggregation, resulting in an average increase of 3.1% in B-ACC of the test set. By adopting the positive sample queues to assist intra-client contrastive learning, consistent performance improvements are achieved. Compared to $\ell_{\text{intra}}$, utilizing $\ell_{\text{inter}}$ is more beneficial, leading to an average increase of 7.4% in B-ACC compared to FedAvg. More specifically, the classification accuracy of the rare diseases increases from 56.3% to 69.1% while the corresponding accuracy of the common diseases decreases from 80.1% to 79.9%. It demonstrates the effectiveness of $\ell_{\text{inter}}$ for feature alignment of the rare diseases. Meanwhile, feature alignment through $\ell_{\text{inter}}$ can be counter-productive for the common disease, as their latent features may not be highly separable and consistent. Comparatively, jointly using $\ell_{\text{intra}}$ and $\ell_{\text{inter}}$ achieves the best overall performance. As discussed above, $\ell_{\text{intra}}$ targets the common diseases while $\ell_{\text{inter}}$ focuses more on the rare diseases. Therefore, FedRare is beneficial for not only the common diseases but also the rare diseases.

### TABLE IV: Quantitative results of FedRare with various combinations of different components.

| Components          | B-ACC (%) |
|---------------------|-----------|
| FedAvg $\ell_{\text{intra}}$ Queue $\ell_{\text{inter}}$ | Test Validation (classes) |
| ✓                   | ✓         | ✓         | ✓         | 0-6 | 0-4 | 5-6 |
| ✓                   | ✓         | ✓         | ✓         | 70.1| 73.3| 80.1| 56.3 |
| ✓                   | ✓         | ✓         | ✓         | 73.2| 74.2| 81.4| 56.4 |
| ✓                   | ✓         | ✓         | ✓         | 74.3| 75.2| 81.8| 58.6 |
| ✓                   | ✓         | ✓         | ✓         | 77.5| 76.7| 79.7| 69.1 |
| ✓                   | ✓         | ✓         | ✓         | 79.7| 78.0| 82.0| 68.0 |

B. Evaluation of Each Hyper-Parameter
One crucial hyper-parameter in FedRare is the dimension $Q$ of the latent features used in both intra- and inter-client contrastive learning. Quantitative results of FedRare with different values of $Q$ are summarized in Table V. According to Table V, the selection of $Q$ largely affects the performance ceilings of FedRare, where setting $Q$ to 256 achieves the best performance. Despite the performance gaps under different values of $Q$, FedRare consistently outperforms other FL frameworks, demonstrating the robustness of FedRare on medical image classification for both common and rare diseases.

Another important hyper-parameter in FedRare is $T$ in Eq. [17] determining the weight of each client in latent feature aggregation. According to the definitions in Eqs. [16] and [17] given class $c$, if $T$ is approaching 0, the federated latent feature $\hat{z}_c$ would bias to one specific client. In contrast, if $T$ is approaching $\infty$, the federated latent feature $\hat{z}_c$ is the average of clients’ latent features. Quantitative results of FedRare with varying $T$ are provided in Table VI. Instead of relying on certain clients or all clients for averaging, selecting those highly-confident clients to calculate the federated latent features for inter-client contrastive learning is more helpful.

### TABLE V: Quantitative results of FedRare on the official test set with various latent feature dimensions $Q$ for both intra- and inter-client contrastive learning.

| $Q$ | 128 | 256 | 512 |
|-----|-----|-----|-----|
| B-ACC (%) | 77.7 | 79.7 | 76.4 |

### TABLE VI: Quantitative results of FedRare on the official test set with different values of $T$ in Eq. [17] (i.e. weight calculation of each client).

| $T$ | 0.01 | 0.1  | 1    | 10   | 100  |
|-----|------|------|------|------|------|
| B-ACC (%) | 78.2 | 78.6 | 79.7 | 79.1 | 79.1 |

