# Federated CLIP for Resource-Efficient Heterogeneous Medical Image Classification

**Yihang Wu**[1], **Ahmad Chaddad**[1,2*]

[1]AIPM, School of Artificial Intelligence, Guilin University of Electronic Technology, China
[2]The Imaging, Vision and Artificial Intelligence Laboratory, École de Technologie Supérieure, Canada

## Abstract

Despite the remarkable performance of deep models in medical imaging, they still require source data for training, which limits their potential in light of privacy concerns. Federated learning (FL), as a decentralized learning framework that trains a shared model with multiple hospitals (a.k.a., FL clients), provides a feasible solution. However, data heterogeneity and resource costs hinder the deployment of FL models, especially when using vision language models (VLM). To address these challenges, we propose a novel contrastive language-image pre-training (CLIP) based FL approach for medical image classification (FedMedCLIP). Specifically, we introduce a masked feature adaptation module (FAM) as a communication module to reduce the communication load while freezing the CLIP encoders to reduce the computational overhead. Furthermore, we propose a masked multi-layer perceptron (MLP) as a private local classifier to adapt to the client tasks. Moreover, we design an adaptive Kullback-Leibler (KL) divergence-based distillation regularization method to enable mutual learning between FAM and MLP. Finally, we incorporate model compression to transmit the FAM parameters while using ensemble predictions for classification. Extensive experiments on four publicly available medical datasets demonstrate that our model provides feasible performance (e.g., 8% higher compared to second best baseline on ISIC2019) with reasonable resource cost (e.g., 120× faster than FedAVG).

**Code** — https://github.com/AIPMLab/FedMedCLIP

## Introduction

With the development of deep learning (DL) in medical imaging, privacy concerns have hindered collaboration between organizations (Zeng et al. 2024). Federated learning (FL) has emerged as a solution that allows decentralized training without sacrificing patient privacy, advancing medical artificial intelligence (AI) applications while maintaining data security (Zhu et al. 2024; Chaddad, Wu, and Desrosiers 2024; Chaddad et al. 2023). For example, in (McMahan et al. 2017), they propose a simple federated aggregation method called FedAVG, and the experimental results show that it provides feasible performance without accessing raw data from other clients. Similarly to FedAVG, in (Li et al. 2020), they propose using a regularization term to measure the discrepancies between the global and local models, thus improving performance in local clients (FedProx).

However, despite advances in FL, it faces two challenges, 1) heterogeneous data between clients, and 2) communication and computational load during local training and global aggregation (Wu, Desrosiers, and Chaddad 2024). Basically, heterogeneous data (e.g., feature shifts) can lead to severe performance degradation on local clients, while communication and computational costs can prohibit the implementation of FL systems in low-resource devices. This is especially important in the era of vision language models such as contrastive language image pre-training (CLIP), which require large computational and communication loads (e.g., $\sim 10^8$ parameters inside) (Radford et al. 2021). This leads to open question that *"How to adapt the CLIP models effectively in FL context with reasonable cost and feasible generalization performance."*

Recent studies have introduced parameter efficient pretraining (PEFT) techniques to adapt the CLIP into FL frameworks. For example, in FedCLIP (Lu et al. 2023), they propose an adapter as a communication module while inserting it after the CLIP encoders. Furthermore, they freeze the CLIP encoder parameters to save computational overhead. The experimental results on the multi-domain dataset OfficeHome indicate that it achieves feasible performance compared to FedAVG, while reducing resource costs. Similarly, prompt learning based approaches such as promptFL (Guo et al. 2023), FedAPT (Su et al. 2024) have demonstrated remarkable performance for natural image classification tasks in heterogeneous settings. Despite these advancements, the key idea remains the same: *Balancing utility and model size in FL.* However, none of these studies involved experiments with medical data, where heterogeneity is common in medical imaging (e.g., modality). Furthermore, in (Huix et al. 2024), they suggest that CLIP exhibits poor recall rate $\sim 50\%$ for the classification of skin cancer, indicating that there exists a large domain gap between natural and medical.

While vanilla CLIP underperforms in medical datasets, its pre-trained backbone remains a valuable starting point. Motivated by previous challenges, our goal is to build a practical adaptation framework that unlocks this potential for medi-

*Correspondence: ahmad8chaddad@gmail.com

cal use without full re-pretraining. Specifically, we use pretrained CLIP as feature extractors (the encoders are frozen), while introducing a novel masked feature adaptation module (FAM) as communication module. Furthermore, we propose a masked MLP for local tasks, while keeping it local without aggregation. To improve the prediction similarity between FAM and MLP, we incorporate a class-wise KL based distillation approach to minimize the differences between the predicted probabilities of FAM and MLP. To enrich the model predictions, we use ensemble predictions obtained from both FAM and MLP. In addition, we introduce model compression to decrease the communication overhead. Finally, the FAMs are aggregated using a simple average aggregation technique. The contributions of this paper can be summarized as follows.

1. *Algorithm:* We propose a novel CLIP-based FL framework for medical image classification tasks (FedMedCLIP). Specifically, we propose a masked FAM, a masked local classifier and a class-wise KL based regularization technique to improve the performance in heterogeneous data setting with reasonable computational and communication cost. Furthermore, we introduce a model compression technique to compress the model parameters before sending them to the global server.
2. *Empirical analysis:* We perform extensive experiments on four medical datasets (e.g., brain tumor, skin cancer) to show the usefulness of the proposed approach in: (i) robustness to heterogeneity FL, (ii) generalization ability to unseen clients, (iii) efficiency of computation and communication procedure, (iv) adaptability with different network architectures and (v) resilience against adversarial perturbations.

## Related Work

**Federated learning.** Federated learning serves as a fundamental framework for training robust models without sharing raw data. For example, FedProx introduced a regularization term to measure the discrepancies between the global and local models, thereby improving performance in local clients (Li et al. 2020). Similar to FedProx, the differences between the local model of the previous round and the global model are calculated using cosine similarities to optimize the local models are proposed in MOON (Li, He, and Song 2021). FedFocal extended the focal loss in FL framework to sovle the class imbalance challenge existed in medical datasets (Dipankar, Ankur, and Sumit 2020). In addition, FedProto focused on maximizing feature-level consistency between local and global models to improve the overall performance (Tan et al. 2022). SCAFFOLD introduced global gradient calibration and controlled variates to correct local optimization directions (Karimireddy et al. 2020). There are also recent studies devoted to FL designs (Qin et al. 2023; Yang et al. 2024; Yu et al. 2025). However, these methods are not suitable for VLMs such as CLIP because they require a large amount of parameters to be transmitted.

