FACMIC: Federated Adaptative CLIP Model for Medical Image Classification

Yihang Wu[1], Christian Desrosiers[2], and Ahmad Chaddad[1,2]

[1] Laboratory for Artificial Intelligence for Personalised Medicine, School of Artificial Intelligence, Guilin University of Electronic Technology, China
[2] Laboratory for Imagery Vision and Artificial Intelligence, ÉTS Montreal, Canada
ahmad8chaddad@gmail.com

Abstract. Federated learning (FL) has emerged as a promising approach to medical image analysis that allows deep model training using decentralized data while ensuring data privacy. However, in the field of FL, communication cost plays a critical role in evaluating the performance of the model. Thus, transferring vision foundation models can be particularly challenging due to the significant resource costs involved. In this paper, we introduce a federated adaptive Contrastive Language Image Pretraining (CLIP) model designed for classification tasks. We employ a light-weight and efficient feature attention module for CLIP that selects suitable features for each client's data. Additionally, we propose a domain adaptation technique to reduce differences in data distribution between clients. Experimental results on four publicly available datasets demonstrate the superior performance of FACMIC in dealing with real-world and multisource medical imaging data. Our codes are available at https://github.com/AIPMLab/FACMIC.

Keywords: Federated learning · Domain adaptation · Medical imaging.

1 Introduction

The success of deep learning (DL) highly depends on the availability of large amounts of data for training. However, there has been a growing focus on data privacy and security in recent years, with some organizations implementing regulations and laws such as the EU General Data Protection Regulation (GDPR) [1]. Collecting raw data is often impractical in this case, which poses a challenge to the feasibility of centralized DL approaches. Federated learning (FL) has emerged as a new distributed learning method to deal with this challenge and has been widely adopted in various applications [2]. FL enables model aggregation without directly accessing the raw user data from different clients. As pioneering work in FL, FedAVG efficiently combines distributed information through a simple but effective averaging algorithm [3]. This method ensures that raw data remain on the local client, preserving data privacy and security.

The performance of FL models is typically influenced by two main factors: data distribution shifts and communication costs [4]. Shifts in data distribution can negatively impact model prediction accuracy, particularly in the medical

field, where variations in imaging equipment can introduce discrepancies. In most FL solutions, communication costs are proportional to the number of model parameters to transmit. This impedes the use of large Vision Language Models (VLMs) such as CLIP that contain more than 10^8 parameters (ViT-B/32) [5].

Work on FL with foundation models has started recently. In [6], the authors proposed sharing the prompts instead of models to reduce communication costs. An important limitation of this approach is that it does not address the issue of heterogeneity in the data distribution. In [7], a specialized model compression technique is introduced to reduce transmission costs. However, this technique still has substantial computational costs. Furthermore, the results in [8] show that foundation models like CLIP or BLIP fail for medical image classification tasks (e.g., 52.7% classification recall for CLIP using the ISIC2019 dataset). This leads us to consider how federated foundation models can be enhanced in terms of effectiveness and efficiency. In this paper, we propose the Federated adaptive CLIP model for the medical image classification (FACMIC) task. FACMIC adds a light-weight feature-attention module on top of CLIP to help the model focus on useful features in the data for each client. It also introduces a domain adaptation (DA) strategy to minimize data discrepancies between clients. Our model can quickly converge and achieve improved performance through adaptation, which greatly reduces training time. The contributions of our work are summarized as follows.

- We propose a novel federated adaptive CLIP model that combines feature attention with an adaptation technique to address both problems of communication costs and distribution discrepancy. Although past studies have applied FL in medical imaging, to our knowledge, we are the first to explore FL on VLMs in this context.
- We test the performance of our method in both real-world and simulated multi-source cases, and show its superior performance compared to state-of-the-art approaches for brain tumor and skin cancer classification tasks.

