Federated Learning for Medical Image Analysis: A Survey

Hao Guan[a], Mingxia Liu[a,*]

[a]*Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA*

Abstract

Machine learning in medical imaging often faces a fundamental dilemma, namely the small sample size problem. Many recent studies suggest using multi-domain data pooled from different acquisition sites/datasets to improve statistical power. However, medical images from different sites cannot be easily shared to build large datasets for model training due to privacy protection reasons. As a promising solution, federated learning, which enables collaborative training of machine learning models based on data from different sites without cross-site data sharing, has attracted considerable attention recently. In this paper, we conduct a comprehensive survey of the recent development of federated learning methods in medical image analysis. We first introduce the background knowledge of federated learning for dealing with privacy protection and collaborative learning issues in medical imaging. We then present a comprehensive review of recent advances in federated learning methods for medical image analysis. Specifically, existing methods are categorized based on three critical aspects of a federated learning system, including client end, server end, and communication techniques. In each category, we summarize the existing federated learning methods according to specific research problems in medical image analysis, and also provide insights into the motivations of different approaches. In addition, we provide a review of existing benchmark medical imaging datasets and software platforms for current federated learning research. We also conduct an experimental study to empirically evaluate typical federated learning methods for medical image analysis. This survey can help to better understand the current research status, challenges and potential research opportunities in this promising research field.

Keywords: Federated learning, Machine learning, Medical image analysis, Data privacy

1. Introduction

Medical image analysis has been greatly pushed forward by computer vision and machine learning (Barragán-Montero et al., 2021; Cheplygina et al., 2019; Guan and Liu, 2022; Litjens et al., 2017). The remarkable success of modern machine learning methods, *e.g.*, deep learning (LeCun et al., 2015), can be attributed to the building and release of grand-scale natural image databases, such as ImageNet (Deng et al., 2009) and Microsoft Common Objects in Context (MS COCO) (Lin et al., 2014). Unlike natural image analysis, the field of medical image analysis still faces the fundamental challenge of the "small-sample-size" problem (Raudys et al., 1991; Vabalas et al., 2019).

Based on small sample data, it is difficult for us to estimate real data distributions, greatly hindering the building of robust and reliable learning models for medical image analysis. An intuitive and direct solution to this small sample size problem is to pool images from multiple sites together and build larger datasets to train high-quality machine learning models. However, sharing medical imaging data between different sites is intractable due to strict privacy protection policies such as Health Insurance Portability and Accountability Act (HIPAA) (US Department of Health and Human Services, 2020) and General Data Protection Regulation (GDPR) (General Data Protection Regulation, 2019). For example, the United States HIPAA has rigidly restricted the exchange of personal health data and images (US Department of Health and Human Services, 2020). Thus, directly sharing and pooling medical images across different sites/datasets is typically infeasible in real-world practice.

As a promising solution for dealing with the small-sample-size problem and protecting individual privacy, federated learning (McMahan et al., 2017; Bonawitz et al., 2019; Kairouz et al., 2021) has become a spotlight research topic in recent years, which aims to train machine learning models in a collaborative manner without exchanging/sharing data among different sites. As an emerging machine learning paradigm, federated learning deliberately avoids demand for all the medical data residing in one single site. Instead, as shown in Fig. 1, it depends on model aggregation/fusion techniques to jointly train a global model which is then sent/broadcast to each site for fine-tuning and deployment.

There have been several survey papers on federated learning (Li et al., 2021, 2020b; Yang et al., 2019; Rahman et al., 2021; Zhang et al., 2021a; Yin et al., 2021b), but further technical details about facilitating federated learning in medicine and healthcare are not yet covered. Several recent surveys introduce the applications of federated learning in medicine and healthcare areas (Antunes et al., 2022; Rajendran et al., 2021; Nguyen et al., 2022; Pfitzner et al., 2021; Rieke et al., 2020). However, some of them focus on electronic health records (Antunes et al., 2022; Rajendran et al., 2021) or internet of medical things (Aouedi et al., 2023), without paying attention to medical imaging. And

*Corresponding author: mingxia_liu@med.unc.edu

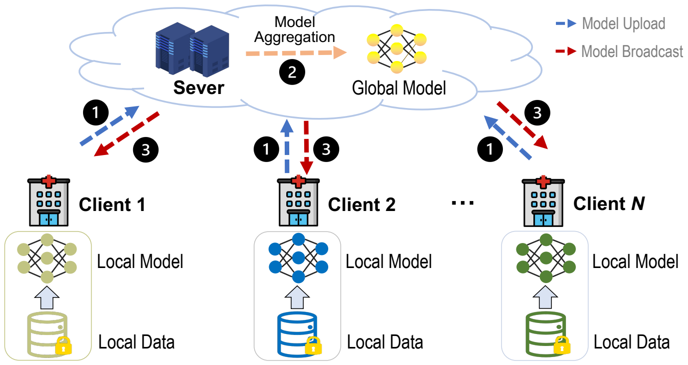

Figure 1: Overview of federated learning (FL) for medical image analysis, including a server and multiple clients. Each selected client trains a model on its local dataset. The server collects the local models and calculates a global model that is broadcast to all the selected clients for deployment.

some survey papers cover very broad areas (Nguyen et al., 2022; Pfitzner et al., 2021), without detailed introduction on federated learning in medical image analysis.

To fill this gap, we review and discuss recent advances in federated learning for medical image analysis in this paper. Our survey paper has significant differences from most previous ones in the following aspects. ***First***, we summarize the existing methods from a system perspective. Specifically, we categorize different approaches into three groups: (1) client-end learning methods, (2) server-end learning methods, and (3) server-client communication methods. Different from previous surveys that are based on multiple research issues in federated learning, this categorization can be more intuitive and clear to picture federated learning. ***Second***, when elaborating on the methods in each group, we have designed a novel "question-answer" paradigm to introduce the motivation and mechanism of each method. We deliberately extract the common questions behind different methods and pose them first in each subsection. These questions stem from the characteristics of medical imaging, thus this "question-oriented" approach of introduction is helpful for providing more insights into different methods. ***Third***, we emphasize the implementation of federated learning techniques for medical image analysis. Specifically, we introduce popular software platforms and benchmark medical imaging datasets for federated learning research in medical imaging. ***In addition***, we also conduct an experiment on a benchmark medical image dataset to illustrate the utility and effectiveness of several typical federated learning methods.

