Label-Efficient Self-Supervised Federated Learning for Tackling Data Heterogeneity in Medical Imaging

Rui Yan, Liangqiong Qu, Qingyue Wei, Shih-Cheng Huang, Liyue Shen, Daniel Rubin, Lei Xing, Yuyin Zhou

***Abstract*— The collection and curation of large-scale medical datasets from multiple institutions is essential for training accurate deep learning models, but privacy concerns often hinder data sharing. Federated learning (FL) is a promising solution that enables privacy-preserving collaborative learning among different institutions, but it generally suffers from performance deterioration due to heterogeneous data distributions and a lack of quality labeled data. In this paper, we present a robust and label-efficient self-supervised FL framework for medical image analysis. Our method introduces a novel Transformer-based self-supervised pre-training paradigm that pre-trains models directly on decentralized target task datasets using masked image modeling, to facilitate more robust representation learning on heterogeneous data and effective knowledge transfer to downstream models. Extensive empirical results on simulated and real-world medical imaging non-IID federated datasets show that masked image modeling with Transformers significantly improves the robustness of models against various degrees of data heterogeneity. Notably, under severe data heterogeneity, our method, without relying on any additional pre-training data, achieves an improvement of 5.06%, 1.53% and 4.58% in test accuracy on retinal, dermatology and chest X-ray classification compared to the supervised baseline with ImageNet pre-training. In addition, we show that our federated self-supervised pre-training methods yield models that generalize better to out-of-distribution data and perform more effectively when fine-tuning with limited labeled data, compared to existing FL algorithms. The code is available at `https://github.com/rui-yan/SSL-FL`.**

This work was partially supported by the National Institutes of Health (NIH) under grants R01CA256890, R01CA227713, and U01CA242879.

R. Yan is with the Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: ruiyan@stanford.edu).

L. Qu is with the Department of Statistics and Actuarial Science and the Institute of Data Science, The University of Hong Kong, Hong Kong, 999077 (e-mail: liangqqu@hku.hk).

D.L. Rubin is with the Department of Biomedical Data Science, Stanford University, Stanford, CA 94305 USA (e-mail:rubin@stanford.edu).

Q. Wei and L. Shen are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: qywei@stanford.edu; liyues@stanford.edu).

SC. Huang is with the Department of Biomedical Informatics, Stanford University, Stanford, CA 94305 USA (e-mail: mschuang@stanford.edu).

L. Xing is with the Department of Radiation Oncology, Stanford University, Stanford, CA 94305 USA (e-mail: lei@stanford.edu).

Y. Zhou is with the Department of Computer Science and Engineering at University of California, Santa Cruz, CA 95064 (e-mail: zhouyuyiner@gmail.com).

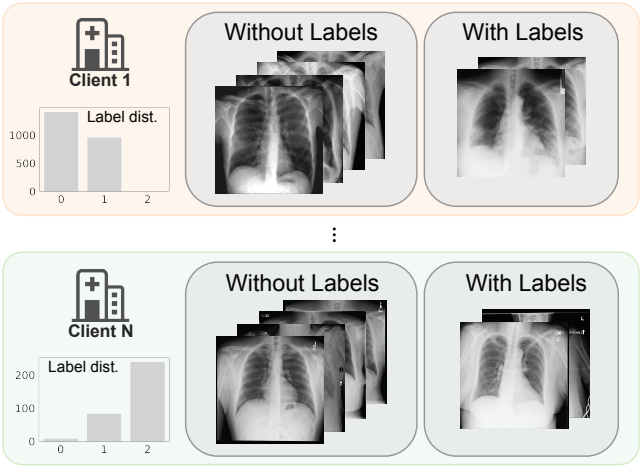


Fig. 1: Data heterogeneity and label deficiency of medical image datasets from different institutions.

***Index Terms*— Federated Learning, Self-supervised Learning, Vision Transformers, Data Efficiency**

I. Introduction

FEDERATED learning (FL) is a paradigm that allows model training using data distributed across multiple sites without explicit data sharing [1]. Compared to models trained at individual sites, federated models can be trained with a much more diverse and larger-scale dataset, which can result in superior performance and stronger generalizability. Therefore, this training paradigm has been widely adopted for critical medical applications such as the detection of brain tumors [2] and COVID-19 [3], [4], and applied to various types of data, including medical imaging data, electronic health records and sensor data [2], [5], [6].

As a decentralized approach, FL suffers from performance degradation due to data heterogeneity and label deficiency [6]–[8]. As shown in Fig. 1, data heterogeneity and label deficiency are particularly pronounced in medical image datasets of real-world applications. Regarding data heterogeneity, for example, some hospitals may have more data from patients at an early stage while the others may collect the data with severe conditions only (*i.e.*, label distribution skew). This is also referred to as statistical heterogeneity or non-identically distributed (non-IID) data partitions; large hospitals usually have more patient

data than community clinics (*i.e.*, quantity skew); and images at each hospital are acquired with different imaging acquisition protocols and on different patient populations (*i.e.*, feature distribution skew). In terms of label deficiency, some sites may not have enough bandwidth or incentive for a complete labor-intensive data labeling and thus only a small proportion of all available medical images may be labeled.

While several research efforts [9]–[12] have been devoted to addressing the challenges caused by data heterogeneity, current approaches tend to deteriorate in performance when using strongly skewed data distributions [13]–[15]. To handle extremely non-IID data partitions, recent studies [15] suggest that Vision Transformers (ViTs) [16] are better alternatives to convolution neural networks (CNNs), which have become the standard architecture used in the FL framework for image data. Qu *et al.* [15] reveal that simply replacing CNNs with ViTs outperforms even the state-of-the-art optimization-based FL methods. However, the success of such models largely relies on supervised ImageNet pre-training, which could suffer from domain discrepancy when fine-tuning with medical images and can be further improved by self-supervised pre-training on a centrally shared large-scale in-domain medical dataset [17]. However, such centrally shared datasets rarely exist in the medical domain due to privacy and ownership concerns. Therefore, it is desired to build a self-supervised FL framework that collaboratively learns a global model by leveraging all available unlabeled data without sharing data among institutions.

Label deficiency is a common challenge in medical imaging. This can make it difficult to train accurate deep learning models on medical imaging data, as these models typically require large amounts of labeled data to learn effectively. To address this issue, various approaches such as semi-supervised and self-supervised learning methods [17]–[22] have been proposed to allow models to learn from partially labeled or unlabeled data. However, many of these methods assume that the data is centralized, which are not practical for decentralized data. To enable privacy-enhancing model training on decentralized medical data, Yang *et al.* [23] combine semi-supervised learning strategies such as consistency loss [24] with FL, referred to as Semi-FL. Recently, several federated contrastive learning (FCL) methods [25]–[27] have been proposed. However, FCL methods often yield sub-optimal results when the data is highly heterogeneous or limited at local clients.

In this paper, we propose a robust self-supervised FL framework to address these challenges as shown in Fig. 2. To the best of our knowledge, this is the first work that simultaneously tackles the issues of data heterogeneity and label deficiency for medical imaging in FL leveraging masked image modeling as the self-supervised task. Self-supervised pre-training has been demonstrated to be an effective solution to alleviate the need for large-scale labeled pre-training datasets and potentially generalizes better across various tasks [28]. Moreover, unlike supervised learning which relies heavily on the label information, self-supervised pre-training learns the intrinsic features of images in local clients without labels, embodying less label-specific inductive bias and thus, less susceptible to label distribution skewness. The proposed method learns visual representations more effectively across non-IID clients, even when data are limited at some clients. To this end, we design a distributed self-supervised learning paradigm to improve FL in medical imaging, which consists of two essential steps: (1) *federated self-supervised pre-training*, which exploits knowledge from decentralized unlabeled data based on masked image modeling in a distributed setting; (2) *federated supervised fine-tuning*, which then transfers this knowledge to the target tasks by fine-tuning the federated models.

Specifically, we implement two masked image modeling methods, BEiT [29] and MAE [30], as the SSL module in our federated framework. We evaluate their performance in both centralized and federated settings across diverse medical imaging tasks and demonstrate that BEiT and MAE together with Transformers are robust to distribution shifts and facilitate effective representation learning with limited amounts of data.

We conduct extensive experiments under different degrees of data heterogeneity and with different fractions of labeled data on diverse medical datasets including diabetic retinopathy images, dermatology images and chest X-rays to validate the broad effectiveness of our self-supervised FL framework.