2 Material and methods

2.1 Problem formulation

In our federated learning setting, we have a set of N clients $\{C_1, C_2, \cdots, C_N\}$, each client C_i having a private data set $\mathcal{D}_i = \{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^{n_i}$. As in similar studies [9], we assume that the data of separate clients have the same input and output space, but follow different distributions, i.e., $P(\mathcal{D}_{i'}) \neq P(\mathcal{D}_i)$, $\forall i' \neq i$. Each dataset $\mathcal{D}_i$ consists of three non-overlapping parts, namely a training set $\mathcal{D}_i^{train}$, a validation set $\mathcal{D}_i^{val}$ and a test set $\mathcal{D}_i^{test}$. Our goal is to train a robust global model $f_\theta(\cdot)$ while preserving data privacy and security. This model should provide good testing performance on the test data of every client, i.e.,

$$\min_f \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i^{test}} \sum_{j=1}^{n_i^{test}} \ell(f_\theta(\mathbf{x}_{i,j}^{test}), y_{i,j}^{test}), \tag{1}$$

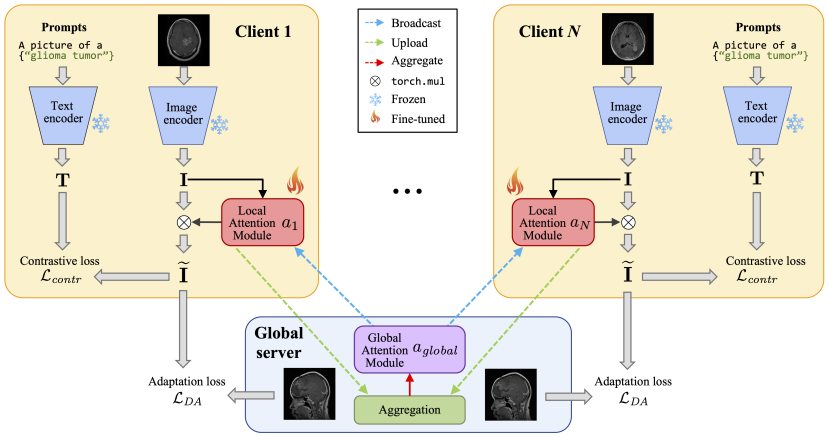

Fig. 1: The proposed FACMIC framework. Each client trains its model separately, optimizing only the parameters of its local attention module (a_i) using contrastive and domain adaptation losses. After receiving the local client parameters, the server aggregates them into a global attention module (a_{global}) whose parameters are broadcasted back to clients.

based on a given loss function ℓ. For generalization, we assume that there exist Q different clients $\{M_1, M_2, \cdots, M_Q\}$ with data $\mathcal{D}_i^M = \{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^{n_i}$, n_i being the number of samples in each client. Our goal is for the model to achieve a good performance on clients that *were excluded* from the local training stage, i.e.,

$$\min_f \frac{1}{M}\sum_{i=1}^{M}\frac{1}{m_i}\sum_{j=1}^{m_i}\ell(f_\theta(\mathbf{x}_{i,j}), y_{i,j}). \tag{2}$$

2.2 Our federated adaptive CLIP framework

Our federated learning framework for CLIP-based medical image classification is illustrated in Figure 1. This framework comprises three key components: a feature attention module for efficient FL, a feature adaptation strategy that addresses the problem of data distribution shifts across different clients, and a global aggregation strategy to combine the learned information from multiple clients. We present these components in what follows.

Training the attention module. We use a pretrained CLIP model comprising an image encoder $e_I(\cdot)$ and a text encoder $e_T(\cdot)$, to extract features from the data for each client C_i. For a training example $\mathbf{x}_j \in \mathcal{D}_i^{train}$, we denote as $\mathbf{I}_j = e_I(\mathbf{x}_j) \in \mathbb{R}^D$ 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 features $\mathbf{T}_j = e_T(\mathbf{x}_j) \in \mathbb{R}^D$.

Pretrained foundation models hold the ability to extract a rich set of features, however, not all of those are suitable for learning a specific task. This is particularly true for detecting and classifying abnormal regions such as lesions in medical images, as these regions are absent in normal images and typically

represent a small part of the image. To identify the regions of focus for locally-trained models, we introduce a client feature attention module, denoted as $a_i(\cdot)$. This attention module takes as input image features $\mathbf{I}$ and returns an attention mask $a_i(\mathbf{I}) \in [0,1]^D$. This mask is then used to generate masked images features $\widetilde{\mathbf{I}} = a_i(\mathbf{I}) \otimes \mathbf{I}$, where $\otimes$ is the Hadamard (element-wise) product.

We measure the probability that an example $\mathbf{x}_j$ belongs to a class c using the cosine similarity between the image features of $\mathbf{x}_j$ and the text features $\mathbf{T}_c$ corresponding to the prompt of c:

$$p(\mathrm{Y}{=}c \,|\, \mathbf{x}_j) \;=\; \frac{\exp(s_{j,c}/\tau)}{\sum_{c'=1}^{K} \exp(s_{j,c'}/\tau)}, \quad \text{with } s_{j,c} = \frac{\langle \widetilde{\mathbf{I}}_j, \mathbf{T}_c \rangle}{\|\widetilde{\mathbf{I}}_j\| \cdot \|\mathbf{T}_c\|} \tag{3}$$

where τ is the softmax temperature parameter.