The remainder of this paper is organized as follows. In Section 2, we introduce the background and motivation of federated learning. We summarize existing federated learning studies for medical image analysis in Section 3. In Section 4, software platforms that support federated learning system development are presented. In Section 5, we introduce medical image datasets that have been widely used in federated learning research. We conduct an experimental study in Section 6 to compare several federated learning methods. Challenges and potential research opportunities are discussed in Section 7. Finally, we conclude this survey paper in Section 8.

2. Background

2.1. Motivation

2.1.1. Privacy Protection in Medical Image Analysis

Personal data protection has become an important issue in the digital era. Many governments have introduced tough new laws and regulations on privacy data protection, such as the CCPA in the United States (California Consumer Privacy Act (CCPA), 2018) and GDPR in Europe (General Data Protection Regulation, 2019). In these laws, data protection has been recognized as a fundamental right of natural persons. Collecting, sharing, and processing of personal data are strictly constrained, and violating these laws and regulations may face high-cost penalties (Satariano, 2019).

With these strict restrictions from laws, medical images, one of the most important privacy information, cannot be easily shared among different sites/datasets. To this end, federated learning, a distribution-oriented machine learning paradigm without cross-site data sharing, has emerged as a promising technique for developing privacy-preservation machine learning models, thus paving the way for the applications of medical artificial intelligence (AI) in real-world practice.

2.1.2. Medical Image Data Limitation and Bias

The traditional way to train machine learning models is to use medical images from a specific site/dataset. It has at least two following drawbacks.

(1) Due to the cost of imaging and labeling, the amount of images in local datasets is usually small. This is the well-known "small-sample-size" problem (Raudys et al., 1991; Vabalas et al., 2019). This problem may lead to sub-par learning performance of a model, and produce results that lack statistical significance.
(2) Data from a specific site/dataset may be biased in distribution and not representative of the true data distribution. For instance, it is not unusual that medical sites contain unbalanced data.

Federated learning helps address these limitations, aiming to "pool" medical images together in a distributed way, thereby greatly increasing the sample size. This can effectively take advantage of available data from multiple sites to enhance statistical power of machine learning models.

2.2. Problem Formulation of Federated Learning

Suppose there are N independent clients (sites) with their own datasets $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_N\}$, respectively. Each of the clients (sites) cannot get access to others' datasets. Federated learning (FL) aims to collaboratively train a machine learning model $\mathcal{M}^*$ by gathering information from those N clients (sites) without exchanging/sharing their raw data. The ultimate output of FL is the learned model $\mathcal{M}^*$ which is broadcast to each client for deployment, and the generalizability of $\mathcal{M}^*$ by FL should outperform each local model $\mathcal{M}_i$ (typically with the same model architecture as $\mathcal{M}^*$) learned through local training.

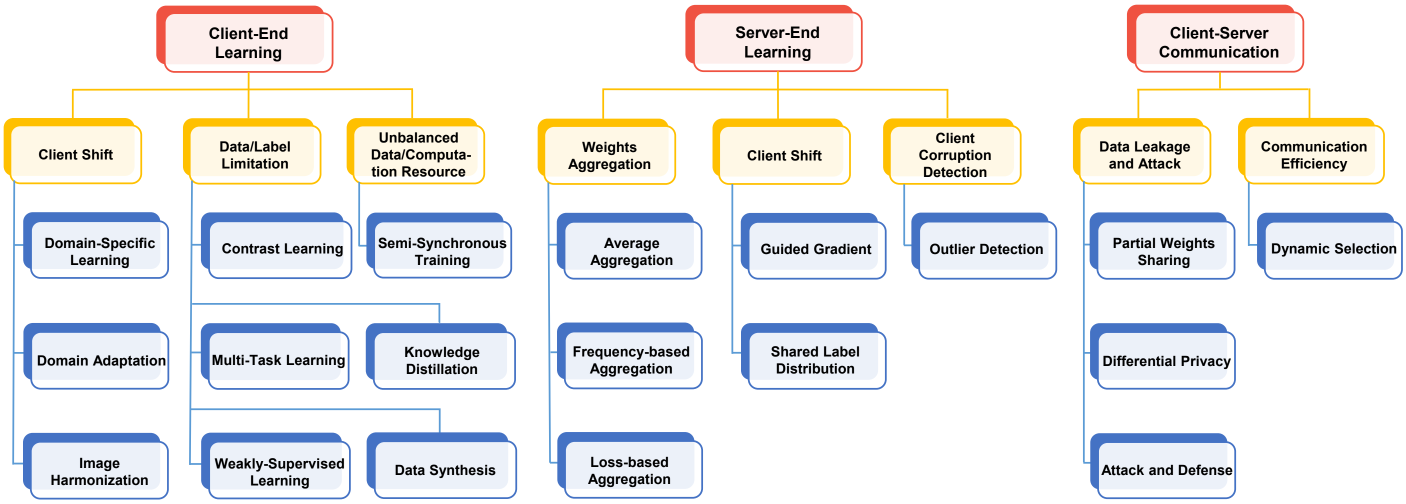

Figure 2: Overview of federated learning (FL) methods for medical image analysis.

2.3. Typical Process of Federated Learning

In Fig. 1, we illustrate the typical process of federated learning that is embodied in a "client-server" architecture. This process encompasses the *Federated Averaging algorithm (FedAvg)* proposed by McMahan*et al.* (McMahan et al., 2017). It serves as the foundation of most popular algorithms for federated learning. A server in a federation triggers and orchestrates the entire training process (without accessing clients' private data) until a certain stop criterion is met, with key components introduced as follows.

1) **Client Selection.** The server selects a set of clients that meet certain requirements. For example, a medical site/dataset might only check in to the server when it can correctly get access to the intranet of a federation with relatively good bandwidth.
2) **Local Training.** Every selected client locally trains a machine learning model through optimization methods (*e.g.*, stochastic gradient descent) based on its local data. In the beginning, the model weights can be initialized both on each client or by the server.
3) **Model Upload.** All the clients that have been selected upload their model (*e.g.*, weights) to the server.
4) **Model Aggregation.** The server computes/updates a global model by aggregating all client models.
5) **Broadcast.** The server sends/broadcasts the current shared global model (*e.g.*, weights) to the selected clients. After downloading and deploying the shared model, a client may continue to update/fine-tune it locally using its private data.

3. Federated Learning for Medical Image Analysis

3.1. Methods Overview: A System Perspective

Federated learning (FL) provides a generic framework for distributed learning with privacy preservation. Most existing machine learning methods can be plugged and integrated into an FL framework. Federated learning is concerned with multiple issues such as data, learning models, privacy protection mechanisms, and communication architecture. As shown in Fig 2, from a system perspective, we categorize existing FL approaches for medical image analysis into three groups: 1) client-end methods, 2) server-end methods, and 2) communication methods. In each group, different methods are clustered according to the specific research problems they aim to address.