**CLIP.** For example, FACMIC extended FedCLIP to medical classification by introducing domain adaptation technique and a novel adapter as communication module (Wu, Desrosiers, and Chaddad 2024). However, finding a publicly available source domain is a challenge for real-world applications. PromptFL demonstrated the usefulness of sharing prompts instead of models in the FL context (Guo et al. 2023). The experiments show that promptFL yields remarkable performance compared to vanilla methods such as FedAVG. Furthermore, based on promptFL, FedAPT proposed adaptive prompts for local clients (i.e., local prompts with global prompts) for global aggregation (Su et al. 2024), and the experimental results showed that their method provides comparable performance compared to the centralized approach. However, their method introduces considerable computational overhead during inference. Shi et al. proposed client-side knowledge distillation and server-side federated contrastive learning guided by CLIP to address the heterogeneity existing in local clients (Shi et al. 2024). However, they do not validate their model under the feature shift condition (e.g., OfficeHome). In addition, FAA-CLIP adapted the CLIP with an adapter and a domain adversarial classifier in FL (Wu et al. 2025). Experimental results show that it achieves feasible classification performance.

Unlike previous static methods like FACMIC that rely on static aggregation, our approach dynamically learns both shared and client-specific features using a masked FAM, client-specific MLP, and class-wise KL loss to balance consistency. This design jointly enhances personalization and generalization, achieving superior performance.

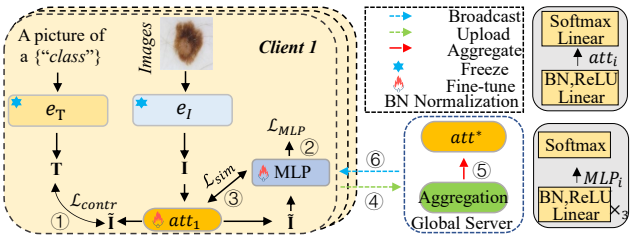


Figure 1: Framework of our approach. The number from ① to ⑥ indicates the corresponding terms of Eq. (5), Eq. (6), Eq. (7), upload, Eq. (13) and download, respectively.

## Methodology

**General framework.** Figure 1 shows the framework of our method. It consists of three parts: 1) local training and inference, 2) model compression and decompression, and 3) global aggregation.

**Local training and inference.** For feature extraction, we use a pretrained CLIP model, comprising an image encoder $e_I(\cdot)$ and a text encoder $e_T(\cdot)$ for each client $C_i$. For a training example $\mathbf{x}_j \in \mathcal{D}_i^{train}$, we denote as $\mathbf{I}_j = e_I(\mathbf{x}_j) \in \mathrm{R}^D$ the $D$-dimensional vector of image features. For text features, we use the standard prompt "`a picture of a {class}`" as input to the text encoder to obtain the features $\mathbf{T}_j = e_T(\mathbf{x}_j) \in \mathrm{R}^D$. To effectively adapt the CLIP model for specific tasks, motivated by (Gao et al. 2024), we introduce a FAM to the CLIP, denoted $att_i(\cdot)$. This FAM takes as input image features $\mathbf{I}$ and returns an attention mask $att_i(\mathbf{I}) \in [0,1]^D$. This mask is then used to generate the

features of the masked images $\tilde{\mathbf{I}} = att_i(\mathbf{I}) \otimes \mathbf{I}$, where $\otimes$ is the Hadamard product (element-wise). After obtaining image and text features, the probability that an example $\mathbf{x}_j$ belongs to a class $c$ can be computed using the cosine similarity $s_{j,c}$ between the image features of $\mathbf{x}_j$ and the text features $\mathbf{T}_c$ corresponding to the prompt of $c$:

$$p(\mathrm{Y}=c \,|\, \mathbf{x}_j) \;=\; \frac{\exp(s_{j,c}/\tau)}{\sum_{c'=1}^{K} \exp(s_{j,c'}/\tau)}, \;\; \text{with } s_{j,c} = \frac{\langle \tilde{\mathbf{I}}_j, \mathbf{T}_c \rangle}{\|\tilde{\mathbf{I}}_j\| \cdot \|\mathbf{T}_c\|} \tag{1}$$

where $\tau$ is the pre-defined softmax temperature parameter.

To reduce the computational load, the CLIP encoders are frozen, and we introduce a masked linear layer to perform transformation. The motivation of using masks is that a masked FAM helps to efficiently learn sparse but dominant feature representations across clients (Kim et al. 2024). Suppose $W$ and $b$ are the weight matrix and bias, respectively. First, we calculate the mean magnitude $u_i$ of each layer:

$$u_i = \frac{1}{n_{in}} \sum_{j=1}^{n_{in}} |W_{i,j}| \tag{2}$$

After obtaining $u_i$, we generate the mask as follows:

$$m_i = \mathcal{S}(u_i - \kappa_i) = \begin{cases} 1, & \text{if } \; \mathrm{u}_i \geq \kappa_i \\ 0, & \text{if } \; \mathrm{u}_i < \kappa_i \end{cases} \tag{3}$$

where $\kappa$ is a learnable threshold, $\mathcal{S}$ is a sign function. Finally, the $W$ are masked using the masks $\mathcal{M}$ and each linear layer can be formulated as follows:

$$\hat{W} = W \odot (\mathcal{M} \cdot \mathbf{1}^{\mathrm{T}}), \; y = \hat{W}x \; + \; (b \odot \mathcal{M}) \tag{4}$$

where $x$ indicates the input.