Keeping the image and text encoders frozen, we train the local adapter modules by minimizing a contrastive loss $\mathcal{L}_{contr}$ that pushes together the image and text features from the same training example and pulls apart non-matching ones. Following [10], we compute the contrastive loss over batches of size B. Let $\mathbf{S}$ be the $B \times B$ matrix where $s_{j,j'}$ is the cosine similarity between image features $\widetilde{\mathbf{I}}_j$ and $\mathbf{T}_{j'}$, as measured in Eq (3). 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:

$$\mathcal{L}_{contr} \;=\; -\frac{1}{B} \sum_{j=1}^{B} \frac{1}{2} \Big(\log p_{j,j} \;+\; \log q_{j,j} \Big). \tag{4}$$

Local feature adaptation. Discrepencies in the local data distribution of clients may affect the training of the global model via federated learning. To address this problem, we propose a client-specific feature adaptation strategy based on the Local Maximum Mean Discrepancy (LMMD) method [11]. This domain adaptation method aligns the class-wise distribution statistics of the data from a source domain $\mathcal{D}_s$ and a target domain $\mathcal{D}_t$, by minimizing the following loss:

$$\mathcal{L}_{DA} = \frac{1}{K} \sum_{c=1}^{K} \Big\| \sum_{\mathbf{x}_i^s \in \mathcal{D}_s} \omega_{i,c}^s \, \phi(\mathbf{x}_i^s) \;- \sum_{\mathbf{x}_j^t \in \mathcal{D}_t} \omega_{j,c}^t \, \phi(\mathbf{x}_j^t) \Big\|_{\mathcal{H}}^2, \;\; \omega_{i,c}^{s|t} = \frac{\mathbb{1}(y_i = c)}{\sum_{y_j \in \mathcal{D}^{s|t}} \mathbb{1}(y_j = c)}. \tag{5}$$

Here, $\phi(\cdot)$ is a mapping function to a Hilbert space $\mathcal{H}$, while $\omega_{i,c}^s$, $\omega_{j,c}^t$ are weights that measure the membership of example $\mathbf{x}_i^s$ and $\mathbf{x}_j^t$ in class c.

In our setting, we suppose that the shift occurs only on the image features, thus $\mathbf{x}{=}\mathbf{I}$ in Eq. (5). Furthermore, since we cannot compute $\phi(\cdot)$ directly (as it maps features to a high-dimensional space), we use the kernel trick and refor-

mulate the loss as

$$\mathcal{L}_{DA} = \frac{1}{K}\sum_{c=1}^{K}\Big[\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\omega_{i,c}^{s}\,\omega_{j,c}^{s}\,k(\mathbf{I}_i^s,\mathbf{I}_j^s) \;+\; \sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\omega_{i,c}^{t}\,\omega_{j,c}^{t}\,k(\mathbf{I}_i^t,\mathbf{I}_j^t) - 2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\omega_{i,c}^{s}\,\omega_{j,c}^{t}\,k(\mathbf{I}_i^s,\mathbf{I}_j^t)\Big] \tag{6}$$

where n_s, n_t are the number of samples in the source domain and target domain, respectively, and $k(\cdot,\cdot) = \langle\phi(\cdot),\phi(\cdot)\rangle$ is the kernel function.

Although we defined our domain adaptation loss, we still need to specify the source and target domains used in this formulation. When performing a local adaption for client C_i, we set the source domain to be the data of this client, i.e., $\mathcal{D}_s = \mathcal{D}_i$. However, to preserve privacy, we cannot exchange data directly between clients, therefore we cannot use the data of other clients for the target domain. Instead, we use a global set of unlabeled images as the target domain data. This constraint can be easily satisfied since there are many publicly available datasets in medical imaging and no labels are needed for this reference data.

To compute the weights $\omega_{i,c}^{s|t}$ in Eq. (5), we employ a pseudo-label strategy to estimate the class labels. Specifically, we obtain the class probabilities of a target image $\mathbf{I}_j^t$ using Eq. (3), and assign this image to the class with the highest probability. Our final loss to update the feature attention parameters of each client is the combination of the contrastive loss and domain adaptation loss,

$$\mathcal{L} = \mathcal{L}_{contr} \;+\; \lambda\mathcal{L}_{DA}, \tag{7}$$

where λ is a hyper-parameter controlling the trade-off between these two loss terms.