3.2. Client-End Learning

3.2.1. Client End: Domain Shift Among Clients

Problem: *Different imaging sites often have significant cross-site data distribution variance caused by different scanning settings and/or subject populations, so how to avoid its negative influence on model training?*

In practice, multi-site medical images may have significantly different data distributions (data heterogeneity), which is the well-known "domain shift" problem (Guan and Liu, 2022) (also referred to as "client shift" in an FL system). As shown in Fig. 3, the three imaging sites have significantly different intensity distributions (in terms of both region-wise and global intensity). In an FL system, domain shifts may cause difficult convergence of the global model and performance degradation of some clients. In the following, we present the relevant studies that focus on reducing domain shift among clients for FL research.
(1) Domain-Specific Learning. Federated learning aims to train a global model that fits well with all clients. Due to cross-site data heterogeneity, the global model may not be able to achieve good performance for all clients. One strategy is fine-tuning the global model using domain-specific (local) data to make it more suitable for a specific client. This method is also known as customized/personalized FL (Wicaksana et al., 2022a; T Dinh et al., 2020; Tan et al., 2022).

Feng et al. (2022) propose an encoder-decoder structure within a federated learning framework for magnetic resonance (MR) image reconstruction. A globally shared encoder is maintained on the server end to learn domain-invariant representations, while a client-specific decoder is trained with local data to take advantage of domain-specific properties of each client. Similar strategies can also be found in (Zhang et al., 2022; Wicaksana et al., 2022a). Chakravarty et al. (2021) propose a federated learning framework that with the combination of a Convolutional

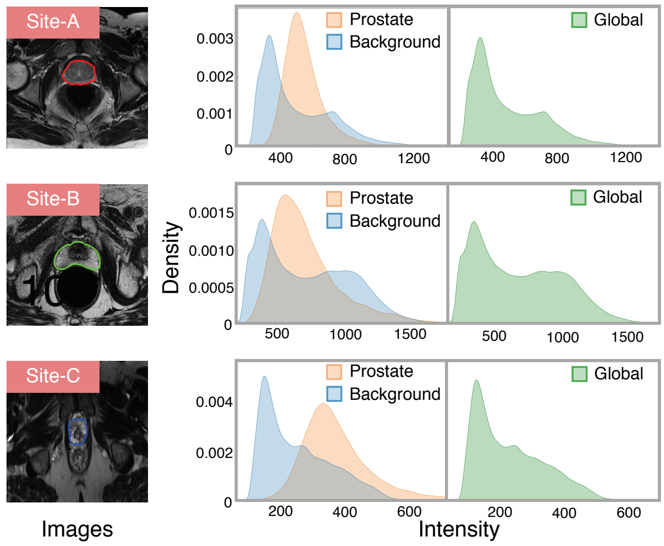

Figure 3: Domain shift among different medical sites. Region-wise and global intensity distribution of different sites for prostate MR images. Image courtesy to Xiao *et al.* (Xiao et al., 2022).

Neural Network (CNN) and a Graph Neural Network (GNN) to tackle the domain shift problem among clients and apply it to chest X-ray image classification. Specifically, model weights of the CNN are shared across clients to learn site-independent features. To address site-specific data variations, a local GNN is built and fine-tuned with local data in each client for disease classification. In this way, both site-independent and site-specific features can be learned. Xu et al. (2022) propose an ensemble-based framework to deal with the client shift for medical image segmentation. Their framework is composed of a global model, personalized models, and a model selector. Instead of only using the global model to fit all the client data, they propose to leverage all the produced personalized models to fit different client data distributions through a model selector. Jiang et al. (2023) propose to train a locally adapted model that accumulates both global gradients (aggregated from all clients) and local gradients (learned from local data) to optimize the model performance on each client. This helps effectively avoid biased performance of the global model on different clients caused by client domain shift. Ke et al. (2021) build a federated learning framework based on Generative Adversarial Network (GAN) to facilitate harmonization (color normalization) of histopathological images. In this method, each client trains a local discriminator to capture client-specific image style, while the server maintains and updates a global generator model to generate domain-invariant images, thus achieving histopathological image harmonization. Similarly, Wagner et al. (2022) propose a GAN model for histopathological image harmonization. In their method, a reference dataset is assumed to be accessible for all clients, which can help the training of all the local GANs at each client.

(2) Domain Adaptation. Domain adaptation (Guan and Liu, 2022; Wilson and Cook, 2020; Kouw and Loog, 2019) is a sound machine learning technique that aims to reduce domain shift among different datasets and enhance the generalizability of a learning model. Many FL studies resort to various domain adaptation algorithms for improved learning performance.

Li et al. (2020d) propose to use domain adaptation methods to align the domain distribution differences among clients. In their method, data in each client are added with noise to achieve privacy protection. A domain discriminator/classifier is trained on these data with noises to reduce domain shift. Dinsdale et al. (2022) propose a domain adaptation-based federated learning framework to remove domain shift among clients caused by different scanners. In their framework, image features are assumed to follow Gaussian distributions, and the mean and standard deviation of the learned features can be shared among clients. During the training of each client model, a label classifier and a domain discriminator are jointly trained to learn features that are domain-invariant, *i.e.*, removing domain shift. Andreux et al. (2020) leverage batch normalization (BN) in a deep neural network to handle client shift. Motivated by BN-based domain adaptation, Li et al. (2018) propose to only share BN parameters (with domain invariant information) in federated training while keeping the BN statistics local because these statistics are assumed to contain domain-specific information. Guo et al. (2021) propose a federated learning method for MRI reconstruction, where the learned intermediate latent features among different clients are aligned with the distribution of latent features of a reference site.

(3) Image Harmonization. Qu et al. (2022) propose a generative replay strategy to handle data heterogeneity among clients. They first train an auxiliary variational autoencoder (VAE) to generate medical images which resemble the input images. Then each client can optimize their local classifier using both the real local data and synthesized data with similar data distribution of other clients. In this way, domain shift can be reduced. Yan et al. (2020) employ cycleGAN (Zhu et al., 2017) to minimize the variations among clients. One client (site) with low data complexity is selected as a reference, then cycleGAN is used to harmonize images from other clients to the reference site. Jiang et al. (2022) propose a frequency-based harmonization method to reduce domain differences among clients. In this method, images are transformed into the frequency domain and phase components are just kept locally, while the average amplitudes from each client are shared and are then normalized to harmonize all the client images.