Specifically, the local FAMs are trained by minimizing a contrastive loss $\mathcal{L}_{contr}$ that pushes the image and text features from the same training example together and pulls apart the non-matching ones. In practice, $\mathcal{L}_{contr}$ is calculated on batches of size $B$. Following (Wu, Desrosiers, and Chaddad 2024), let $\mathbf{S}$ be the $B \times B$ matrix where $s_{j,j'}$ is the cosine similarity between the image features $\tilde{\mathbf{I}}_j$ and $\mathbf{T}_{j'}$ as measured in Eq (1). We compute an image probability matrix $\mathbf{P} = \mathit{Softmax}(\mathbf{S}/\tau) \in [0,1]^{B \times B}$ and a text probability matrix $\mathbf{Q} = \mathit{Softmax}(\mathbf{S}^{\top}/\tau) \in [0,1]^{B \times B}$. The contrastive loss is then formulated as follows:

$$\mathcal{L}_{contr} \;=\; -\frac{1}{B} \sum_{j=1}^{B} \frac{1}{2} \Big( \log p_{j,j} \; + \; \log q_{j,j} \Big). \tag{5}$$

Although a global model can provide robust performance, global aggregation can reduce the client-specific features learned by FAM, which can degrade the performance in clients. Therefore, we propose a MLP as the local classifier for all clients. Unlike traditional MLPs, the proposed MLP replaces the linear layer with a masked linear layer, as defined in Eq. 4. This allows the MLP to focus on task-specific features while ensuring the sparsity of the model structure. Specifically, the local MLP generates raw logits $\mathbf{O}^m \in \mathbb{R}^{B \times C}$, then it goes through a Softmax function to output the probability matrix $\mathbf{P}^m \in [0,1]^{B \times C}$ according to the number of classes $C$ using $\tilde{\mathbf{I}}$. Then, we measure the $\mathcal{L}_{MLP}$ as follows:

$$\mathcal{L}_{MLP} = -\frac{1}{B} \sum_{j=1}^{B} \mathcal{L}_{CE}(p_j, y) \tag{6}$$

If the quality of the clients data is low, it can degrade the performance of the local MLP. Therefore, we use the KL divergence to let the global FAM and the local MLP learn mutually. Specifically, we designed a class-wise KL regularization for local training because applying KL without considering class information can lead to misadaptation due to client heterogeneity. This helps the FAM and MLP learn class-wise information mutually (i.e., enrich the FAM with client-specific features and vice versa) (Cui et al. 2024). Let $\mathbf{P}^v \in [0,1]^{B \times C}$ be the predicted similarities of the CLIP decoder, the loss of consistency regulation ($\mathcal{L}_{sim}$) can be formulated as follows.

$$\mathcal{L}_{sim} = \frac{1}{2C} \left( \sum_{c=1}^{C} \sum_{i=1}^{B} \hat{q}_i^{(c)} \log \frac{\hat{q}_i^{(c)}}{\hat{p}_i^{(c)}} + \sum_{c=1}^{C} \sum_{i=1}^{B} \hat{p}_i^{(c)} \log \frac{\hat{p}_i^{(c)}}{\hat{q}_i^{(c)}} \right) \tag{7}$$

where $\hat{q}_i^{(c)}$ and $\hat{p}_i^{(c)}$ are defined as:

$$\hat{q}_i^{(c)} = \frac{\exp(s_i^{(c)}/T)}{\sum_{b=1}^{B} \exp(s_b^{(c)}/T)}, \hat{p}_i^{(c)} = \frac{\exp(o_i^{(c)}/T)}{\sum_{b=1}^{B} \exp(o_b^{(c)}/T)} \tag{8}$$

where $T$ is a scale parameter. $s$ and $o$ are the logits from FAM and MLP. To more precisely regularize FAM and MLP, we design a dynamic weight parameter $\varpi$ to balance the impact of Eq. 7 as follows:

$$\mathcal{L}_{sim} = \frac{1}{C} \left( \sum_{c=1}^{C} \sum_{i=1}^{B} (\varpi \hat{q}_i^{(c)} \log \frac{\hat{q}_i^{(c)}}{\hat{p}_i^{(c)}} + (1-\varpi) \hat{p}_i^{(c)} \log \frac{\hat{p}_i^{(c)}}{\hat{q}_i^{(c)}}) \right) \tag{9}$$

Here, we propose a training-agnositic $\varpi$ as follows.

$$\varpi = \frac{\mathcal{H}(p^v)}{\mathcal{H}(p^m) + \mathcal{H}(p^v)} \tag{10}$$

where $\mathcal{H}$ is the entropy function. By minimizing $\mathcal{L}_{sim}$, it can help the private model to learn from the global model, and vice versa. This enhances their overall performance and consistency.

Finally, each local model optimizes the following loss function with a hyperparameter $\lambda$:

$$\mathcal{L} = \mathcal{L}_{contr} + \mathcal{L}_{MLP} + \lambda \cdot \mathcal{L}_{sim} \tag{11}$$

During inference, since the global model benefits more from knowledgeable clients while the local classifier fits more about local tasks, we propose to use ensemble predictions to improve the classification ability. Basically, for each sample, the final predicted probability $p^{ens}$ can be measured as:

$$p^{ens} = \varpi \cdot p^{MLP} + (1-\varpi) \cdot p^{FAM} \tag{12}$$

**Model compression and decompression.** We compress the model parameters before sending it to the global server. Specifically, the compression pipeline converts model parameters (i.e., weights) to float16, packs metadata using network byte order, serializes parameters as binary data, and

applies zlib compression (Gailly and Adler 2004) to ensure storage efficiency. When local clients receive the compressed model parameters, decompression is performed to update the local model. Decompression reverses the compression steps by decompressing the zlib data, parsing metadata in big-endian format, converting binary data to tensors, adjusting precision from float16 to float32, and restoring parameters to the model state dictionary.