Our main contributions are summarized as follows:

- We design a privacy-preserving federated self-supervised pre-training framework that uses masked image modeling to learn visual representations from decentralized data. Our proposed framework can tackle data heterogeneity and label deficiency at once.
- Our results on diverse medical datasets demonstrate that the proposed method is more label-efficient and robust to non-IID data compared to ImageNet supervised baselines and existing FL algorithms.
- For evaluation of a real-world distribution, we construct a federated chest X-rays benchmark called COVID-FL by curating the data from 8 different medical sites for testing the model's robustness in a realistic federated setting.

II. RELATED WORK

A. Federated Learning

Federated learning (FL) is a distributed training technique that trains machine learning models on private data across decentralized clients. As a standard FL algorithm, FedAvg [1] performs local model training via stochastic gradient descent updates at each client, followed by a model aggregation at the server. However, FedAvg generally suffers from performance degradation due to the weight divergence issue caused by the non-IID data partitions (*i.e.*, data heterogeneity) [10], [31]. Many efforts have been devoted to addressing this issue, including regularizing the local model learning in parameter space [10], [12], sharing a subset of data among clients [14], guiding local training with knowledge distillation [32], [33], and introducing effective global model aggregation strategies [11], [34]. Recent studies tackle data heterogeneity from the perspective of model initialization, suggesting that pre-training alleviates the drastic accuracy drop caused by data heterogeneity [35], [36]. Moreover, [37] shows that Vision Transformer [16] pre-trained on ImageNet leads to significant

performance gain on non-IID data, outperforming its CNNs counterpart and optimization-based FL methods. However, these FL methods assume fully labeled samples are available, which is not always feasible in the medical domain. An FL method that could handle both data heterogeneity and limited annotations is desired.

B. Self-supervised Learning

Self-supervised learning (SSL), a method that exploits unlabeled data by using the data itself to provide the supervision, has gained popularity because of its ability to learn effective image representations [38], [39] and avoid the cost of annotating large-scale datasets. The core of SSL lies in adopting pretext tasks as self-supervision and then using the learned representations for different downstream tasks. Various pretext tasks have been proposed such as image inpainting [40] and jigsaw puzzle [41], which have also been shown to be beneficial to medical imaging tasks [42]. A more recent research strand focuses on contrastive learning, which learns visual representations by forcing the representations of positive pairs to be closer while far apart for negative pairs. To obtain enough informative negative pairs, contrastive learning methods such as MoCo [43] and SimCLR [44] rely on a large memory bank or batch size. BYOL [45] instead trains one network to predict representations of the same image under a different augmented view obtained from the other network.

With the recent advance in Vision Transformer (ViT) [16], multiple works such as BEiT [29] and MAE [30] have been proposed to learn visual representations by signal reconstruction given corrupted images. We refer to this type of methods as masked image modeling. As opposed to contrastive learning, masked image modeling does not heavily depend on a large sample size or certain compositions of data augmentation while achieving competitive performance [30].

C. Federated Self-supervised Learning

Federated self-supervised learning is referred to as applying self-supervised pre-training to learn representations from unlabeled decentralized data. It has attracted increasing attention in FL community given its capability to facilitate model learning when labeled data are limited at local clients.

Previous works [25]–[27] mainly consider contrastive learning as the self-supervised task, known as federated contrastive learning (FCL). Note that contrastive learning needs large and diverse data to generate enough informative negative samples to train a good model, while a client in FL generally lacks of data diversity, especially for non-IID cases. Regarding this issue, FedCA [25] shares local data features among clients. FedEMA [27] updates local models by using the exponential moving average of the global model. Several works [46], [47] have been proposed for medical imaging tasks by using MoCo as the self-supervised method. To address the limited local data diversity, [46] shares local negative samples, and [47] shares the metadata of local image representations among clients. Most FCL works [25]–[27], [47], however, consider only label distribution skew in their heterogeneity experiments, where the number of images at each client is sufficient (*e.g.*, greater than 10000) and each client contains the same number of images. It remains elusive how these FCL methods perform under highly heterogeneous data partitions with limited local data sample size (*e.g.*, less than 200 images for certain clients).

In our work, we develop the first federated self-supervised pre-training framework that employs masked image modeling as the self-supervised task. We demonstrate that masked image modeling (*i.e.*, MAE and BEiT) coupled with Transformers is robust to distribution shifts and could learn effectively when data are relatively limited. The proposed pre-training scheme significantly advances the capability of federated models over highly heterogeneous data partitions.

III. Methodology

A. Problem Statement

Our work aims at building a robust model that collaboratively learns from decentralized clients without data sharing. Specifically, our goal is to improve the model performance in FL especially for non-IID client data and in limited label scenarios. Suppose there are N clients. Each client $k \in \{1, .., N\}$ has a local dataset $\mathcal{D}^k$. To learn a generalized global model over $\mathcal{D} = \bigcup_{k=1}^{N} \{\mathcal{D}^k\}$, the global objective function is defined as follows:

$$\arg\min_{w} \mathcal{L}(w) = \sum_{k=1}^{N} \frac{|\mathcal{D}^k|}{|\mathcal{D}|} \mathcal{L}_k(w), \tag{1}$$

and the local objective function $\mathcal{L}_k(w)$ in client k measuring the local empirical loss over data distribution $\mathcal{D}^k$ is defined as:

$$\mathcal{L}_k(w) = \mathbb{E}_{x \sim \mathcal{D}^k}[\ell_k(w; \boldsymbol{x})], \tag{2}$$

where ℓ_k is the loss function used for client k, and w denotes the global model parameters to be learned.

The focus of our work is to address the data heterogeneity issue in FL, given that the data across different clients are usually non-IID, *i.e.*, $\mathcal{D}^m$ and $\mathcal{D}^n$ $(m \neq n)$ follow different distributions $P_m(\boldsymbol{x}, y)$ and $P_n(\boldsymbol{x}, y)$. Furthermore, considering that some local clients may not have sufficient labeled data due to the lack of resources, in this paper, we also investigate how FL performs under limited annotation, *i.e.*, local dataset $\mathcal{D}^k$ comprises labeled data $\mathcal{D}_l^k = \{(\boldsymbol{x}, y)\}$ and unlabeled data $\mathcal{D}_u^k = \{\boldsymbol{x}\}$, where $|\mathcal{D}_l^k|$ is relatively small.

B. Generalized Framework

To address this important problem, we propose a generalized self-supervised FL framework to enhance both the robustness and the performance of federated models when learning from decentralized data with statistical heterogeneity. Our framework comprises two stages: a federated self-supervised pre-training stage and a supervised federated fine-tuning stage, as shown in Fig. 2 and Alg. 1. During the self-supervised stage, the model exploits knowledge from decentralized data by pre-training with masked image modeling in a distributed setting. In the supervised federated fine-tuning stage, the knowledge is transferred from the previous stage to the target task by fine-tuning the federated models.