3.2.2. Client End: Limited Data and Labels

Problem: *Medical imaging datasets are often small-sized and lack label/annotation information, so how to avoid their negative influence on model training (e.g., biased training)?*

In real-world practice, there are often limited medical images in one client (site), and labeled images are even fewer due to the high cost of image annotation/labeling. A client model may be badly trained with limited labeled data, which can cause negative influences on the entire federation. Therefore, how to alleviate the small-sample-size problem is an important topic of federated learning in medical image analysis.

(1) Contrast Learning. Contrastive learning (Chaitanya et al., 2020; He et al., 2020; Misra and Maaten, 2020) is a self-supervised method that can learn useful representations of images by using unlabeled data. A model trained with contrast learning can provide good initialization for further fine-tuning (with a few labeled data) on downstream tasks. Contrast learning has been introduced into federated learning for handling medical data shortage (Wu et al., 2022, 2021). Wu et al. (2022, 2021) use contrast learning to pretrain (initialize) the encoder of a U-Net in each client, then the global U-Net is fine-tuned with limited labeled data. In this way, the negative influence caused by the shortage of labeled medical images can be largely reduced. Similar strategies can be found in (Dong and Voiculescu, 2021).

(2) Multi-Task Learning. Multi-task learning (Smith et al., 2017; Zhang and Yang, 2021) typically solves multiple but related learning tasks at the same time, which can exploit commonalities across tasks. When the training data for each task are small-sized, jointly learning of different tasks can actually share data which is an effective approach for data augmentation. Smith et al. (2017) propose a novel optimization framework, *i.e.*, MOCHA, which extends classic multi-task learning in the federated environment. MOCHA is based on a bi-convex alternating method and is guaranteed to converge. Huang et al. (2022b) propose a federated multi-task framework in which several related tasks, *i.e.*, attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), and schizophrenia (SCZ), are jointly trained. In this method, encoders for each task in clients are federated to derive a global encoder that can learn common knowledge among related mental disorders.

(3) Weakly-Supervised Learning. Weakly-supervised learning (Zhou, 2018) is an extensive group of methods that train a model under weak supervision. Weak supervision information typically includes three types. 1) *Incomplete supervision*. Only a small subset of labeled training data is provided while the other data has no labels. Semi-supervised learning (Yang et al., 2022; Van Engelen and Hoos, 2020) is a popular solution for such scenarios. 2) *Inexact supervision*. Only coarse-grained labels are provided for the training data. Multiple-instance learning (Quellec et al., 2017; Carbonneau et al., 2018) is a representative method to handle this problem. 3) *Inaccurate supervision*. Not all the provided labels are correct. Learning from noisy labels (Song et al., 2022; Frénay and Verleysen, 2013) is the corresponding technique.

Yang et al. (2021a) introduce semi-supervised learning into the federated learning framework which can leverage unlabeled data to assist the federated training. For unlabeled data in a client, the global model assigns them pseudo labels. Meanwhile, it also outputs predictions on augmented data of the original unlabeled data. A consistency loss is utilized on these predictions to further adjust the global model weights. Lu et al. (2022) use multiple-instance learning for local model training on the task of pathology image classification. Whole slide images (WSIs) and weak annotation (*e.g.*, patient or not) are used as the input, with no region-based labels provided. And multiple patches (instances) of a WSI are fed into a network for training. Kassem et al. (2022) build a semi-supervised FL system for surgical phase recognition based on laparoscopic cholecystectomy videos. The key idea is to leverage the temporal information in labeled videos to guide unsupervised learning on unlabeled videos.

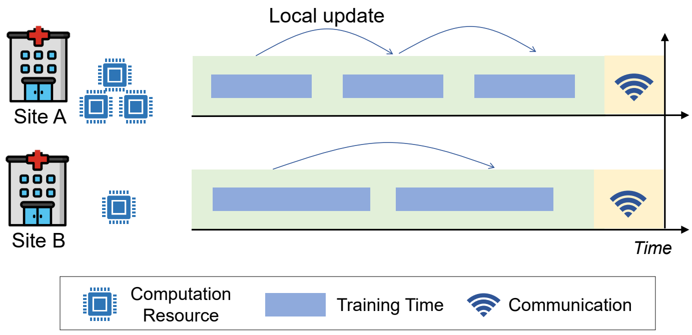

Figure 4: Different local updates for clients with different computation and data resources.

(4) Knowledge Distillation. Kumar et al. (2021a) leverage knowledge distillation for COVID-19 detection using chest X-ray images. The network trained on similar data (other chest X-ray image datasets) is used as a “teacher”, while the client model is a “student”. By matching the softmax activation output of the teacher, the student (client model) can learn useful knowledge for the task. In this way, it can help alleviate the demand for large data during the federation learning process.

(5) Data Synthesis. Zhu and Luo (2022) propose a federated learning framework with virtual sample synthesis for medical image analysis. Given an image $\mathbf{x}$ in the client, the authors first use Virtual Adversarial Training (Miyato et al., 2018) to generate synthetic samples that are similar to $\mathbf{x}$, and then use all the synthesized data for local model training/updating. Peng et al. (2022) propose a federated graph learning framework for brain disease prediction, where a Graph Convolutional Network (GCN) is used as the local learning model. Considering the missing nodes and edges when separating the global graph into local graphs, the authors leverage network inpainting to predict the missing nodes and their associated edges. This helps complete the graphs for GCN training in each client, with results suggesting its effectiveness in graph data synthesis and augmentation.

3.2.3. Client End: Unbalanced Data and Computation Resource

Problem: *Different medical sites/datasets may have significantly different data scales and computation resources (e.g., number of GPUs) in real-world practice, so how to reduce its influence on federated training?*

In the standard template of federated learning, *i.e.*, FegAvg, the learner in each client conducts a predefined number of local training epochs (with equal batches and learning rate) before reaching a synchronization time point when it sends its model to the server. However, if the clients have significantly different computation and data resources, this may lead to computational inefficiencies and slow convergence of model optimization. For example, each client is supposed to conduct 50 epochs’ updates before sharing its weight. In such a setting, a client with advanced GPUs may take 1 second, while a client with weak computation utility may take 100 seconds. In such a case, the stronger client will have to spend 99 seconds waiting for weight sharing.

Aiming at handling the computational and data scale heterogeneity among clients, Stripelis et al. (2021) propose a Semi-Synchronous Training strategy in federated learning and apply it to the task of brain age prediction. As shown in Fig. 4, in their method, each client conducts a variable number of updates (epochs) between synchronization time points which depend on its computational power and data scale. Higher computation power or fewer local data will lead to more local updates (epochs).