| Method | Client | | | | | | Global | AVG |
|---|---|---|---|---|---|---|---|---|
| AS | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_{glo}$ | |
| CLIP$_{zs}$ | 31.95 | 23.98 | 24.31 | 17.71 | 20.12 | 33.47 | 17.17 | 24.1 |
| Individual | 74.82 | 52.03 | 74.42 | 67.71 | 69.14 | 59.48 | 17.17 | 59.25 |
| FedAVG | 78.35 | 60.47 | 76.39 | 58.33 | 75.00 | 71.77 | 84.54 | 72.12 |
| LoRA | 72.10 | 59.45 | 76.50 | 55.21 | 71.29 | 72.18 | 82.26 | 69.86 |
| PromptFL | 74.21 | 48.26 | 76.50 | 62.50 | 76.95 | 69.76 | 84.42 | 70.37 |
| CocoOp | 73.94 | 49.13 | 74.77 | 61.46 | 75.59 | 68.35 | 82.63 | 69.41 |
| FedCLIP | 69.54 | 56.98 | 74.88 | 57.29 | 70.12 | 71.37 | 79.91 | 68.58 |
| FACMIC | 69.19 | 56.10 | 74.54 | 60.42 | 67.58 | 66.94 | 74.23 | 67.0 |
| LP++ | 67.87 | 27.18 | 68.17 | 59.38 | 74.22 | 57.66 | 77.52 | 61.71 |
| FedAPT | 77.15 | 51.09 | 73.80 | 67.65 | **81.70** | 68.60 | **85.43** | 72.21 |
| Ours | **84.45** | **71.92** | **82.69** | **84.31** | 79.00 | **79.40** | 81.01 | **80.4** |

Table 1: Accuracy (%) in ISIC2019. **Bold** means the best, while Underline indicates the second best. AVG is the average value of the client accuracy and global accuracy.

**Global aggregation.** We use a simple average aggregation to provide an unbiased aggregated global model. Specifically, the local clients compress their local models and then send them to the global server, while the global server decompresses the models and performs aggregation, then compresses them again to send them back. In each communication round, each client $C_i$ uploads its FAM parameters $v_i^{att}$ to the server. Thereafter, the server combines these parameters into a single vector:

$$v_{global}^{att} = \frac{1}{N}\sum_{i=1}^{N} v_i^{att}. \tag{13}$$

## Experiments

### Datasets

**ISIC2019.** ISIC2019 is a skin cancer classification dataset with various characteristics (e.g., age) (Tschandl, Rosendahl, and Kittler 2018; Codella et al. 2018; Combalia et al. 2019). To simulate data heterogeneous, the dataset was divided based on the "anatomy site (AS)" metadata provided in the original dataset, resulting in seven clients: $C_1$ holds the data whose AS is "anterior torso", while $C_2$, $C_3$, $C_4$, $C_5$, $C_6$ and $C_{glo}$ hold the data whose AS is "head or neck", "lower extremity", "palms or soles", "posterior torso", "upper extremity" and "Nan", respectively.

**Brain tumor.** The BraTS dataset has two classes, namely high-grade glioma and low-grade glioma (Menze et al. 2014; Bakas et al. 2017, 2018; Gong et al. 2024). It has 273 patients for training, while holds 70 patients for testing. Following (Menze et al. 2014), to simulate heterogeneous data, the training set was divided according to the modalities, resulting in four clients: $C_0$ holds the data whose modality is FLAIR, while $C_1$, $C_2$ and $C_3$ hold the data whose modality is T1 weighted, T1 contrast enhancement (T1-CE), and T2 weighted, respectively. Finally, the original test set (with four modalities) is used for global testing.

**Prostate cancer.** Similar to BraTS, this dataset has two classes, namely Muscle-Invasive Bladder Cancer (MIBC) and Non-Muscle-IBC (NMIBC) (Cao et al. 2024). It is a multi-center dataset with T2 weighted modality ($C_1$, patients (n) =160; $C_2$, n=48; $C_3$, n=32; $C_4$, n=35), with a total of 279 patients. The acquisition equipment varies from center to center, i.e. $C_1$ has MAGNETOM Skyra, $C_2$ has UMR 780, $C_3$ has Discovery MR750w 3.0T, and $C_4$ has MAGNETOM Verio.

**ICH.** The RNSA ICH dataset (Flanders et al. 2020), which contains five ICH subtypes, is used for experiments. The same pre-processing strategies as in (Wu et al. 2023) are applied, and images with only a single hemorrhage type are selected. To simulate heterogeneous data, following (Wu et al. 2023), Dirichlet distribution is used to divide the training set to {5,10,15} clients, while the test set is used for global testing. Details can be found in code link.

For each client, we divide the data into training (60%), validation (20%) and test (20%).

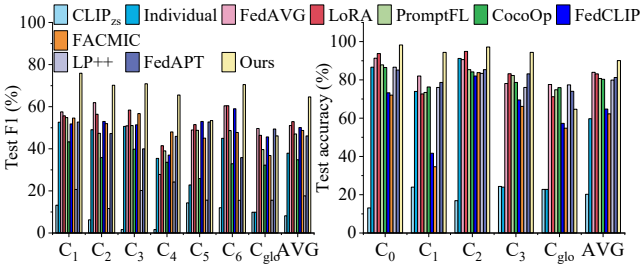


Figure 2: Test metrics on ISIC2019 (**Left**) and BraTS (**Right**) datasets.

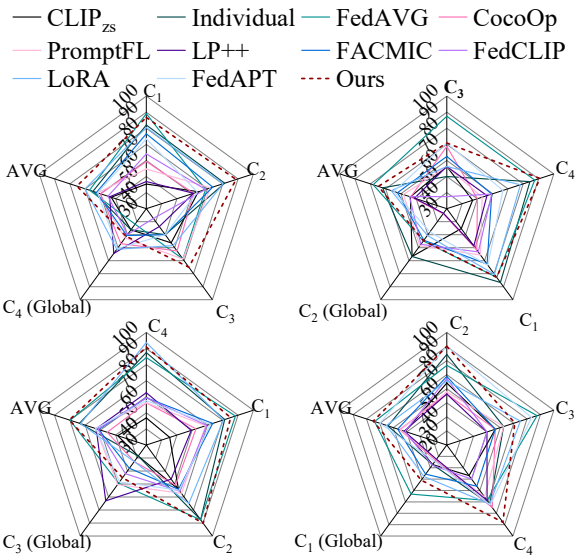


Figure 3: Test accuracy on Prostate dataset.