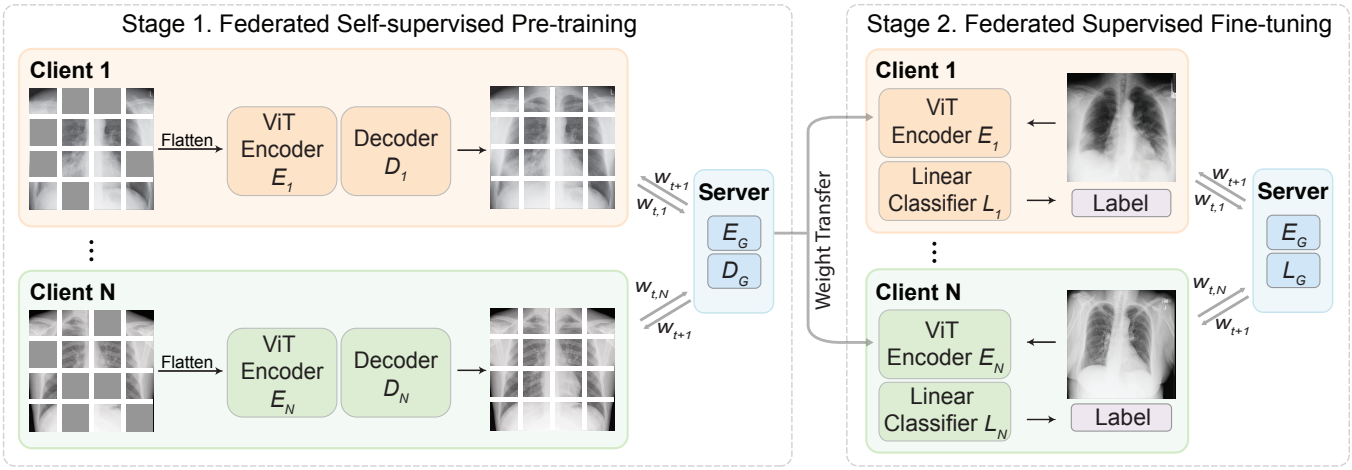


Fig. 2: Overview of the federated self-supervised learning framework. In the pre-training stage (left), masked image modeling is used as the self-supervised task to learn representations from unlabeled images in each client. The pre-training process consists of three steps and ends when it reaches the maximum communication rounds T. At round t, (1) Each client k ($k \in \{1, ..., N\}$) trains its local auto-encoder E_k and D_k with the unlabeled local data; (2) Client k uploads the weights of its auto-encoder $w_{t,k}$ to the central server; (3) The server produces a global auto-encoder E_G and D_G with weights w_{t+1} via model weights averaging and broadcasts the global model back to each local client. In the fine-tuning stage (right), the final pre-trained global encoder E_G^* from the first stage is used to initialize each local encoder E_k. A linear classifier L_k is appended to each local encoder. End-to-end federated fine-tuning is performed on labeled images in each local client.

Specifically, we integrate two popular masked image modeling methods, BEiT [29] and MAE [30], into our generalized federated framework. We denote BEiT and MAE coupled with our framework as Fed-BEiT and Fed-MAE, respectively. The pre-training and fine-tuning details of Fed-BEiT and Fed-MAE are illustrated in Sec. III-C and Sec. III-D.

C. Federated Self-supervised Pre-training

During pre-training, the k-th local model is an autoencoder consisting of an encoder E_k and a decoder D_k. The model is trained using masked image modeling, which involves masking a subset of image patches and reconstructing the original signals in the masked patches. We implement two popular masked image modeling methods, BEiT [29] and MAE [30], as the SSL module in our federated framework. In this section, we describe the main components of our proposed federated pre-training protocol.

For the k-th client, an input image $\boldsymbol{x} \sim \mathcal{D}^k$ ($\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$) is divided into a sequence of image patches $\boldsymbol{x}_p = \{\boldsymbol{x}_p^i\}_{i=1}^P \in \mathbb{R}^{P \times (S^2 \cdot C)}$, where (H, W) is the dimension of the original image, C is the number of channels, (S, S) is the dimension of each image patch and $P = HW/S^2$ is the number of image patches.

1) Masking: We denote the masking ratio as γ, the masked positions as $\mathcal{M}$, and the unmasked positions as $\mathcal{V}$. After randomly masking $\gamma\%$ of image patches, we get $|\mathcal{M}| = \gamma P$ and $|\mathcal{M}|+|\mathcal{V}| = P$. The total image patches can be represented as: $\boldsymbol{x}_p = \boldsymbol{x}_p^{\mathcal{M}} \cup \boldsymbol{x}_p^{\mathcal{V}} = \{\boldsymbol{x}_p^i, i \in \mathcal{M}\} \cup \{\boldsymbol{x}_p^i, i \in \mathcal{V}\}$, where $\boldsymbol{x}_p^{\mathcal{M}}$ represents the masked patches and $\boldsymbol{x}_p^{\mathcal{V}}$ represents the unmasked visible patches. Specifically, BEiT uses block-wise (n-gram) masking [29], and MAE uses random masking.

2) Encoder: We employ ViT [16] as our encoder and apply it to a sequence of image patches as shown in Fig. 2.

- For BEiT, the input to the ViT encoder is

$$\{\boldsymbol{x}_p^{\mathcal{V}}\boldsymbol{E}\} \cup \{\boldsymbol{e}_p^{\mathcal{M}}\} \in \mathbb{R}^{P \times D}, \boldsymbol{E} \in \mathbb{R}^{(S^2 \cdot C) \times D},$$

  where $\boldsymbol{x}_p^{\mathcal{V}}\boldsymbol{E} \in \mathbb{R}^{|\mathcal{V}| \times D}$ is the linear projection of the visible patches to dimension D and $\boldsymbol{e}_p^{\mathcal{M}} = \{\boldsymbol{e}_p^i, i \in \mathcal{M}\} \in \mathbb{R}^{|\mathcal{M}| \times D}$ is a learnable embedding for the masked patches. The output of the encoder is $\{\boldsymbol{h}_i\}_{i=1}^P$ ($\boldsymbol{h}_i \in \mathbb{R}^D$) represents the encoded representations of the i-th patch.
- For MAE, the ViT encoder takes only the linear projection of the visible patches $\boldsymbol{x}_p^{\mathcal{V}}\boldsymbol{E}$ as the input with added position embeddings. The output of the encoder is $\{\boldsymbol{h}_i, i \in \mathcal{V}\}$ where $\boldsymbol{h}_i \in \mathbb{R}^D$ represents the encoded visible patch $i \in \mathcal{V}$.

3) Decoder: Our decoder performs the signal reconstruction task given the encoded representations of the input patches.

- For BEiT, the inputs to the decoder are the encoded representations for all the patches $\{\boldsymbol{h}_i\}_{i=1}^P$ obtained from the last layer of the encoder. The decoder is a single linear layer to predict the visual tokens at the masked positions $\{\boldsymbol{z_i}, i \in \mathcal{M}\}$ which was generated by the DALLE pre-trained dVAE [48] tokenizer.
- For MAE, the inputs to the decoder are the encoded visible patches $\{\boldsymbol{h}_i, i \in \mathcal{V}\}$ along with a learnable vector for the masked patches $\boldsymbol{e}_p^{\mathcal{M}} = \{\boldsymbol{e}_p^i, i \in \mathcal{M}\}$ and position embeddings. The decoder is a lightweight ViT that regresses the pixel values for the masked patches.

4) Loss function: The k-th local encoder E_k and decoder D_k are trained with their local data $\mathcal{D}^k$ to minimize the local objective function $\mathcal{L}_k(w) = \mathbb{E}_{x \sim \mathcal{D}^k}[\ell_k(w; \boldsymbol{x})]$.

- In BEiT, ℓ_k is the cross-entropy loss of the predicted visual tokens of the masked patches $\{\boldsymbol{z}_i, i \in \mathcal{M}\} \in \mathbb{R}$:

$$\ell_k = -\sum_{i \in \mathcal{M}} \frac{1}{|\mathcal{M}|} \log p(\boldsymbol{z}_i; w | \boldsymbol{x}^{\mathcal{M}}), \tag{3}$$

Algorithm 1: Our federated self-supervised learning framework. T is the maximum number of communication rounds, E is the number of local epochs.

Input: local client k, local data $\mathcal{D}^k = \mathcal{D}_l^k \bigcup \mathcal{D}_u^k$
Server Execution:
  initialize w_0
  for *each round* $t = 1, ..., T$ **do**
    $S_t \leftarrow$ (Selection of K clients)
    for *each client* $k \in S_t$ *in parallel* **do**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    ▷ Pre-training stage
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{|\mathcal{D}^k|}{|\bigcup \mathcal{D}^k|} w_{t+1}^k$
    ▷ Fine-tuning stage
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{|\mathcal{D}_l^k|}{|\bigcup \mathcal{D}_l^k|} w_{t+1}^k$
ClientUpdate(k, w_t)
  ▷ Pre-training stage
  Sample batches $\mathcal{B}$ from local data $\mathcal{D}^k$
  for *each local epoch* $i = 1, ..., E$ **do**
    for *batch* $b \in \mathcal{B}$ **do**
      $b^{\mathcal{M}} = \bigcup_{\boldsymbol{x} \in b} \text{Masking}(\boldsymbol{x}) = \bigcup_{\boldsymbol{x} \in b} \boldsymbol{x}^{\mathcal{M}}$
      $w_{t+1}^k \leftarrow w_{t,k} - \eta \nabla \ell_k(w_{t,k}, b^{\mathcal{M}})$
  ▷ Fine-tuning stage
  Sample batches $\mathcal{B}$ from local labeled data $\mathcal{D}_l^k$
  for *each local epoch* $i = 1, ..., E$ **do**
    for *batch* $b \in \mathcal{B}$ **do**
      $w_{t+1}^k \leftarrow w_{t,k} - \eta \nabla \ell_k(w_{t,k}, b)$

- In MAE, ℓ_k is the mean squared error of the predicted pixel values of the masked patches $\{\boldsymbol{x}_p^i, i \in \mathcal{M}\} \in \mathbb{R}^{S \times S \times C}$:

$$\ell_k = \sum_{i \in \mathcal{M}} \frac{1}{|\mathcal{M}|} ((\boldsymbol{x}_p^i - \hat{\boldsymbol{x}}_p^i)^2; w), \tag{4}$$

In the federated pre-training stage, each local client takes E steps of gradient descent to update the local model E_k and D_k by minimizing its local loss $\mathcal{L}_k$ on data $\mathcal{D}^k$. Then, the server takes a weighted average of all the resulting local models to update the global model E_G and D_G, which is further sent back to the local clients for the next training iteration. The whole pre-training process terminates when we reach the maximum number of communication rounds T. Once pre-training is complete, the final pre-trained global encoder (E_G^*) is saved while the final global decoder (D_G^*) is discarded.