3.3. Sever-End Learning

3.3.1. Sever End: Weight Aggregation

Problem: *How to aggregate the weights of clients properly to avoid performance degradation after each client-server communication?*

Chen et al. (2022) propose a Progressive Fourier Aggregation strategy at the server end. Based on previous studies that low-frequency components of parameters form the basis of deep network capability (Liu et al., 2018), only these low-frequency components are aggregated to share knowledge learned from different clients, while the high-frequency parts are disregarded. Li et al. (2022) consider the training loss of each client as the impact factor of the weight aggregation. The client with relatively bad performance caused by uneven data will get a smaller weight for the global weight aggregation.

3.3.2. Sever End: Domain Shift Among Clients

Problem: *The domain shift among clients may cause non-convergence of federated models, so how to avoid this from the server end?*

Hosseini et al. (2023) argue that the data heterogeneity between different medical centers (clients) may lead to a biased global model, *i.e.*, a model that has good performance for some clients while exhibiting inferior performance for the other clients. Thus, they propose a revised optimization objective (motivated by fair resource allocation approaches in wireless network research), to facilitate uniform model performance across all the clients. In their method, the clients for which the global model has inferior performance will contribute more to the total loss function. Fan et al. (2021) leverage the guided-gradient to optimize the global model. After aggregating all the local weights of the clients, only positive values of the aggregated weights are used to update the global. The authors argue that this is helpful for the global gradient descent to go towards the optimal direction, and the guided-gradient can reflect the most influential regions of the medical images.

Luo and Wu (2022) propose a method called federated learning with shared label distribution (FedSLD) for medical image classification by mitigating label distribution differences among clients. In their method, it is assumed that the amount of samples of each category (label distribution) is known for the entire federation. During local training in client i, a weighted cross-entropy loss is designed as the batch loss. The weight is computed as the label distributions in each batch, with respect to their label distributions across the entire federation.

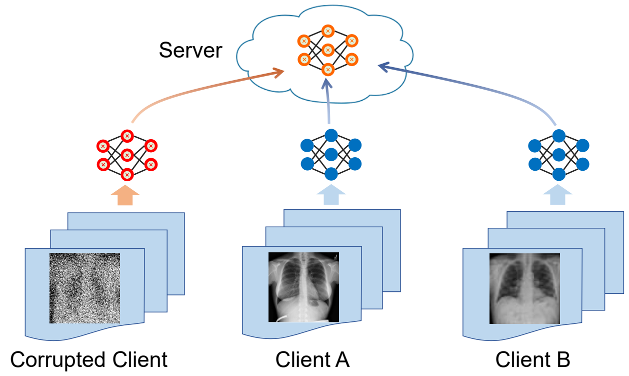

Figure 5: Corrupted clients will lead to a corrupted global model, thus negatively influencing the entire federated learning system.

3.3.3. Sever End: Client Corruption/Anomaly Detection

Problem: *If one or more clients are corrupted by very noisy labels or malicious attacks, how to avoid its negative influence on the entire federation?*

Classic federated learning framework holds the assumption that all the clients work normally. In this context, the term “normal” means that a client is trained with correctly labeled data or the client is honest without malicious attack. In real-world practice (as shown in Fig. 5), however, a client may be trained with “dirty” data that have noisy labels or suffered from poisoning attacks from malicious parties. How to deal with this issue is critical for ensuring the safety of a federated learning system. Alkhunaizi et al. (2022) propose a sever-end outlier detection method, called Distance-based Outlier Suppression (DOS), which is robust to client corruption/failure. In this method, the weight of each client is calculated based on an anomaly score for the client using Copula-based outlier detection. A client with a high outlier score will get a tiny weight during model aggregation, thus reducing the negative influence of corrupted clients. Experimental results on clients with noisy labels demonstrate the effectiveness of this method.

3.4. Client-Server Communication

3.4.1. Data Leakage and Attack

Problem: *How to avoid data leakage and privacy violation during the interaction/communication between the server and clients during federated training?*

Protection of data privacy, *i.e.*, ensuring the data of each client are not seen and accessed by other clients/sever, is the main concern and motivation of federated learning systems. Prior studies have shown that, even without intersite data sharing, pixel-level images can be reconstructed or recovered by the leaked gradients of a machine learning model (Geiping et al., 2020; Yin et al., 2021a; Zhu et al., 2019). Therefore, it is critical to study advanced techniques to proactively avoid data leakage during communication between the server and multiple clients. Many studies focus on this topic in recent years.

(1) Partial Weights Sharing. Yang et al. (2021c) consider that sharing an entire model (network) may not fully protect privacy, and thus propose sharing a partial model for federated learning on medical datasets. Specifically, clients only share the feature-learning part of a model for aggregation on the server while keeping the last several layers private. Similar strategies can also be found in (Li et al., 2019).

(2) Differential Privacy. Gradient information of a deep neural work may contain individual privacy that can be reconstructed by malicious parties. Differential privacy (Dwork et al., 2014) could limit the certainty in inferring an individual's presence in the training dataset. And several recent studies (Li et al., 2020d; Lu et al., 2022; Malekzadeh et al., 2021) propose to add Gaussian random noise to the computed gradients on the patients' imaging data in each client/site, thus protecting privacy from the server and other clients.

(3) Attack and Defense. Kaissis et al. (2021) apply gradients attack (Geiping et al., 2020) to a medical image classification system, and conduct an empirical study on its capability of reconstructing training images from clients in an FL system. Hatamizadeh et al. (2023) propose a gradient inversion algorithm to estimate the running statistics (*i.e.*, running mean and variance) of BN layers to match the gradient from real images and the synthesized ones, thus generating synthesized images that are very similar to the original ones. They further propose a method to measure and visualize the potential data leakage.

3.4.2. Communication Efficiency

To improve communication efficiency, Zhang et al. (2021b) propose a dynamic fusion-based federated learning approach for COVID-19 diagnosis. Their framework dynamically selects the participating clients for weight fusion according to the performance of local client models, and then performs model aggregation based on participating clients' training time If a client does not upload the updated model within a certain waiting time, it will be excluded by the central server for this aggregation round.

4. Software Platforms and Tools

In this section, we review several popular and influential federated learning platforms. These software platforms provide application interfaces (APIs) for the development of FL systems, which can boost the efficiency and robustness of building large FL systems.

4.1. PySyft

PySyft (Ziller et al., 2021)[1] is an open-source FL library enabling secure and private machine learning by wrapping popular deep learning frameworks. It is implemented by Python and can run on Linux, MacOS, and Windows systems. PySyft has attracted more than 8,000 stars and 1,900 forks on GitHub[2], which shows its popularity. Budrionis et al. (2021) carry out an empirical study using PySyft on a medical dataset. Their experimental results demonstrate that the performance of machine learning models trained with federated learning is comparable to those trained on centralized data.