**Implementation details.** Pretrained ViT-B/32 provided by OpenAI (Radford et al. 2021) is used as the CLIP backbone

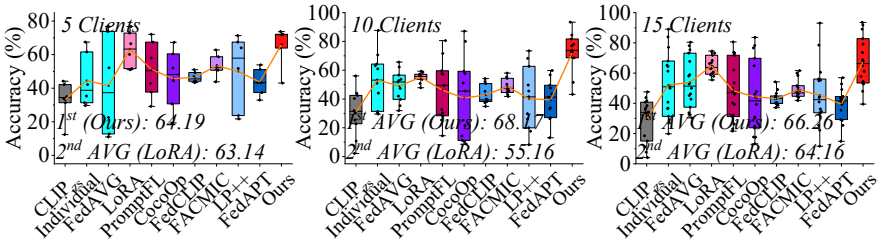


Figure 4: Test accuracy on ICH dataset. Each box in the first row represents the range of the client accuracies (from min to max), while the square connected by orange line represents the average accuracy.

| Alg. | ISIC2019 | | BraTS | | Prostate | | ICH | |
|---|---|---|---|---|---|---|---|---|
| | Computation (min) | | | | Communication (GB) | | | |
| Individual | 35.03 | 3.52 | 11.72 | 75.74 | 0 | 0 | 0 | 0 |
| FedAVG | 95.58 | 28.19 | 98.00 | 84.50 | 7.569 | 3.03 | 10.784 | 7.07 |
| LoRA | 71.49 | 7.201 | 17.80 | 40.20 | 0.259 | 0.052 | 0.279 | 0.174 |
| PromptFL | 50.01 | 2.10 | 2.05 | 61.47 | 0.015 | 0.0013 | 0.0015 | 0.018 |
| CocoOp | 86.83 | 11.10 | 21.80 | 90.87 | 0.017 | 0.0016 | 0.0038 | 0.018 |
| FedCLIP | 52.75 | 3.36 | 5.60 | 59.75 | 0.071 | 0.018 | 0.054 | 0.094 |
| FACMIC | 75.11 | 4.64 | 15.48 | 55.36 | 0.094 | 0.0181 | 0.096 | 0.081 |
| LP++ | 13.85 | 0.47 | 4.56 | 15.84 | 0.0002 | 0.000002 | 0.000012 | 0.000023 |
| FedAPT | 72.495 | 4.23 | 12.01 | 115.30 | 0.037 | 0.0029 | 0.014 | 0.038 |
| Ours | 68.56 | 2.85 | 8.01 | 42.15 | 0.063 | 0.012 | 0.054 | 0.039 |

Table 2: Computation and communication costs for baselines and our approach.

with a batch sise of 32. We use the AdamW optimizer to fine-tune the parameters of both the FAM and the MLP. To provide a stable learning procedure as suggested in (Wu, Desrosiers, and Chaddad 2024), the initial learning rate of the FAM is set to $5 \times 10^{-5}$ for Prostate and BraTS, while $5 \times 10^{-4}$ for ISIC2019. For MLP, it is set to $1 \times 10^{-3}$ (ISIC2019) and $1 \times 10^{-4}$ (Prostate and BraTS). The $\lambda$ and $T$ are set to 0.04 and 2 for all datasets, respectively. The optimization process uses a weight decay of $0.02$ and beta parameters of $(0.99, 0.98)$. The exponential learning rate scheduler is used with a gamma of 0.97 for local training. We set the communication round to 100 (ISIC2019) and 50 (BraTS, Prostate and ICH), while the local training epoch is set to one. We select one vanilla FL technique (FedAVG), and seven PEFT based federated approaches, namely FedCLIP, PromptFL, CocoOp with FedAVG (Zhou et al. 2022), LP++ with FedAVG (Huang et al. 2024), LoRA with FedAVG (rank is set to three) (Zanella and Ben Ayed 2024), FedAPT and FACMIC. For CocoOp, only the parameters of the prompt learner are aggregated, whereas for LP++, only the linear probes parameters are aggregated. For non-federated methods, the Individual without aggregation and $\text{CLIP}_{zs}$ are selected. The random seed was set to 0 to eliminate the impact of different seeds. All experiments are based on the Windows 11 operating system, and feature an Intel 13900KF CPU with 128 GB of RAM and an RTX 4090 GPU. The FAM has $\sim 5 \times 10^5$ parameters.

**Results.** Table 1, Figure 2, Figure 3, and Figure 4 show the test accuracy (ACC) and F1 score for the ISIC2019, BraTS, Prostate and ICH datasets. We highlight the following points: 1) For all medical datasets, the original pre-trained CLIP shows poor generalization ability (e.g., $24.1\%$ AVG test ACC on ISIC2019), consistent with the results described in (Huix et al. 2024). This suggests that the original CLIP has limited domain knowledge. 2) The use of FL techniques can improve the performance of CLIP on medical datasets (e.g., using FedAVG provides 72.12% AVG ACC on ISIC2019). However, they still face challenges such as class imbalance, as it shows a lower F1 score (e.g. 51. 03% AVG on ISIC2019), indicating overfitting in unbalanced data. Furthermore, for the clients with sufficient data (e.g., $C_1$ in Prostate), the learned knowledge can improve the performance of FAM and MLP in other clients (e.g., $86.72\%$ and $88.73\%$ ACC in $C_2$ and $C_3$, client:$\{C_1$,$C_2$,$C_3\}$, global: $C_4$). In addition, for the clients with limited samples (e.g., $C_3$ in Prostate), the usefulness of FAM is weak and it will negatively degrade the feature representations, thereby reducing the performance of local clients (e.g., $37.5\%$ ACC on $C_3$ with FedCLIP). However, the proposed approach improves performance on $C_3$ to 75% by training a FAM with a local MLP mutually. A similar situation can be found for $C_4$ in the following setting: client:$\{C_2$,$C_3$,$C_4\}$, global: $C_1$. 3) Considering the average performance, the proposed approach yields the highest AVG metrics (e.g., 80.4% ACC and 64.65% F1 score on ISIC2019, 90.07% ACC on BraTS) compared to other baselines. Futhermore, FedAVG shows large differences on ICH dataset (e.g., larger box size in Figure 4), while PEFT methods such as FedCLIP provide smaller discrepancies. However, these approaches show considerable performance degradation as the number of clients increases (e.g., $\sim 5\%$ ACC drop with FACMIC from 5 to 10 clients). Unlike others, our method provides stable test metrics despite client settings (e.g., consistently higher than 66% AVG), suggesting that the introduction of masked MLP and KL based regularizations is a robust solution for a large-scale client setting. 4) Regarding the