D. Supervised Federated Fine-tuning

In the federated fine-tuning stage, as shown in Fig. 2, we initialize the local encoder E_k of the k-th client with the pre-trained global encoder E_G^* obtained from the first stage, and append a linear classifier L_k upon the encoder. The entire model is then fine-tuned on the local labeled data. Specifically, we use average pooling to extract the learned representations from the local encoder, which are then fed to a linear classifier L_k (*i.e.*, a softmax layer) to minimize the cross-entropy loss for image classification tasks.

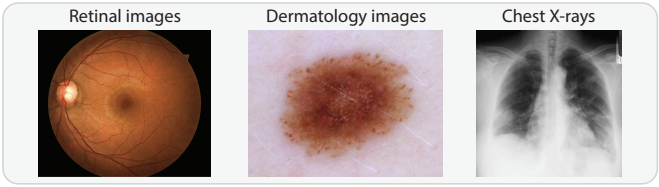


Fig. 3: Preview of retinal images, skin images and chest X-rays

IV. Experiments

In this section, we present experiments on a variety of simulated and real-world federated medical datasets to evaluate the effectiveness of our methods.

We first provide details on the datasets and experimental setup, then compare the robustness of our methods to data heterogeneity with baselines that are pre-trained on ImageNet. Additionally, we analyze the generalizability of the proposed method to out-of-distribution data and investigate its label efficiency through fine-tuning with different fractions of labels. Furthermore, we compare the performance of our methods to previous FL methods, including (1) federated self-supervised pre-training baselines and (2) optimization-based FL methods and semi-supervised FL methods, in terms of robustness to non-IID data and label efficiency.

A. Dataset

We evaluate the performance of our methods on three popular tasks in the medical imaging domain: (1) detecting diabetic retinopathy from retinal fundus images, (2) diagnosing skin lesions from dermatology images, and (3) identifying pneumonia and COVID-19 from chest X-rays. These tasks vary in terms of image modality, image acquisition, label distribution, and other factors. For example, retinal images are obtained using fundus cameras, dermatology images are captured with digital cameras, and chest X-rays are acquired using X-ray scanners. Fig. 3 illustrates the visual differences among these three medical datasets.

Retina Dataset. We evaluate FL methods on the Kaggle Diabetic Retinopathy competition dataset[1], which contains 35,126 retinal fundus images acquired from various cameras. The original images are divided into five categories (normal, mild, moderate, severe, and proliferating). We preprocessed the dataset by binarizing the labels into Normal and Diseased, and randomly selecting 9,000 balanced images as the training set and 3,000 images as the test set.

Dermatology Dataset. This dermatology dataset is referred to as *Derm* in this paper. It includes images from ISIC17, 19, 20[2], with approximately 5,000 images in the Melanoma (malignant) class from these three datasets and 5,000 randomly selected images in the benign classes from ISIC19. The Derm dataset is then randomly divided into a training set of approximately 7,500 images and a test set of approximately 2,500 images.

COVID-FL dataset. To evaluate the performance of our methods on a real-world federated data partitions, we create *COVID-FL*, a dataset in which each client only contains

[1]https://www.kaggle.com/c/diabetic-retinopathy-detection
[2]https://challenge.isic-archive.com/data/

data from a single real-world site without any overlap. Our COVID-FL dataset includes 20,018 chest X-ray scans from eight different publicly available data repositories: (1) BIMCV-COVID19 [49] (the Valencia Region Image Bank, Spain), (2) ml-workgroup[3] (the Institute for Diagnostic and Interventional Radiology, Hannover, Germany), (3) SIRM[4] (the Italian Society of Medical and interventional Radiology COVID-19 Database, Italy), (4) Eurorad[5] (the European Society of Radiology), (5) MIDRC-RICORD-1c[6] (the RSNA International COVID-19 Open Radiology Database), (6) the RSNA Pneumonia Detection Challenge dataset[7], (7) the Guangzhou pediatric dataset [50] (from the Guangzhou Women and Children's Medical Center, China), and (8) the Cohen dataset[8] with duplicated images removed.

In our COVID-FL dataset, each data site represents a single medical institution in order to mimic the real-world federated scenarios. Each site may be missing one or more classes. For example, BIMCV only contains images of COVID-19 infections, while Guangzhou pediatric only includes images of normal and non-COVID-19 pneumonia patients (Fig. 6a). Data from different sites were acquired using different machines and on different patient populations, resulting in heterogeneity in intensity distribution (Fig. 6b). This simulates the real-world scenario where data is collected from various institutions with different equipment and patient populations.

The COVID-FL dataset is further divided into an 80%-20% train-test split, yielding 16,044 training images and 3,974 test images. Each data site has the same proportion of train and test sets. The test set can be considered as a combination of the held-out data in each client (*i.e.*, hospital).

Skin-FL dataset. The Skin-FL dataset contains skin lesion images and was created to evaluate the generalization of the model to out-of-distribution data. Following [51], after removing duplicates, the training set of Skin-FL consists of 22,888 images from four datasets as shown in Fig. 7a: 784 images from Derm7pt [52], 8,012 images from HAM10000 [53], 1,839 images from PAD-UFES [54], and 12,253 images from ISIC19. There are eight classes in total in the training set: Actinic keratosis (AK), Benign keratosis (BKL), Melanoma (MEL), Melanocytic nevus (NV), Vascular lesion (VASC), Squamous cell carcinoma (SCC), Basal cell carcinoma (BCC), and Dermatofibroma (DF). Specifically, Derm7pt has six classes except for AK and SCC; HAM10000 includes seven classes except for SCC; PAD-UFES contains six classes except for DF and VASC; ISIC19 contains all eight classes.

Moreover, we use 33,126 images from ISIC20 [55] as our out-of-distribution test set to investigate how our proposed method generalizes to unseen clients. This test set contains several classes that are not included in the training set, so we binarize the predictions during fine-tuning and the labels of our test set into Benign and Malignant following [51]. This task is very challenging due to the severe class imbalance (Fig. 7b).

[3]https://github.com/ml-workgroup/covid-19-image-repository
[4]https://www.sirm.org/category/senza-categoria/COVID-19/
[5]https://eurorad.org/
[6]https://doi.org/10.7937/91ah-v663
[7]https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/
[8]https://github.com/ieee8023/covid-chestxray-dataset

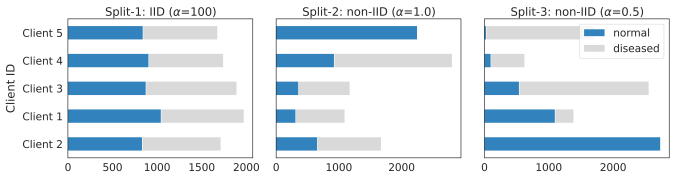


Fig. 4: Retina: visualization of statistical heterogeneity among clients, where the x-axis is the number of training samples, and the y-axis is client IDs.

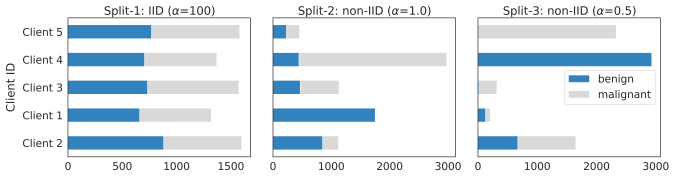


Fig. 5: Derm: visualization of statistical heterogeneity.