VI. Conclusion
In this paper, we study medical image classification for both common and rare diseases under the federated learning (FL) setting and propose a novel FL framework FedRare. Considering the relatively limited data amount of rare diseases, the data heterogeneity problem is challenging. To address this, we first propose intra-client contrastive learning, coupled with data re-sampling, to extract highly-separable latent features for all classes. Then, the federated latent features of classes are calculated based on the automatically-selected clients and sent back to clients for guidance. By taking the federated latent features as additional samples for inter-client contrastive learning, the latent feature distributions of diseases, including the rare diseases, become more consistent, leading to better
convergence. Experimental results on the publicly-available dataset demonstrate the superior performance of FedRare for medical image classification of both common and rare diseases. We believe the analysis of heterogeneous data brought by rare diseases will inspire future work on medical image classification in clinical applications.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” 2016, arXiv:1602.05629.

[2] C. Zhang et al., “A survey on federated learning,” Knowl-Based Syst., vol. 216, no. 106775, 2021.

[3] W. Y. B. Lim et al., “A simple framework for contrastive learning of visual representations,” in Proc. ICML, 2020.

[4] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. ICCV, 2017, pp.2242-2251.

[5] Y. Liu, J. Peng, J. J. Q. Yu, and Y. Wu, “PPGAN: Privacy-preserving generative adversarial network,” in Proc. ICPDS, 2019, pp. 985-989.

[6] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, “Big self-supervised models are strong semi-supervised learners,” in Proc. NeurIPS, 2020, pp. 22243-22255.

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. CVPR, 2020, pp. 9729-9738.

[8] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” 2020, arXiv:2003.04297.

[9] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” in Proc. ICCV, 2021, pp. 9640-9649.

[10] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, “Federated adversarial domain adaptation,” in Proc. ICLR, 2019.

[11] Q. Dou et al., “Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study,” NPJ Digit. Med., vol. 4, no. 1, pp. 1-11, 2021.

[12] N. Rieke et al., “The future of digital health with federated learning,” NPJ Digit. Med., vol. 3, no. 1, pp. 1-7, 2020.

[13] W. Zhang et al., “Dynamic-fusion-based federated learning for COVID-19 detection,” IEEE Internet Things J., vol. 8, no. 21, pp. 15884-15891, 2021.

[14] S. Silva et al., “Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data,” in Proc. ISBI, 2019, pp. 270-274.

[15] D. Li, A. Kar, N. Ravikumar, A. F. Frangi, and S. Fidler, “Federated simulation for medical imaging,” in Proc. MICCAI, 2020, pp. 159-168.

[16] O. Choudhury et al., “Differential privacy-enabled federated learning for sensitive health data,” 2019, arXiv:1910.02578.

[17] 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. CVPR, 2021, pp. 2423-2432.

[18] X. Li et al., “Multi-site fMRI analysis using privacy-preservation federated learning and domain adaptation: ABIDE results,” Med. Image Anal., vol. 65, no. 101765, 2020.

[19] Y. Zhao et al., “Federated learning with Non-IID data,” 2018, arXiv:1806.00582.

[20] F. Sadler, S. Wiedemann, K. -R. Müller, and W. Samek, “Robust and communication-efficient federated learning from Non-IID data,” IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 9, pp. 3400-3413, 2019.

[21] C. Briggs, Z. Fan, and P. Andras, “Federated learning with hierarchical clustering of local updates to improve training on Non-IID data,” in Proc. IJCNN, 2020, pp. 1-9.

[22] H. Wang, Z. Kaplan, D. Niu, and B. Li, “Optimizing federated learning on Non-IID data with reinforcement learning,” in Proc. INFOCOM, 2020, pp. 1698-1707.

[23] P. Tschandl et al., “The HAM10000 dataset, a large collection of multisource dermatoscopic images of common pigmented skin lesions,” Sci. Data, vol. 5, no. 180161, 2018.

[24] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, “Supervised contrastive learning for pre-trained language model fine-tuning,” in Proc. ICLR, 2021.

[25] N. Codella et al., “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC),” 2019, arXiv:1902.03368.

[26] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, 2019, pp. 6105-6114.

[27] P. Yao et al., “Single model deep learning on imbalanced small datasets for skin lesion classification,” IEEE Trans. Med. Imaging, 2021.

[28] N. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 11, 2008.