| Components | | | | Clients | | | | | | Global | AVG |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{contr}$ | $\mathcal{L}_{MLP}$ | $\mathcal{L}_{sim}$ | Aggregation | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_{glo}$ | |
| ✓ | ✘ | ✘ | ✓ | 74.28 | 63.53 | 78.36 | 58.82 | 72.64 | 74.80 | **81.13** | 71.94 |
| ✓ | ✓ | ✘ | ✓ | 84.36 | 69.03 | **84.40** | 75.49 | 77.65 | 77.60 | 80.52 | 78.44 |
| ✘ | ✓ | ✘ | ✘ | 81.23 | 64.40 | 80.98 | 81.37 | 76.69 | 72.60 | 17.17 | 67.77 |
| ✓ | ✓ | ✓ | ✘ | 81.67 | 69.75 | 82.80 | **84.31** | **80.35** | 76.80 | 17.17 | 70.41 |
| ✓ | ✓ | ✓ | ✓ | **84.45** | **71.92** | 82.69 | **84.31** | 79.00 | **79.40** | 81.01 | **80.4** |

Table 3: Ablations on each component in our approach. **Bold** represents the best.

global generalization ability, the proposed FAM yields feasible ACC on ISIC2019 (81.01% ACC), while providing lower ACC on BraTS (64.66%) compared to other PEFT methods such as FedAPT (74.04%). This suggests that the potential of FAM is limited where the feature shifts are considerablely large (e.g., different modalities). We argue that the pre-trained CLIP encoders has limited prior knowledge, thus simply using a FAM on the global site leads to ∼4% lower ACC on ISIC2019 dataset compared to prompt-based methods such as FedAPT, while for local clients, benefited by the MLP, it learns robust feature patterns, thus providing the best test metric.

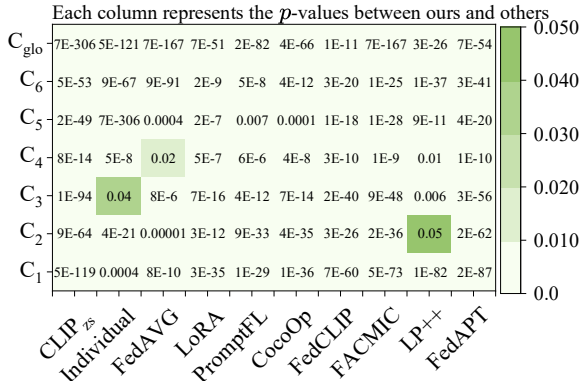


Figure 5: The $p$-value between our model and the other baselines on the ISIC 2019 dataset using Wilcoxon signed-rank test. These values are based on the test set.

**Computation and communication overhead.** Table 2 reports the total computation time and communication cost for the baselines and the proposed approach to achieve the best validation metrics. For Prostate, only one client setting is considered (client:{$C_1$,$C_2$,$C_3$}, global: {$C_4$}), while for ICH, the results are based on 10 clients. As illustrated, FedAVG has a large computational and communication overhead (e.g., 95.58 min, 7.569 GB on ISIC2019), while the proposed method provides a feasible resource cost (e.g., 120× faster than FedAVG). Furthermore, our method yields comparable computational and communication load compared to baselines such as FedAPT on ISIC2019 (68.56 min, 0.063 GB vs. 72.495 min, 0.037 GB), and FedCLIP on BraTS (2.85 min, 0.012 GB vs. 3.36 min, 0.018 GB). These results highlight the potential of the proposed designs for resource-efficient FL frameworks.

**Ablations on components.** We validate the usefulness of each component in the proposed approach in the ISIC2019 data set. As illustrated in Table 3, introducing $\mathcal{L}_{MLP}$ improves the overall performance in local clients (e.g., ∼ 7% ACC improvement), while adding $\mathcal{L}_{sim}$ provides 1.32% average ACC improvement compared to without $\mathcal{L}_{sim}$. We note that for certain clients, such as $C_1$, the use of $\mathcal{L}_{sim}$ slightly reduces the performance (e.g., 1.21%). This suggests that FAM can reduce the learned knowledge of the local MLP in class imbalance situation. Furthermore, the usefulness of $L_{sim}$ on the global site performance is limited since it lacks a private local MLP (i.e., it can not provide ensemble prediction). Overall, the use of all these components leads to the best AVG (80.4%).

**Statistical Significance Analysis.** Figure 5 shows the $p$-value (measured with test set, significance level equals 0.05) between ours and other baselines on the ISIC2019 dataset using Wilcoxon signed-rank test (Rey and Neuhäuser 2011). As illustrated, the proposed method demonstrates significant performance improvements over other approaches (e.g., $p < 10^{-11}$ on $C_2$ compared to LoRA).

**Influence of $T$.** We explored the performance on the ISIC2019 dataset with different $T$ values used in Eq. 8. As illustrated in Figure 8, smaller $T$ leads to a decrease in performance (e.g., an ACC of 78.03% on $C_5$), while larger $T$ such as 10 results in a lower AVG of 78.94%.

**Impact of model compression.** We explored the impact of performance losses due to float 16 conversion during model compression on ISIC2019 dataset. As shown in Figure 8, the AVG is 80.48% without compression, slightly higher than with compression. However, with compression, the model size is reduced from 2.01 (FAM itself) to 1.36MB. These results suggest that compressing the weights of the model does not considerably impact its performance.

**Sensitivity analysis.** We performed experiments using different $\lambda$ values on the ISIC2019 dataset to analyze its sensitivity. Figure 6 shows the test ACC with various $\lambda$ values ($\lambda \in [0.01, 0.1]$). As illustrated, $\lambda$ equal to 0.04 leads to the highest overall ACC (AVG=80.4) compared to other settings.