B. Experiment Setup

1) Construction of non-IID dataset: We model IID and non-IID data distributions using a Dirichlet distribution following [32], [56], [57] for Retina and Derm dataset. Compared to real federated data partitions, simulated data partitions allow for a more flexible and thorough investigation of the model behavior, as they can be easily manipulated to test different degrees of data heterogeneity. Suppose a dataset has J classes, we randomly partition the data into N local clients by simulating

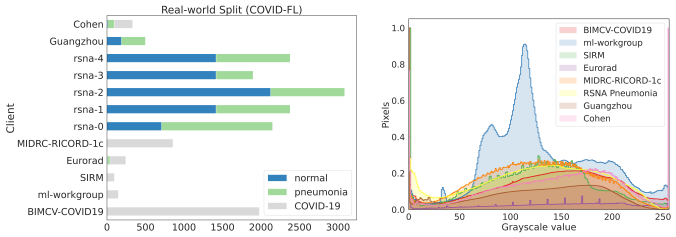
$$\boldsymbol{p}_j = \{p_{j,1}, ..., p_{j,N}\} \sim \mathrm{Dir}_N(\alpha)$$

where $p_{j,k} \in (0,1)$ and $||\boldsymbol{p}_j||_1 = 1$ ($j \in [1, J]$, $k \in [1, N]$).

We assign a proportion $p_{j,k}$ of the instances of class j to client k. The concentration parameter α in the Dirichlet distribution $\mathrm{Dir}(\alpha)$ controls the degree of heterogeneity, with smaller values of α leading to higher data heterogeneity. We simulate three sets of data partitions with α values of ($\alpha = \{100, 1.0, 0.5\}$) for both the Retina and Derm datasets, each of which consists of $N = 5$ simulated clients (see Fig. 4 and 5). Based on the level of data heterogeneity, these three partitions are referred to as Split-1 (IID), Split-2 (moderate non-IID) and Split-3 (severe non-IID).

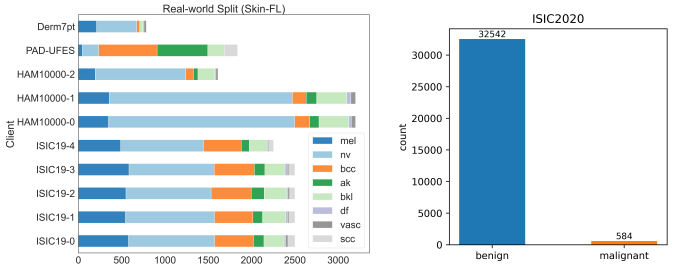
COVID-FL and Skin-FL are two real-world federated data sets that exhibit both label distribution skewness and feature distribution skewness. Given that some clients contain significantly more data than others, we partition those clients into sub-clients without any overlap, resulting in a total of 12 clients in COVID-FL and 10 clients in Skin-FL. The distributions of these two datasets are shown in Fig. 6a and 7a.

2) Data Augmentation: During pre-training, we apply random scaling and crop patches of size 224×224 from the original images for all datasets, followed by random color jittering and random horizontal flipping. The random scaling factor is chosen from a range of [0.4, 1.0] for COVID-FL and [0.2, 1.0] for the other datasets. During fine-tuning, we apply random scaling and cropping the images patches of size 224×224 and perform random rotation with a degree of 10 and random horizontal flipping. The random scaling factor is chosen from a range of [0.8, 1.2] for COVID-FL and [0.6, 1.0] for the other datasets.



(a) Label distribution on COVID-FL (b) Chest X-ray intensity distribution at each client

Fig. 6: COVID-FL: data heterogeneity among clients.



(a) Label distribution on Skin-FL train set (b) Out-of-distribution test set

Fig. 7: Skin-FL: data heterogeneity among clients.

3) Self-supervised FL pre-training setup: All methods are implemented using Pytorch and deployed in a distributed training system using DistributedDataParallel (DDP). ViT-B [16] is chosen as the backbone for the proposed models. Following the setup in BEiT [29] and MAE [30], the input is split into 14×14 image patches and the same number of visual tokens for BEiT and 16×16 patches for MAE. In our main experiment, we randomly mask at most 40% of total image patches for BEiT and 60% for MAE. AdamW with $\beta 1 = 0.9$, $\beta 2 = 0.999$ is used for optimization.

We use the same set of hyperparameters for both centralized and federated learning in each task. The base learning rate (η) and batch size (B) vary among tasks based on hyperparameter tuning. More details can be found in Table I. Fed-BEiT pre-training runs for 1000 communication rounds with a warmup period of 10 epochs; Fed-MAE pre-training runs for 1600 communication rounds with a warmup period of 5 epochs. Both methods employ a cosine learning rate decay of 0.05.

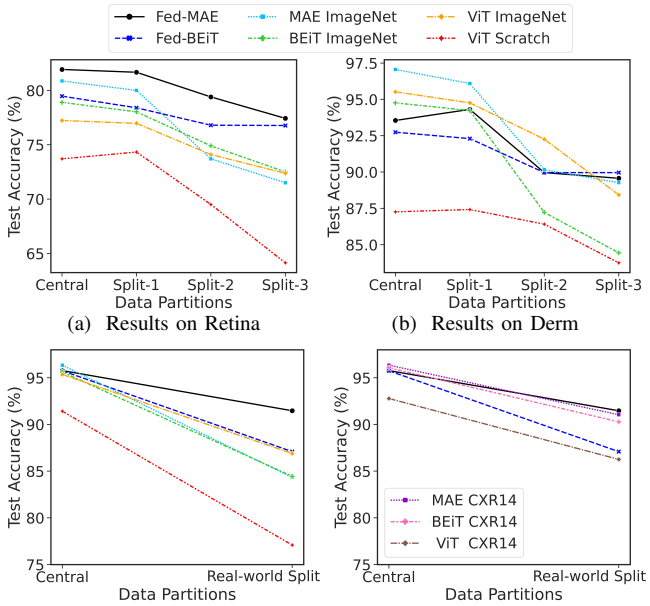
In terms of the pre-training schedule, we observe that a larger number of communication rounds generally leads to more improvement, but the improvement becomes much less prominent after a certain number of rounds. Here, we use 1000 and 1600 communication rounds as the default values for Fed-BEiT and Fed-MAE pre-training, respectively. Note that we also conduct ablation studies on the number of communication rounds and investigate its impact on model performance in Sec. IV-D.

4) Supervised FL fine-tuning for downstream tasks: During federated fine-tuning, the model is fine-tuned for 100 communication rounds with a base learning rate starting at 3e-3 and a batch size of 256 for all tasks, except for COVID-FL, whose batch size is set to 64.

5) Evaluation Metrics: We use accuracy as the evaluation metric for classification on the Retina, Derm and COVID-FL datasets. For the Skin-FL dataset, due to its severe class imbalance issues, we use F1-score as the evaluation metric.

TABLE I: Table of hyper-parameters (base learning rate η and batch size B) in experiments on the Retina, Derm, COVID-FL and Skin-FL datasets during federated self-supervised pre-training.

Dataset	Fed-BEiT		Fed-MAE	
	η	B	η	B
Retina	1.5e-3	256	1e-3	128
Derm	3e-3	128	7.5e-4	128
COVID-FL	1.5e-3	64	3.75e-4	64
Skin-FL	1.5e-3	128	7.5e-4	128



(a) Results on Retina (b) Results on Derm

(c) Results on COVID-FL (left: comparison with training from scratch and with ImageNet pre-training; right: comparison with CXR14 pre-training)

Fig. 8: Comparison of model performance in terms of test accuracy w.r.t. data heterogeneity for the Retina, Derm and COVID-FL datasets.

C. Results

To evaluate the proposed federated self-supervised pre-training methods (Fed-BEiT and Fed-MAE), we compare them with four baseline approaches, including: (1) no pre-training (ViT scratch), (2) ImageNet supervised pre-training (ViT ImageNet) [15], (3) ImageNet pre-training using BEiT [29] (BEiT ImageNet), and (4) ImageNet pre-training using MAE [30] (MAE ImageNet).