[1]https://github.com/OpenMined/PySyft

[2]https://github.com

4.2. OpenFL

The Open Federated Learning (OpenFL)[3] is an open-source FL framework initially developed for use in medical imaging. OpenFL is built through a collaboration between Intel and the University of Pennsylvania (UPenn) to develop the Federated Tumor Segmentation (FeTS) platform[4]. OpenFL supports model training with PyTorch and TensorFlow. Foley et al. (2022) provide several use cases of OpenFL in medicine, such as tumor segmentation and respiratory distress syndrome prediction.

4.3. PriMIA

The Privacy-preserving Medical Image Analysis (PriMIA) (Kaissis et al., 2021) is an open-source framework for privacy-preserving decentralized deep learning with medical images. PriMIA is built upon the PySyft ecosystem which supports Python and PyTorch for deep learning development. It is compatible with a wide range of medical imaging data formats. The source code, documentation as well as publicly available data can be found online (`https://zenodo.org/record/4545599`). For example, Kaissis et al. (2021) use PriMIA to perform classification on pediatric chest X-rays and achieve good results.

4.4. Fed-BioMed

Fed-BioMed[5] is an open-source federated learning software for real-world medical applications. It is developed by Python and supports multiple machine learning toolkits such as PyTorch, Scikit-Learn, and NumPy. It can also be used in cooperation with PySyft. Silva et al. (2020) use Fed-BioMed to conduct multi-center analysis for structural brain imaging data (MRI) across different datasets and verify its effectiveness.

4.5. TFF

The TensorFlow Federated (TFF)[6] is an open-source framework for general-purpose federated learning developed by Google. TFF is implemented by Python. Its strength lies in that it can be seamlessly integrated with TensorFlow[7]. Users of TensorFlow and Keras[8] could easily construct federated learning systems using TFF.

4.6. FATE

The Federated AI Technology Enabler (FATE)[9] is a general-purpose federated learning framework developed by WeBank. It is implemented by Python and can run on Linux/Mac systems on a single host or on multiple nodes. The FATE provides a collection of machine learning algorithms within the federated framework, including logistic regression, tree-based models, and deep neural networks.

Due to the increasing and extensive influence of federated learning, many software platforms and frameworks have been proposed to date. More comparative reviews and evaluations can be found in (Kholod et al., 2020; Li et al., 2021).

[3]https://github.com/securefederatedai/openfl

[4]https://www.fets.ai

[5]https://fedbiomed.gitlabpages.inria.fr

[6]https://github.com/tensorflow/federated

[7]https://github.com/tensorflow/tensorflow

[8]https://keras.io

[9]https://github.com/FederatedAI/FATE

Table 1: Overview of benchmark datasets for federated learning research on different medical image analysis tasks.

Studies	Task	Dataset	Modality	Model
Brain				
(Peng et al., 2022)	ASD, AD classification	ABIDE, ADNI	fMRI	GCN
(Gürler and Rekik, 2022)	Brain connectivity prediction	OASIS	MRI	GNN
(Islam et al., 2022)	Brain tumor classification	UK Data Service	MRI	CNN
(Dinsdale et al., 2022)	Age prediction	ABIDE	MRI	CNN (VGG)
(Qi et al., 2022)	Intracranial hemorrhage diagnosis	RSNA	CT	CNN (DenseNet)
(Stripelis et al., 2021)	Brain age prediction	UK Biobank	MRI	CNN
(Liu et al., 2021b)	Intracranial hemorrhage diagnosis	RSNA	CT	CNN (DenseNet)
(Fan et al., 2021)	ASD classification	ABIDE	MRI	CNN
(Li et al., 2020d)	ASD classification	ABIDE	fMRI	MLP
(Sheller et al., 2019)	Brain tumor segmentation	BraTS	MRI	U-Net
(Li et al., 2019)	Brain tumor segmentation	BraTS	MRI	CNN
Chest				
(Hatamizadeh et al., 2023)	Image generation (attack)	COVID-19 CXR ChestX-ray14	Chest X-ray	CNN (ResNet)
(Yan et al., 2023)	Classification	COVID-FL	Chest X-ray	Transformer
(Alkhunaizi et al., 2022)	Classification	CheXpert	Chest X-ray	CNN
(Dong et al., 2022)	Classification	ChestX-ray14	Chest X-ray	CNN
(Chakravarty et al., 2021)	Classification	CheXpert	Chest X-ray	CNN, GNN
Lung				
(Yang et al., 2021c)	COVID-19 diagnosis	COVIDx	Chest X-ray	CNN
(Feki et al., 2021)	COVID-19 diagnosis	Local dataset	Chest X-ray	CNN
(Kumar et al., 2021a)	COVID-19 diagnosis	COVID-19 CXR	Chest X-ray	CNN
(Dong and Voiculescu, 2021)	COVID-19 diagnosis	COVID-19 CXR	Chest X-ray	CNN
(Yang et al., 2021a)	Segmentation	Local dataset	CT	CNN
Heart				
(Linardos et al., 2022)	Cardiac diagnosis	ACDC, M&M	MRI	CNN
(Qi et al., 2022)	Cardiac segmentation	M&M, Emidec	MRI	U-Net
(Li et al., 2020a)	Cardiac image synthesis	Local dataset	CT	GAN
(Wu et al., 2022)	Cardiac segmentation	ACDC	MRI	U-Net
Breast				
(Agbley et al., 2023)	Breast tumor classification	BreakHis	Pathology	CNN
(Wicaksana et al., 2022b)	Breast tumor segmentation	BUS etc.	Ultrasound	U-Net
Skin				
(Yan et al., 2023)	Skin lesion classification	ISIC	Dermoscopy	Transformer
(Wicaksana et al., 2022a)	Skin lesion classification	HAM10000	Dermoscopy	CNN
(Alkhunaizi et al., 2022)	Skin lesion classification	HAM10000	Dermoscopy	CNN
(Qi et al., 2022)	Skin lesion classification	HAM10000	Dermoscopy	CNN (DenseNet)
(Liu et al., 2021b)	Skin lesion classification	HAM10000	Dermoscopy	CNN (DenseNet)
(Bdair et al., 2021)	Skin lesion classification	HAM10000 ISIC	Dermoscopy	CNN (Efficient-Net)
(Chen et al., 2021)	Skin lesion classification	HAM10000 ISIC	Dermoscopy	CNN (VGG)
Eye				
(Yan et al., 2023)	Diabetic classification	Retina	Color retinal image	Transformer
(Qiu et al., 2023)	Fundus segmentation	RIM-ONE etc.	Color retinal image	CNN (MobileNet)
(Qu et al., 2022)	Diabetic classification	Retina	Color retinal image	VAE, CNN
Abdomen				
(Zhu et al., 2023b)	Prostate segmentation	PROMISE12 NCI-ISBI 2013	MRI	U-Net
(Qiu et al., 2023)	Prostate segmentation	PROMISE12	MRI	CNN (MobileNet)
(Xu et al., 2023)	Tumor segmentation	LiTS etc.	CT	U-Net
(Wicaksana et al., 2022a)	Cancer classification	ProstateX	MRI	CNN
(Luo and Wu, 2022)	Cancer classification	MedMNIST	CT	CNN
(Liu et al., 2022)	Polyp detection	GLRC	Colonoscopy	CNN
(Yan et al., 2020)	Cancer classification	ProstateX	MRI	GAN
(Roth et al., 2021)	Prostate segmentation	MSD-Prostate PROMISE12 ProstateX NCI-ISBI 2013	MRI	U-Net
Histology				
(Hosseini et al., 2023)	Cancer classification	TCGA	Pathology	CNN (DenseNet)
(du Terrail et al., 2023)	Cancer classification	Local dataset	Pathology	CNN
(Lu et al., 2022)	Cancer classification	TCGA	Pathology	CNN
(Adnan et al., 2022)	Cancer classification	TCGA	Pathology	CNN (DenseNet)
(Luo and Wu, 2022)	Cancer classification	MedMNIST	Pathology	CNN
(Wagner et al., 2022)	Image harmonization	PESO	Pathology	GAN
(Ke et al., 2021)	Image harmonization	TCGA etc.	Pathology	GAN
Others				
(Feng et al., 2022)	MRI reconstruction	fastMRI, BraTS	MRI	U-Net
(Elmas et al., 2022)	MRI reconstruction	fastMRI, BraTS, IXI	MRI	GAN
(Guo et al., 2021)	MRI reconstruction	fastMRI, BraTS, IXI	MRI	U-Net