**Robustness.** We validate the robustness to gradient based adversarial attacks. Specifically, we consider the fast gradient sign method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) and projected gradient descent (PGD) (Madry et al. 2018) for experiments on the ISIC2019 dataset. The magnitude of adversarial attacks is set to 0.1 for FGSM and PGD. As illustrated in Figure 6, PEFT approaches such as FedCLIP and LoRA exhibit a considerable performance decrease (e.g. ∼ 10% AVG under FGSM attack), while the proposed method shows a higher AVG of 33.85%, indicating its robustness to adversarial attacks compared to PEFT

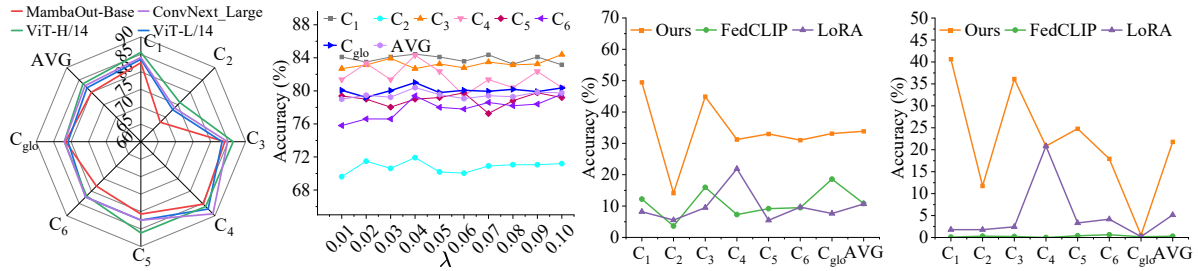


Figure 6: Test accuracy on ISIC2019 dataset with varying backbones (**Left**), $\lambda$ values (**2-column**), FGSM (**3-coumn**) and PGD (**Right**).

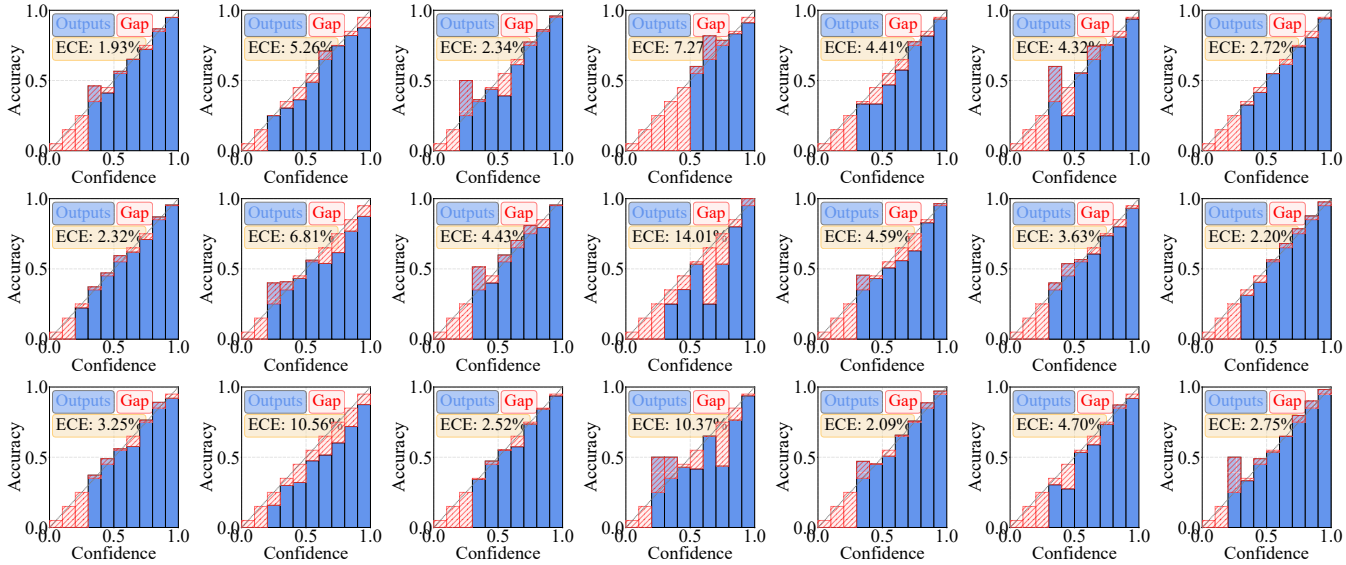


Figure 7: Reliability plot for the proposed approach (**first row**) and baselines (LoRA, **second row**; PromptFL, **third row**) on ISIC2019 dataset. The clients and global are represented from the left column to the right column.

approaches. Similar situation can be found with PGD attack (e.g., the proposed method provides $\sim 25\%$ higher ACC on $C_1$ compared to LoRA).

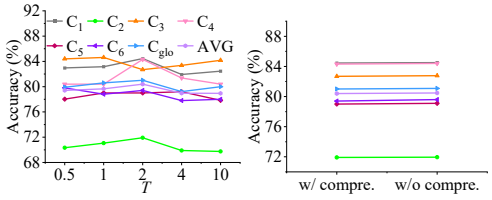


Figure 8: Test accuracy of the proposed method on ISIC2019 dataset with different $T$ values (**Left**) and model compression (**Right**).

**Backbones.** We changed the network backbones with deeper/different architectures, such as ViT-L/14, ViT-H/14, ConvNext_Large provided in OpenCLIP (Cherti et al. 2023; Ilharco et al. 2021) and MambaOut-Base (Yu and Wang 2025) to validate its adaptability. Figure 6 shows the test ACC with these backbones. The proposed approach exhibits remarkable adaptability with deeper networks such as ViT-L/14 (81.83% AVG) and MambaOut-Base (80.2% AVG).

**Calibration.** Figure 7 shows the reliability diagrams (Vaicenavicius et al. 2019) on the ISIC2019 dataset for the proposed method and baselines. As illustrated, the proposed approach demonstrates a lower ECE value on local clients (e.g., 1.93% on $C_1$, 2.34% on $C_3$ compared to LoRA (2.32% on $C_1$ and 4.43% on $C_3$) and PromptFL (3.25% on $C_1$ and 2.52% on $C_3$). In addition, our method achieves comparable ECE on global site (2.72%) compared to LoRA (2.20%) and PromptFL (2.75%).

## Conclusion

In this study, we explored the potential of VLMs for medical imaging in FL and proposed a novel CLIP-based FL framework. We introduced a masked FAM as the communication module while freezing the CLIP encoders to reduce computational and communication overhead, while using masked MLPs to adapt the local client with class-wise KL. Finally, ensemble predictions are obtained to improve local performance. Experimental results in skin-, brain- and prostate-related classification tasks demonstrate the remarkable performance of our approach compared to SOTA methods.