Global aggregation. The last part of our proposed FACMIC framework is the aggregation strategy to combine the parameters of different clients into a single global model. This strategy works as follows. In each round, each client C_i uploads its attention module parameters θ_i^a to the server. Thereafter, the server combines these parameters into a single vector θ_{global}^a using a weighted average

$$\theta_{global}^a = \sum\nolimits_{i=1}^{N}\omega_i\cdot\theta_i^a, \quad \omega_i = \frac{n_i^{train}}{\sum_{i'=1}^{N} n_{i'}^{train}}. \tag{8}$$

Next, the server broadcasts the global attention module parameters back to each client. Since the attention module has only a small amount of parameters compared to the CLIP encoders, this strategy has very low communication and computational costs.

3 Experiments

3.1 Datasets

Brain Tumor. We conduct experiments on a public MRI brain tumor classification dataset, denoted as BT [12]. BT has four different classes, namely glioma

Table 1: Samples of each client under non-iid conditions for BT and SC datasets.

Clients	$\alpha = 0.3$		$\alpha = 0.6$		$\alpha = 0.9$	
	BT	SC	BT	SC	BT	SC
Client1	2075	218	1355	387	1480	1156
Client2	389	1374	908	487	423	335
Client3	406	647	607	1365	967	748
Global	394	118	394	118	394	118

tumor, meningioma tumor, no tumor, and pituitary tumor. The training set has 2,870 samples, while the testing set has 394 samples.

Skin Cancer. We also use a public skin cancer (denoted as SC) dataset, obtained from The International Skin Imaging Collaboration (ISIC) [13]. The data set contains the following diseases: actinic keratosis (AK), basal cell carcinoma (BCC), dermatofibroma (DF), melanoma (MEL), nevus (NV), pigmented benign keratosis (PBK), seborrheic keratosis (SK), squamous cell carcinoma (SCC) and vascular lesion (VL). The training set contains 2,239 samples, and the testing set 118 samples. For both the BT and SC dataset, we divide the training set into three clients, following both iid and non-iid conditions.

Real Multi-Source Skin Cancer. We build this dataset (denoted as Real) from three sources, SC, HAM10000, and ISIC2019 [14–16], where each source is treated as an individual client. Since ISIC2019 lacks a test set, we divided it into two parts, one for training and the other for testing, with a ratio of 8:2. We selected the common classes for this dataset: AK, BCC, DF, MEL, NV, PBK, and VL. The global testing set contains 6,233 samples, while Client 1 (SC) has 1,971 samples, Client 2 (ISIC2019) holds 19,766 samples and Client 3 (HAM10000) possesses 8,512 samples. For the BT, SC and Real datasets, each client's data was then divided into three subsets: a training set, a validation set, and a testing set (8:1:1). Finally, we evaluated the global model using the global testing set.

Data under iid condition. For BT and SC, we randomly allocate the training set to each client with an equal number of samples. After this split, in BT, every client receives 956 samples, while in SC, every client holds 746 samples.

Non-iid data. We employ a Dirichlet distribution with concentration parameters $[\alpha, \alpha, \alpha]$, $\alpha \in \{0.3, 0.6, 0.9\}$, as a conjugate prior to generate non-iid data for the clients. For each class, we sample from this distribution and use the sampled values (one for each client) to divide the class examples among the clients. Table 1 shows the results after division for the BT and SC datasets.

3.2 Implementation details

The ViT-B/16 pre-trained model is adopted as the CLIP backbone in our framework. Our light-weight attention module is composed of five layers: a first linear layer, a batch normalization layer, a LeakyReLU layer, a second linear layer, and a softmax activation function. During training, we keep the CLIP encoders frozen and optimize only the attention module's parameters using Adam with

Table 2: Global testing accuracy (ACC%) and balanced accuracy (BACC%) (highest) for BT, SC and Real dataset. The BACC metric is only adopted for Real dataset while the other results indicate ACC. The best results are marked in **bold**. Average indicates the mean value of ACC on all clients. Note: ✘ indicates no partition on data.