5. Medical Image Datasets for Federated Learning

In this section, we introduce the benchmark datasets that have been commonly used in federated learning for medical image analysis. For clarity, these datasets are presented in terms of different research objects/organs.

5.1. Medical Image Data Usage Overview

For most existing FL research in medical image analysis, there are typically two ways of using different imaging datasets for simulation and experiment. The first way is to directly use databases from different medical sites/centers (Li et al., 2020d; Dayan et al., 2021). These databases are typically research projects that are built through multi-center cooperation. Thus, they are ideal choices to set up a FL simulation environment. Another popular way to build an FL experiment platform is to split a very large-scale medical image dataset into several subsets (Chakravarty et al., 2021; Alkhunaizi et al., 2022), where each subset is treated as a client dataset.

5.2. Brain Images

5.2.1. ADNI

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005; Jack Jr et al., 2008) is the largest and most influential benchmark for the research of Alzheimer's Disease (AD), including ADNI-1, ADNI-2, ADNI-GO and ADNI-3. Structural brain MRI, functional MRI, and positron emission tomography (PET) from 1,900+ subjects and 59 centers are provided for analysis and research.

5.2.2. ABIDE

Autism Brain Imaging Data Exchange (ABIDE) initiative (Di Martino et al., 2014) is a benchmark database for research on Autism spectrum disorder. ABIDE contains both structural and functional brain images independently collected from more than 24 imaging laboratories/sites around the world.

5.2.3. BraTS

Multimodal Brain Tumor Image Segmentation Benchmark (BraTS) (Menze et al., 2014) is a benchmark dataset for brain tumor segmentation. BraTS is updated regularly for the Brain Tumor Segmentation Challenge[10]. It contains brain MRIs acquired by various scanners from around 19 independent institutions.

5.2.4. RSNA Brain CT

Radiological Society of North America (RSNA) (Flanders et al., 2020) is a large-scale multi-institutional CT dataset for intracranial hemorrhage detection. RSNA contains 874,035 images which are compiled and archived from three different institutions, *i.e.*, Stanford University (Palo Alto, USA), Thomas Jefferson University Hospital (Philadelphia, USA), and Universidade Federal de So Paulo (So Paulo, Brazil).

[10]https://www.med.upenn.edu/cbica/brats

5.2.5. UK BioBank

UK Biobank (Miller et al., 2016) is a large-scale brain imaging dataset that consists of around 100,000 participants with brain imaging in structural, functional, and diffusion modalities.

5.2.6. IXI

IXI Dataset[11] consists of around 600 MR images from healthy subjects. All the images are acquired from three different hospitals (using different scanners or scanning parameters) in London.

5.3. Chest/Lung/Heart Images

5.3.1. CheXpert

CheXpert (Irvin et al., 2019) is a large-scale dataset including 224,316 chest radiographs of 65,240 patients. These images are acquired from Stanford University Medical Center.

5.3.2. ChestX-ray

The ChestX-ray (also known as ChestX-ray14)[12] is a large and publicly-available medical image dataset that contains 112,120 X-ray images (in frontal-view) of 30,805 patients with 14 disease labels. It is expanded from the ChestX-ray8 dataset (Wang et al., 2017) by adding six thorax diseases, including Edema, Emphysema, Fibrosis, Hernia, Pleural, and Thickening.

5.3.3. COVID-19 Chest X-ray

The COVID-19 Chest X-ray (also known as COVID-19 CXR) (Chowdhury et al., 2020)[13] is a publicly-available database of chest X-ray images, containing 3,616 COVID-19 positive cases, 10,192 normal controls, 6,012 lung opacity (non-COVID infection), and 1,345 viral pneumonia cases.

5.3.4. COVIDx

The COVIDx dataset (Wang et al., 2020) is a large-scale and fully accessible database comprising 13,975 chest X-ray images of 13,870 patients. COVIDx includes 358 chest X-ray images from 266 COVID-19 patient cases, 8,066 normal cases, and 5,538 non-COVID-19 pneumonia cases.

5.3.5. ACDC

Automatic Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018) is a large publicly available and fully annotated dataset for cardiac MRI assessment. This dataset consists of 150 patients that are divided into 5 categories in terms of well-defined characteristics based on physiological parameters.