## Acknowledgments

This research was funded by the National Natural Science Foundation of China grant number 82260360, the Innovation Project of GUET Graduate Education 2025YCXS244 and the Guangxi Science and Technology Base and Talent Project (2022AC18004, 2022AC21040).

## References

Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J. S.; Freymann, J. B.; Farahani, K.; and Davatzikos, C. 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 1–13.

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; et al. 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*.

Cao, K.; Zou, Y.; Zhang, C.; Zhang, W.; Zhang, J.; Wang, G.; Zhang, C.; Lyu, J.; Sun, Y.; Zhang, H.; et al. 2024. A multicenter bladder cancer MRI dataset and baseline evaluation of federated learning in clinical application. *Scientific Data*, 11(1): 1147.

Chaddad, A.; Lu, Q.; Li, J.; Katib, Y.; Kateb, R.; Tanougast, C.; Bouridane, A.; and Abdulkadir, A. 2023. Explainable, domain-adaptive, and federated artificial intelligence in medicine. *IEEE/CAA Journal of Automatica Sinica*, 10(4): 859–876.

Chaddad, A.; Wu, Y.; and Desrosiers, C. 2024. Federated learning for healthcare applications. *IEEE internet of things journal*, 11(5): 7339–7358.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Codella, N. C.; Gutman, D.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, 168–172. IEEE.

Combalia, M.; Codella, N. C.; Rotemberg, V.; Helba, B.; Vilaplana, V.; Reiter, O.; Carrera, C.; Barreiro, A.; Halpern, A. C.; Puig, S.; et al. 2019. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

Cui, J.; Tian, Z.; Zhong, Z.; Qi, X.; Yu, B.; and Zhang, H. 2024. Decoupled kullback-leibler divergence loss. *Advances in Neural Information Processing Systems*, 37: 74461–74486.

Dipankar, S.; Ankur, N.; and Sumit, R. 2020. Fed-Focal Loss for imbalanced data classification in Federated Learning. In *IJCAI*.

Flanders, A. E.; Prevedello, L. M.; Shih, G.; Halabi, S. S.; Kalpathy-Cramer, J.; Ball, R.; Mongan, J. T.; Stein, A.; Kitamura, F. C.; Lungren, M. P.; et al. 2020. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3): e190211.

Gailly, J.-l.; and Adler, M. 2004. Zlib compression library.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Gong, Z.; Xu, T.; Peng, N.; Cheng, X.; Niu, C.; Wiestler, B.; Hong, F.; and Li, H. B. 2024. A Multi-Center, Multi-Parametric MRI Dataset of Primary and Secondary Brain Tumors. *Scientific Data*, 11(1): 789.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guo, T.; Guo, S.; Wang, J.; Tang, X.; and Xu, W. 2023. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*.

Huang, Y.; Shakeri, F.; Dolz, J.; Boudiaf, M.; Bahig, H.; and Ben Ayed, I. 2024. LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23773–23782.

Huix, J. P.; Ganeshan, A. R.; Haslum, J. F.; Söderberg, M.; Matsoukas, C.; and Smith, K. 2024. Are Natural Domain Foundation Models Useful for Medical Image Classification? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7634–7643.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP. If you use this software, please cite it as below.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, 5132–5143. PMLR.

Kim, M.; Saad, W.; Debbah, M.; and Hong, C. S. 2024. SpaFL: Communication-efficient federated learning with sparse models and low computational overhead. *Advances in Neural Information Processing Systems*, 37: 86500–86527.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Lu, W.; Hu, X.; Wang, J.; and Xie, X. 2023. FedCLIP: Fast Generalization And Personalization For CLIP in Federated Learning. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018*, 1–23.

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

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10): 1993–2024.

Qin, Z.; Deng, S.; Zhao, M.; and Yan, X. 2023. FedAPEN: personalized cross-silo federated learning with adaptability to statistical heterogeneity. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1954–1964.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rey, D.; and Neuhäuser, M. 2011. Wilcoxon-signed-rank test. In *International encyclopedia of statistical science*, 1658–1659. Springer.

Shi, J.; Zheng, S.; Yin, X.; Lu, Y.; Xie, Y.; and Qu, Y. 2024. CLIP-Guided Federated Learning on Heterogeneity and Long-Tailed Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14955–14963.

Su, S.; Yang, M.; Li, B.; and Xue, X. 2024. Federated adaptive prompt tuning for multi-domain collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15117–15125.

Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8432–8440.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Vaicenavicius, J.; Widmann, D.; Andersson, C.; Lindsten, F.; Roll, J.; and Schön, T. 2019. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*, 3459–3467. PMLR.

Wu, N.; Yu, L.; Yang, X.; Cheng, K.-T.; and Yan, Z. 2023. FedIIC: Towards robust federated learning for class-imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 692–702. Springer.

Wu, Y.; Chaddad, A.; Desrosiers, C.; Daqqaq, T.; and Kateb, R. 2025. FAA-CLIP: Federated Adversarial Adaptation of CLIP. *IEEE Internet of Things Journal*, 12(12): 21091–21102.

Wu, Y.; Desrosiers, C.; and Chaddad, A. 2024. FACMIC: Federated Adaptative CLIP Model for Medical Image Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 531–541. Springer.

Yang, Z.; Zhang, Y.; Zheng, Y.; Tian, X.; Peng, H.; Liu, T.; and Han, B. 2024. FedFed: Feature distillation against data heterogeneity in federated learning. *Advances in Neural Information Processing Systems*, 36.

Yu, W.; Chen, S.; Tong, Y.; Gu, T.; and Gong, C. 2025. Modeling inter-intra heterogeneity for graph federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22236–22244.

Yu, W.; and Wang, X. 2025. Mambaout: Do we really need mamba for vision? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4484–4496.

Zanella, M.; and Ben Ayed, I. 2024. Low-Rank Few-Shot Adaptation of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1593–1603.

Zeng, S.; Guo, P.; Wang, S.; Wang, J.; Zhou, Y.; and Qu, L. 2024. Tackling data heterogeneity in federated learning via loss decomposition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 707–717. Springer.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhu, M.; Yang, Q.; Gao, Z.; Liu, J.; and Yuan, Y. 2024. Stealing Knowledge from Pre-trained Language Models for Federated Classifier Debiasing. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 685–695. Springer.