We pre-train methods directly on decentralized target task data in a distributed setting, while the four baseline approaches (except for ViT Scratch which does not require pre-training) are pre-trained on a centralized large-scale dataset ImageNet-22K [58] under centralized settings. Furthermore, on COVID-FL, we conduct experiments using pre-trained models trained on the large dataset ChestX-ray14 (CXR14) [59] as additional baselines (ViT CXR14, BEiT CXR14 and MAE CXR14), similar to the three ImageNet pre-training baselines. CXR14 consists of 112,120 chest X-ray images, which is seven times larger in size than the COVID-FL training set.

All of the methods in this comparison utilize ViT-B [16] as their backbone for fairness. The federated fine-tuning process is run for 1000 communication rounds for models trained from random initialization, and 100 rounds for the others.

TABLE II: Test accuracy for federated fine-tuning on Retina Central (centralized), Split-1 (IID), Split-2 (moderate non-IID) and Split-3 (severe non-IID). The best result is bolded and the second-best result is marked with a line underneath.

Pre-training			Test Accuracy (%)			
Method	Setup	Dataset	Central	Split1	Split2	Split3
None	*None*	*None*	73.70	74.33	69.50	64.13
Supervised	Centralized	ImageNet	77.23	76.97	74.10	72.37
BEiT [29]	Centralized	ImageNet	78.90	78.03	74.90	72.50
MAE [30]	Centralized	ImageNet	80.87	80.00	73.70	71.50
Fed-BEiT	Distributed	Retina	79.47	78.40	76.80	76.77
Fed-MAE	Distributed	Retina	**81.93**	**81.67**	**79.40**	**77.43**

TABLE III: Test accuracy for federated fine-tuning on the Derm dataset with various degrees of data heterogeneity.

Pre-training			Test Accuracy (%)			
Method	Setup	Dataset	Central	Split1	Split2	Split3
None	*None*	*None*	87.26	87.42	86.41	83.75
Supervised	Centralized	ImageNet	95.52	94.76	**92.26**	88.43
BEiT [29]	Centralized	ImageNet	94.76	94.23	87.22	84.44
MAE [30]	Centralized	ImageNet	**97.06**	**96.09**	90.16	89.27
Fed-BEiT	Distributed	Derm	92.74	92.30	89.96	**89.96**
Fed-MAE	Distributed	Derm	93.55	94.31	89.96	89.57

1) More robust to data heterogeneity: Data heterogeneity is a key FL challenge that our work aims to address. Fig. 8 and Table II-IV compare the results of our proposed method and the baselines under different degrees of data heterogeneity for multiple medical image classification tasks.

First, we observe that our proposed method is the only method that is consistently robust across all medical tasks under different levels of data heterogeneity. Specifically, the discrepancy in test accuracy across different data partitions is smallest using our methods, Fed-BEiT and Fed-MAE. The advantage of our method is particularly pronounced when data heterogeneity is severe. In particular, our methods outperform the four baselines when the data distribution is strongly skewed (*i.e.*, Retina and Derm Split-3 and COVID-FL split), with an improvement of 5.06%, 1.53%, and 4.58% in test accuracy on Retina, Derm and COVID-FL, respectively, compared to the supervised baseline with ImageNet pre-training. While our methods outperform all the baselines in severe non-IID scenarios, it is worth noting that they may not always outperform under centralized and IID settings for certain tasks, such as Derm. This is reasonable, as the domain shift between the pre-training dataset ImageNet and the fine-tuning skin images is relatively small in this case. Nonetheless, our method consistently improves model performance under severe data heterogeneity across these three medical tasks.

We further investigate the performance of ImageNet self-supervised pre-training methods, including BEiT ImageNet and MAE ImageNet, which have been shown to outperform their supervised counterparts when fine-tuning with ImageNet in centralized learning [29] [30]. However, we find that these methods are more prone to non-IID data compared to their supervised counterparts and our federated self-supervised learning methods, resulting in a nontrivial decline in performance when the label distribution skewness among clients increases.

We have demonstrated that our proposed method, which conducts federated self-supervised pre-training using only the decentralized target task medical images (with much smaller size compared to other pre-training data such as ImageNet), can achieve comparable results in centralized settings and IID federated settings on most datasets, and outperform all ImageNet pre-training baselines in non-IID federated settings. This shows the potential of our framework to train high-quality federated models in real-world medical applications, where the data distribution across hospitals is typically non-IID.

TABLE IV: Test accuracy for federated fine-tuning on COVID-FL Central (centralized) and Split (non-IID). T represents the communication rounds of federated fine-tuning.

Pre-training			Test Accuracy (%)		T
Method	Setup	Dataset	Central	Split	
None	*None*	*None*	91.42	77.08	1000
Supervised	Centralized	ImageNet	95.35	86.89	100
BEiT [29]	Centralized	ImageNet	95.62	84.45	100
MAE [30]	Centralized	ImageNet	**96.35**	84.32	100
Supervised	Centralized	ChestX-ray14	92.78	82.26	100
BEiT [29]	Centralized	ChestX-ray14	96.07	90.28	100
MAE [30]	Centralized	ChestX-ray14	**96.35**	91.04	100
Fed-BEiT	Distributed	COVID-FL	95.75	87.09	100
Fed-MAE	Distributed	COVID-FL	95.77	**91.47**	100

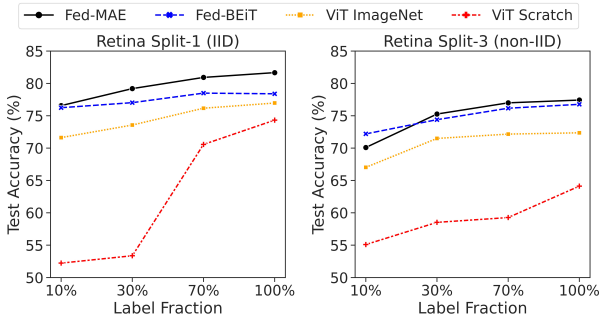


Fig. 9: Test accuracy for Retina under IID and non-IID settings with different fractions of labeled training samples.

Note that for the COVID-FL dataset, we have included three pre-training baselines using a large centralized in-domain dataset, CXR14, as previously mentioned. As shown in Fig. 8c and Table IV, in the non-IID split of COVID-FL, both Fed-BEiT and Fed-MAE outperform the CXR14 supervised pre-training baseline (ViT CXR14). However, the improvement in performance of our proposed method compared to the CXR14 self-supervised pre-training baselines (MAE CXR14 and BEiT CXR14)is not significant. This suggests that, if a large centralized in-domain medical dataset is available, pre-training on it and fine-tuning with downstream target task data may be a good alternative to our proposed method. Nonetheless, it is rare for such datasets to exist for various medical tasks due to privacy and ownership concerns.

2) More label-efficient: We conduct further experiments to evaluate the model performance under limited label scenarios using the Retina dataset. Specifically, we reduce the number of labeled training images by different ratios during federated fine-tuning, taking approximately 70%, 30% and 10% of the labeled samples from each class, resulting in a total of 6000, 3000 and 1000 labeled training data. Fig. 9 shows the test accuracy when fine-tuning with different fractions of labeled

TABLE V: F1-score (%) on Skin-FL dataset.

Method	Backbone	Pre-training Dataset	Non-IID Split	
			Malignant	Benign
FedAvg [23]	ViT-B	*None*	15.4	97.7
FedAvg [23]	EfficientNet	ImageNet	16.1	97.4
FedMatch [60]	EfficientNet	ImageNet	16.0	97.3
FedPerl [61]	EfficientNet	ImageNet	17.8	97.4
FedAvg [23]	ViT-B	ImageNet	23.5	98.1
Fed-BEiT	ViT-B	Skin-FL	**24.2**	98.4
Fed-MAE	ViT-B	Skin-FL	23.6	**98.5**

data. In both IID and non-IID settings, our methods consistently improve the performance compared to the supervised baseline with ImageNet pre-training. The results also show that the model trained from scratch has unsatisfactory test accuracy when the number of labeled images is limited, such as less than 55% when the number of labeled images is 1000.

3) Generalization to out-of-distribution data: One desired property of a well-trained federated model is its ability to generalize to out-of-distribution data. We test the genralization of our proposed methods using Skin-FL and compare them to the baselines stated in [51] and our ViT supervised baselines. Fed-BEiT and Fed-MAE perform slightly better than the supervised baseline with ImageNet pre-training and notably better than all other methods (Table V).

4) Comparison with prior FL methods: In this section, we compare the performance of our proposed methods to previous FL methods in terms of (1) robustness to data heterogeneity and (2) label efficiency.