Method	$\alpha = 0.3$		$\alpha = 0.6$		$\alpha = 0.9$		*iid*		Average		Real	
	BT	SC	BT	SC	BT	SC	BT	SC	BT	SC	ACC	BACC
Centralized	✘	✘	✘	✘	✘	✘	✘	✘	70.30	**61.01**	60.81	47.80
FedAVG	66.24	48.30	66.49	49.15	64.21	54.23	68.52	49.15	66.36	50.21	60.24	50.96
FedProx	66.24	48.30	63.95	51.69	65.23	49.15	70.05	50.84	66.37	49.99	59.52	48.89
MOON	67.25	50.84	63.96	50.00	65.48	53.38	69.29	47.45	66.49	50.42	59.95	50.90
FedFocal	67.00	49.15	65.48	49.15	67.26	54.23	64.21	51.69	65.99	51.05	59.97	49.80
FedCLIP	68.78	53.38	66.24	55.93	67.26	54.23	66.50	50.0	67.19	53.38	58.43	51.17
Ours	**82.74**	**56.78**	**82.23**	**58.47**	**81.73**	**56.78**	**82.23**	**57.62**	**82.23**	57.41	**72.37**	**63.61**

beta parameters set to 0.9 and 0.98, a weight decay of 0.02, a fixed learning rate of 5×10^{-5}, and a batch size of 32.

For FL, we set the number of global training rounds to 100 for BT, 50 for SC and 50 for Real. For each round, we perform a single epoch of local training and aggregate the parameters of all clients. As image preprocessing, we resized the images to 224×224, and normalized their intensity using z-score normalization for both the training and testing phases [17]. For the LMMD loss $\mathcal{L}_{DA}$, we adopt a Gaussian kernel with a bandwidth set to the median pairwise squared distances in the training data following [11], and use a weight of $\lambda = 1$ for this loss term. The environment used for experiments is based on the Windows 11 operating system, and features an Intel 13900KF CPU with 128 GB of RAM and an RTX 4090 GPU.

3.3 Comparison with state-of-the-art methods

To have a comprehensive comparison, we include several related approaches in our experiments: FedAVG [3], MOON [18], FedProx [19], FedFocal [20], FedCLIP [9] and a centralized method. FedAVG fine-tunes and aggregates all the parameters of the CLIP image and text encoders. MOON extends FedAVG with a contrastive loss between the previous model and the current model. The FedProx approach adds a proximal term to FedAVG that allows having slight differences between clients and the server. FedFocal replaces the standard CE loss with a focal loss for FedAVG. FedCLIP adds an adapter to the CLIP model and only considers the parameters of this adapter during the fine-tuning and aggregation steps. Finally, the centralized method is designed with only one client holding all the training data. For a fair comparison, the same experimental setting described above is used for all tested methods. We evaluate performance using classification accuracy (ACC) as the primary metric. Additionally, due to the class imbalance in the Real dataset, we also consider balanced accuracy (BACC) for a more comprehensive assessment [21]. *Additional implementation details and visualization results can be found in the Supplemental materials.*

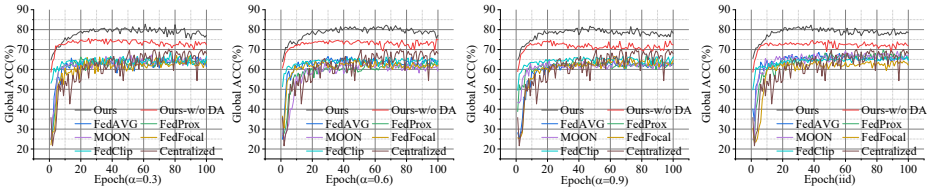

Fig. 2: Global testing accuracy (%) for each round on the BT dataset.

Table 3: Impact of the domain adaptation loss (ACC%).

$\mathcal{L}_{DA}$	$\alpha=0.3$	$\alpha=0.6$	$\alpha=0.9$	*iid*
✓	82.48	82.23	81.73	82.23
✗	75.63	75.12	74.62	74.62

Table 4: Impact of batch size (ACC%).

Batch size	4	8	16	32	Avg
BT (*iid*)	79.69	80.20	81.22	82.23	80.84
Real	63.39	67.32	69.68	72.37	68.19

Table 2 reports the global classification ACC and BACC for the BT, SC and Real datasets. As can be seen, our FACMIC approach yields a better performance than other FL methods across all datasets, outperforming the second-best federated approach by 15.40%, 4.03% and 12.13% in ACC on BT, SC and Real, respectively. FACMIC also achieves the highest BACC in the Real dataset, outperforming the second-best FL approach by 12.44%. Figure 2 shows the global test accuracy measured in each communication round. Our method demonstrates a notable increase in accuracy with a minimal number of epochs, highlighting its effectiveness. Specifically, when aggregating parameters in Eq. 8, it gives a larger weight to clients with more training samples. In the non-iid setting, this enables the global model to learn from the most knowledgeable clients, preventing performance degradation. The lower performance of other methods in the iid setting could be due to the need of fine-tuning the entire image encoder or to the shallower adaptation module architecture (FedCLIP). Our findings suggest that the CLIP model can achieve comparable performance in some medical image classification tasks.

3.4 Ablation study

To validate the effectiveness and generalizability of our model, a series of ablation studies were conducted following the same experimental setting. As reported in Table 3, adding the domain adaptation loss $\mathcal{L}_{DA}$ leads to a better classification accuracy in the BT dataset, in all situations. These results demonstrate the need to address the problem of shifting data distribution among clients.

We also studied the performance of our method for different batch sizes (from 4 to 32), as using large batches is not practical for less capable devices. For sizes of 4 and 8, the adaptation loss is rescaled by a factor of 1/10 since it is too large compared with the contrastive loss in those cases. Table 4 shows the global testing ACC using BT (*iid*) and Real dataset. As one can see, our model is robust to batch size on the BT data, however, using a too small batch size for the large-scale SC data can result in negative adaptation.

Table 5: Global testing accuracy (%) using BT2 dataset.

FedAVG	FedProx	MOON	FedFocal	Centralized	FedCLIP	Ours
84.18	84.72	82.32	79.99	86.12	91.7	**94.42**

To assess our method's generalization ability, we use the fine-tuned global model trained on the BT dataset under iid condition to perform classification directly on another brain tumor dataset (BT2) without any fine-tuning. The BT2 dataset comprises 7023 samples obtained from figshare, SARTAJ, and Br35H [22–25], and has the same classes as the BT dataset. We tested these methods using the whole dataset. As reported in Table 5, our method achieves the highest generalization accuracy of 94.42% on BT2.

4 Conclusion

We explored the usefulness of VLMs for medical imaging in FL and presented a novel FACMIC framework that combines a feature attention module to reduce communication costs and a domain adaptation strategy to minimize data distribution differences between each client. Experimental results in brain tumor and skin cancer classification tasks demonstrate the superior performance of FACMIC compared to state-of-the-art FL approaches.

Acknowledgments. This research was funded by the National NSFC (82260360), the Guilin (20222C264164), and the Guangxi talent (2022AC18004, 2022AC21040).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.
2. Ahmad Chaddad, Yihang Wu, and Christian Desrosiers. Federated learning for healthcare applications. *IEEE Internet of Things Journal*, 11(5):7339–7358, 2024.
3. Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.
4. Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.
5. Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
6. Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 2023.

7. Sixing Yu, J Pablo Muñoz, and Ali Jannesari. Bridging the gap between foundation models and heterogeneous federated learning. *arXiv preprint arXiv:2310.00247*, 2023.
8. Joana Palés Huix, Adithya Raju Ganeshan, Johan Fredin Haslum, Magnus Söderberg, Christos Matsoukas, and Kevin Smith. Are natural domain foundation models useful for medical image classification? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7634–7643, January 2024.
9. Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*, 2023.
10. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
11. Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE transactions on neural networks and learning systems*, 32(4):1713–1722, 2020.
12. Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. Brain tumor classification (mri). `https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri`, 2020.
13. Andrey Katanskiy. Skin cancer dataset. `https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic/data`, 2019.
14. Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.
15. Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. 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)*, pages 168–172. IEEE, 2018.
16. Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.
17. Nanyi Fei, Yizhao Gao, Zhiwu Lu, and Tao Xiang. Z-score normalization, hubness, and few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 142–151, 2021.
18. Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021.
19. Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.
20. Sarkar Dipankar, Narang Ankur, and Rai Sumit. Fed-focal loss for imbalanced data classification in federated learning. In *IJCAI*, 2020.
21. Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

22. Msoud Nickparvar. Brain tumor mri dataset. https://www.kaggle.com/dsv/2645886, 2021.
23. Jun Cheng. brain tumor dataset. https://figshare.com/articles/dataset/brain_tumor_dataset/1512427, 4 2017.
24. Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. Brain tumor classification (mri). https://www.kaggle.com/dsv/1183165, 2020.
25. Ahmed Hamada. Brain tumor mri dataset. https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=no, 2020.