We first compare the performance of our methods to four federated contrastive learning baselines: FedEMA, FedBYOL, FedMoCo and FedMoCov3. FedEMA [27] is the state-of-the-art federated self-supervised per-training method that has achieved the best performance on CIFAR10. FedBYOL and FedMoCo are two baselines that combine BYOL [45] and MoCo [43] with FedAvg [23]. ResNet-50 is used as the backbone for the above three baselines. We also implement a baseline based on MoCov3 [62] with ViT-B as the backbone, referred to as FedMoCov3. For the training details, the pre-training procedure lasts for 1000 communication rounds for FedBYOL and FedEMA, 300 and 150 rounds for FedMoCov3 and FedMoCo, respectively, until convergence. The AdamW optimizer is used with a batch size of 256 with a base learning rate of 0.03 for FedMoCo, and a batch size of 384 and a base learning rate starting at 5e-4 for the other three baselines.

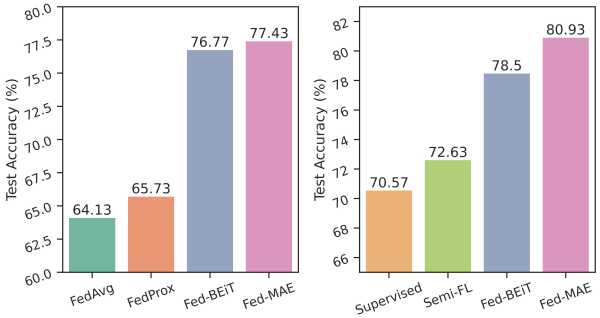
As shown in Table VI, while all of the self-supervised learning methods improve the robustness to data heterogeneity compared to random initialization, our proposed methods outperform all the baselines on the Retina dataset under the heterogeneous data partitions Split-2 and Split-3. Specifically, Fed-MAE and Fed-BEiT surpass the previous state-of-the-art FedEMA by 3.47% and 2.81%, respectively, in test accuracy for the severe non-IID Split-3. In addition, Table VII compares the performance of our methods and the four baselines on the IID Split-1 of the Retina dataset with different fractions of labeled training samples. Our proposed methods outperform all the baselines with limited annotations, with greater improvements in performance when fine-tuning with fewer labeled samples.

TABLE VI: Test accuracy (%) for federated fine-tuning on the Retina dataset using the proposed methods and federated self-supervised pre-training baselines.

Method	Backbone	Central	Split1	Split2	Split3
Rand. init.	ViT-B	73.70	74.33	69.50	64.13
FedMoCov3	ViT-B	79.35	78.06	74.98	72.32
FedMoCo	ResNet-50	77.50	75.80	73.03	70.10
FedBYOL	ResNet-50	80.10	78.43	75.27	72.93
FedEMA [27]	ResNet-50	80.12	78.51	76.08	73.96
Fed-BEiT	ViT-B	79.47	78.40	76.80	76.77
Fed-MAE	ViT-B	**81.93**	**81.67**	**79.40**	**77.43**

TABLE VII: Test accuracy (%) for federated fine-tuning on Retina Split-1 (IID) with different fractions of labeled samples.

Method	Backbone	10%	30%	70%	100%
Rand. init.	ViT-B	52.23	53.37	70.57	74.33
FedMoCov3	ViT-B	72.42	73.34	77.69	78.06
FedMoCo	ResNet-50	65.20	71.73	75.20	75.80
FedBYOL	ResNet-50	72.77	74.27	77.10	78.43
FedEMA [27]	ResNet-50	72.95	74.30	77.02	78.51
Fed-BEiT	ViT-B	76.25	77.03	78.50	78.40
Fed-MAE	ViT-B	**76.57**	**79.20**	**80.93**	**81.67**



(a) under non-IID settings (b) under label-deficient settings

Fig. 10: Comparison of model performance under non-IID and label-deficient settings.

Additionally, we compare our methods to (1) FedProx [10] to evaluate their robustness to non-IID data, and to (2) semi-supervised FL (Semi-FL [23]) to examine their effectiveness with limited labels. ViT-B is used as the backbone for FedProx and Semi-FL.

Fig. 10a compares the test accuracy under the Retina dataset for the severe non-IID partition Split-3. To mitigate the weight divergence caused by data heterogeneity, FedProx [10] adds an L_2 regularization term $\frac{\mu}{2}||w - w_t||^2$ to the local objective function (Eq. 2) during local client updates. We observe that training using FedProx can improve 1.60% of accuracy from the FedAvg baseline after carefully tuning the optimization parameters μ (μ is set to 0.001). However, the gain from using our methods is significantly larger than using FedProx. Specifically, Fed-MAE yields a gain of 13.3% in test accuracy. It is worth noting that the application of self-supervised pre-training in our method is orthogonal to optimization-based FL algorithms such as FedProx. Combining both could potentially further boost the model performance. Fig. 10b shows the test accuracy of different methods when training with 70% labeled data on the IID Split-1 of the Retina dataset. To improve model performance when labeled data is scarce, Semi-FL [23] leverages the unlabeled data in conjunction with supervision from the labeled data. It trains the clients with labeled data in a

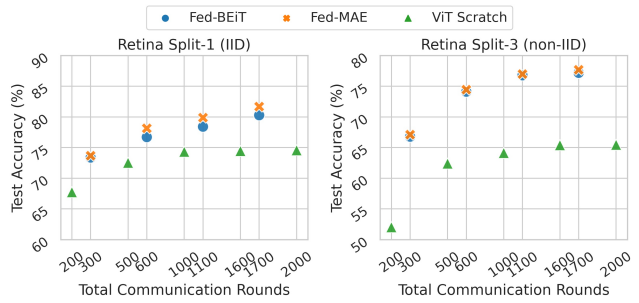


Fig. 11: Ablation study on the number of total communication rounds using Retina Split-1 (IID) and Split-3 (severe non-IID) partitions.

fully supervised manner for 400 epochs and then jointly trains with an extra client holding all the unlabeled data for another 400 epochs. This method was designed for segmentation tasks, and we adapt it to classification tasks. For the unlabeled client, we use a consistency loss function based on data augmentation, calculating the cross-entropy loss between the outputs of the augmented data and the pseudo labels based on the predictions of the original data. According to Fig. 10b, our Fed-MAE outperforms the supervised baseline by 10.36% and the semi-supervised method by 8.3%.

D. Ablation Studies

We perform ablation studies to assess the impact of factors such as communication rounds, training data size, mask ratios, and data augmentations on model performance.

1) Number of communication rounds: Fig. 11 compares the accuracy of our methods (Fed-BEiT and Fed-MAE) and the baseline (ViT Scratch) for different numbers of total communication rounds ($T_{\text{total}} = T_p + T_f$), where T_p and T_f represent the number of communications rounds for pre-training and fine-tuning, respectively. For the proposed methods, we use $T_p \in \{200, 500, 1000, 1600\}$ for pre-training and $T_f = 100$ for fine-tuning, while for the baseline without pre-training, we use $T_p = 0$ and $T_f \in \{200, 500, 1000, 1600, 2000\}$.

On the IID (Split-1) and severe non-IID (Split-3) partitions of the Retina dataset, our methods consistently outperform the baseline. For the baseline without pre-training (represented by the green markers in Fig. 11), the accuracy remains below 75% for Split-1 and 65% for Split-3, even when T_f reaches 2000. By using our proposed pre-training techniques (represented by the blue and orange markers in Fig. 11), the accuracy for the two splits surpasses 75% and 65% with much shorter T_{overall} (less than 600 for Split-1 and less than 300 for Split-3), and continues to increase as T_p increases (reaching around 81% and 77% for Split-1 and Split-3, respectively, when T_p increases to 1600).

Here, we examine the trade-off between model accuracy and communication cost (the number of communication rounds $\times$ the size of the communicated model). During fine-tuning, the size of the communicated model (encoder + linear classifier, 85.8M parameters) is consistent for both our methods and the baseline, while the model communicated during pre-training (encoder + decoder) is larger, with 111.7M parameters for Fed-MAE and 92M parameters for Fed-BEiT, which are $1.3\times$ and $1.11\times$ the size of the model transmitted during fine-tuning. With the size of the communicated model taken into consideration, our proposed methods still outperform the baseline. For example, when T_p is 500 and T_f is 100, our method Fed-BEiT achieves an accuracy of 74.10% on Retina Split-3 with a communication cost of 500×92M+100×85.8M, while the baseline achieves an accuracy of about 63% with the same communication cost at a communication round of 636.

TABLE VIII: Ablation study on the size of the training data.

Number of training images	3000	6000	9000
Fed-BEiT	69.93	74.03	76.77
Fed-MAE	71.70	75.67	77.43
Init. w/ ImageNet weights	71.50	72.17	72.37

2) Training data size: To investigate the effect of the size of the training data on model performance, we reduce the number of total training samples of the Retina Split-3 dataset from 9000 to 6000 and 3000. Table VIII compares the accuracy of our methods and ImageNet pre-training. We find the performance gap between the model pre-trained using our methods and the model initialized with ImageNet weights is reduced when the number of total training images decreases. Therefore, it is applicable to directly use pre-trained ImageNet weights when the total number of training images from all clients combined is limited (*e.g.*, less than 3000).

TABLE IX: Ablation study on mask ratio.

		Mask Ratio				
Dataset	Method	30%	40%	50%	60%	70%
Retina	BEiT	78.53	**79.47**	79.00	78.37	77.60
	MAE	79.73	81.50	81.73	**81.93**	80.90
Derm	BEiT	92.62	**93.27**	92.57	92.04	91.83
	MAE	92.94	93.35	93.60	**93.79**	93.07
COVID-FL	BEiT	95.79	**95.84**	95.67	95.57	95.32
	MAE	**96.55**	96.53	96.23	95.80	95.67

3) Masking ratio: We examine the optimal mask ratio for medical datasets: Retina, Derm and COVID-FL (Table IX). Our results indicate that the optimal mask ratio for BEiT and MAE on Retina and Derm is 40% and 60%, respectively, which is consistent with previous findings on ImageNet [29], [30]. We also observe that the optimal mask ratio for MAE on COVID-FL is 30%. This makes sense because retina and skin images in the Retina and Derm datasets are more similar to natural images than the chest X-ray images in the COVID-FL dataset.

TABLE X: Ablation study on data augmentation.

Data Augmentation	Test Accuracy
random crop + horiz. flip	81.03
random crop + horiz. flip + gray scale + color jitter	**81.73**

4) Data Augmentation: Table X shows the effect of data augmentation on MAE pre-training in centralized settings using the Retina dataset. Adding gray scaling and color jittering improves accuracy by 0.7%, suggesting that task-specific data augmentations may be beneficial for pre-training on medical tasks.

V. Conclusion

In this paper, we propose a privacy-preserving and federated self-supervised learning framework that collaboratively trains models on decentralized data using masked image modeling as the self-supervised task. Our framework is robust to non-IID data distribution across clients, and significantly outperforms state-of-the-art ImageNet supervised pre-training baselines under severe data heterogeneity. It also generalizes well to out-of-distribution data and effectively learns with limited labeled data. Across diverse medical datasets, we show that our proposed method outperforms existing federated self-supervised learning methods, as well as optimization-based and semi-supervised FL methods, under non-IID and label deficient scenarios.

References

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

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

[3] I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, A. Liu, A. B. Costa, B. J. Wood, C.-S. Tsai *et al.*, "Federated learning for predicting clinical outcomes in patients with covid-19," *Nature medicine*, vol. 27, no. 10, pp. 1735–1743, 2021.

[4] W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S. K. Lo, and F.-Y. Wang, "Dynamic-fusion-based federated learning for covid-19 detection," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15 884–15 891, 2021.

[5] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn *et al.*, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence*, vol. 3, no. 6, pp. 473–484, 2021.

[6] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[8] C. He, M. Annavaram, and S. Avestimehr, "Group knowledge transfer: Federated learning of large cnns at the edge," *NeurIPS*, vol. 33, 2020.

[9] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *ICLR*, 2020.

[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, 2020.

[11] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *NeurIPS*, vol. 33, 2020.

[12] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *ICML*, 2020.

[13] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *CVPR*, 2021, pp. 10 713–10 722.

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

[15] L. Qu, Y. Zhou, P. P. Liang, Y. Xia, F. Wang, L. Fei-Fei, E. Adeli, and D. Rubin, "Rethinking architecture design for tackling data heterogeneity in federated learning," in *CVPR*, 2022.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[17] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen *et al.*, "Big self-supervised models advance medical image classification," in *ICCV*, 2021.

[18] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5866–5885, 2021.

[19] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[20] D. Zhang, W. Zeng, J. Yao, and J. Han, "Weakly supervised object detection using proposal-and semantic-level relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[21] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image analysis," *arXiv*, 2022.

[22] X. Li, X. Hu, X. Qi, L. Yu, W. Zhao, P.-A. Heng, and L. Xing, "Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2284–2294, 2021.

[23] D. Yang, Z. Xu, W. Li, A. Myronenko, H. R. Roth, S. Harmon, S. Xu, B. Turkbey, E. Turkbey, X. Wang *et al.*, "Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from china, italy, japan," *Medical image analysis*, vol. 70, p. 101992, 2021.

[24] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *NeurIPS*, vol. 32, 2019.

[25] F. Zhang, K. Kuang, Z. You, T. Shen, J. Xiao, Y. Zhang, C. Wu, Y. Zhuang, and X. Li, "Federated unsupervised representation learning," *arXiv*, 2020.

[26] W. Zhuang, X. Gan, Y. Wen, S. Zhang, and S. Yi, "Collaborative unsupervised visual representation learning from decentralized data," in *ICCV*, 2021, pp. 4912–4921.

[27] W. Zhuang, Y. Wen, and S. Zhang, "Divergence-aware federated self-supervised learning," in *ICLR*, 2021.

[28] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *NeurIPS*, vol. 32, 2019.

[29] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," in *ICLR*, 2021.

[30] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2022.

[31] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-iid data quagmire of decentralized machine learning," in *ICML*, 2020.

[32] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *ICML*, 2021.

[33] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *NeurIPS*, vol. 33, 2020.

[34] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *ICLR*, 2020.

[35] H.-Y. Chen, C.-H. Tu, Z. Li, H.-W. Shen, and W.-L. Chao, "On pre-training for federated learning," *arXiv*, 2022.

[36] J. Nguyen, K. Malik, M. Sanjabi, and M. Rabbat, "Where to begin? exploring the impact of pre-training and initialization in federated learning," *arXiv*, 2022.

[37] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv*, 2020.

[38] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *CVPR*, 2020, pp. 6707–6717.

[39] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[40] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.

[41] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016.

[42] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models genesis," *Medical image analysis*, vol. 67, 2021.

[43] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[45] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *NeurIPS*, vol. 33, pp. 21 271–21 284, 2020.

[46] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu, "Federated contrastive learning for volumetric medical image segmentation," in *MICCAI*. Springer, 2021, pp. 367–377.

[47] N. Dong and I. Voiculescu, "Federated contrastive learning for decentralized unlabeled medical images," in *MICCAI*. Springer, 2021, pp. 378–387.
[48] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021, pp. 8821–8831.
[49] M. d. l. I. Vayá, J. M. Saborit, J. A. Montell *et al.*, "Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients," *arXiv*, Accessed 10 January 10 2022.
[50] D. S. Kermany, M. Goldbaum, and W. Cai *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
[51] T. Bdair, N. Navab, and S. Albarqouni, "Fedperl: Semi-supervised peer learning for skin lesion classification," in *MICCAI*, 2021, pp. 336–346.
[52] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *JBHI*, vol. 23, no. 2, pp. 538–546, 2018.
[53] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.
[54] A. G. Pacheco and R. A. Krohling, "The impact of patient clinical information on automated skin cancer detection," *Computers in biology and medicine*, vol. 116, p. 103545, 2020.
[55] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, pp. 1–8, 2021.
[56] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni, "Bayesian nonparametric federated learning of neural networks," in *ICML*, 2019.
[57] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv*, 2019.
[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[59] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*, 2017, pp. 2097–2106.
[60] W. Jeong, J. Yoon, E. Yang, and S. J. Hwang, "Federated semi-supervised learning with inter-client consistency & disjoint learning," *arXiv*, 2020.
[61] T. Bdair, N. Navab, S. Albarqouni *et al.*, "Semi-supervised federated peer learning for skin lesion classification," *Machine Learning for Biomedical Imaging*, vol. 1, no. April 2022 issue, pp. 1–10, 2022.
[62] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *ICCV*, 2021, pp. 9640–9649.