[11]https://brain-development.org/ixi-dataset
[12]https://www.kaggle.com/datasets/nih-chest-xrays/data
[13]https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

5.3.6. M&M

Multi-Center, Multi-Vendor, and Multi-Disease Cardiac Segmentation (M&Ms) Challenge (Campello et al., 2021)[14] is a publicly available cardiac MRI dataset. This dataset contains 375 participants from 6 different hospitals in Spain, Canada, and Germany. All the cardiac MRIs are acquired by 4 different scanners (*i.e.*, GE, Siemens, Philips, and Canon).

5.4. Skin Images

5.4.1. HAM10000

The "Human Against Machine with 10000 training images" (HAM10000) (Tschandl et al., 2018)[15] is a popular large-scale dataset for diagnosis of pigmented skin lesions. It consists of 10,015 dermatoscopic images from different sources. Cases in this dataset include a collection of all representative diagnostic categories of pigmented lesions.

5.4.2. ISIC

The International Skin Imaging Collaboration (ISIC) challenge dataset (Cassidy et al., 2022)[16] is a large-scale database, containing a series of challenges for skin lesion image analysis. ISIC has become a standard benchmark dataset for dermatoscopic image analysis.

5.5. Others

5.5.1. Eye: Kaggle Diabetic Retinopathy (Retina)

The Kaggle Diabetic Retinopathy (Retina)[17] is a large-scale dataset of color digital retinal fundus images for diabetic retinopathy detection. It includes 17,563 pairs of color digital retinal fundus images. Each image in this dataset is provided a label (a rated scale from 0 to 4) in terms of the presence of diabetic retinopathy, where 0 to 4 represents no, mild, moderate, severe, and proliferative diabetic retinopathy, respectively.

5.5.2. Abdomen: PROMISE12

The MICCAI 2012 Prostate MR Image Segmentation challenge dataset (PROMISE12)(Litjens et al., 2014) is a publicly available dataset for the evaluation of prostate MRI segmentation methods. It consists of 100 prostate MRIs acquired by different scanners from 4 independent medical centers, including University College London in the United Kingdom, Haukeland University Hospital in Norway, the Radboud University Nijmegen Medical Centre in the Netherlands, and the Beth Israel Deaconess Medical Center in the USA.

5.5.3. Histology: TCGA

The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013)[18] is a large-scale landmark cancer genomics database. Whole-slide images for normal controls and cancers are provided for histology and microscopy research.

5.5.4. Knee: fastMRI

The fastMRI (Knoll et al., 2020; Muckley et al., 2021)[19] is a large-scale dataset for medical image reconstruction using machine learning approaches. This dataset contains more than 1,500 knee MRIs (1.5 and 3 Tesla) and DICOM images from 10,000 clinical knee MRIs (1.5 and 3 Tesla).

5.5.5. MedMNIST

MedMNIST (Yang et al., 2021b) is a dataset for medical image classification. Similar to the MNIST dataset[20], all the images in the MedMNIST are stored as the size of 28×28. The MedMNIST includes 10 pre-processed subsets, covering primary modalities (*e.g.*, MR, CT, X-ray, Ultrasound, OCT). As a lightweight dataset with diversity, MedMNIST is good for rapid prototyping machine learning algorithms.

6. Experiment

To empirically evaluate the federated learning performance of different approaches for medical image analysis, we conduct an experiment to assess several representative FL methods and some methods with diverse settings on a popular benchmark dataset.

6.1. Dataset

We conduct the experiment on the popular benchmark ADNI dataset (Mueller et al., 2005; Jack Jr et al., 2008). Two studies/phases in ADNI (*i.e.*, ADNI-1 and ADNI-2) with baseline data are used as two client datasets, where subjects that appear in both ADNI-1 and ADNI-2 are removed from ADNI-2 for independent evaluation. Specifically, ADNI-1 consists of 1.5T T1-weighted structural MRIs of 428 subjects (including 199 patients with AD and 229 normal controls (NCs)), while ADNI-2 contains 3.0T T1-weighted structural MRIs of 360 subjects (including 159 AD patients and 201 NC subjects). We use brain regions-of-interest (ROI) as the features to represent each MRI. The ROI features are calculated based on the mean gray matter volumes of 90 brain regions defined in the AAL atlas (Tzourio-Mazoyer et al., 2002). In all experiments, for each client, 80% of the dataset is randomly selected to construct the training set, while the remaining 20% samples are used for test. To avoid bias caused by random partition, the random partition process is repeated five times, and we record and report the mean and standard deviation results.

6.2. Experimental Setup

The task here is AD vs. NC classification based on structural MRI data. We use four metrics to evaluate the classification performance, including classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the ROC curve (AUC). Logistic Regression (with model weight $\mathbf{w}$) is used as the machine learning model for each FL setting, which has been widely used in medical imaging analysis (Divya and Shantha Selva Kumari, 2021; Bzdok et al., 2015; Wachinger et al., 2016; van Ravesteijn et al., 2009).

[14] https://www.ub.edu/mnms
[15] https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000
[16] https://challenge.isic-archive.com/data
[17] https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data
[18] https://www.cancer.gov/ccg/research/genome-sequencing/tcga
[19] https://fastmri.med.nyu.edu
[20] http://yann.lecun.com/exdb/mnist

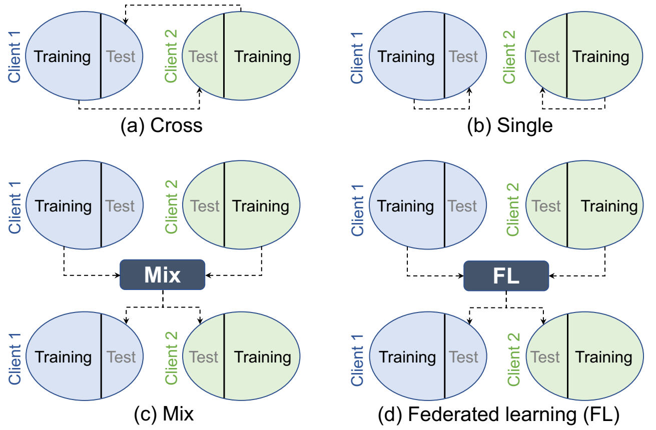

Figure 6: Different settings for performance comparison.

Table 2: Classification results (mean±standard deviation) of different federated learning settings in terms of four metrics. ADNI1-tr: ADNI-1 is adopted as the training set. ADNI2-tr: ADNI-2 is used as the training set.

Client	Method	ACC	SEN	SPE	AUC
ADNI-1	ADNI1-tr	–	–	–	–
	ADNI2-tr	0.818	0.809	0.825	0.886
	Single	0.844±0.013	0.786±0.045	0.895±0.040	0.889±0.018
	Mix	0.870±0.017	0.823±0.045	0.923±0.050	0.901±0.018
	FedAvg	0.860±0.028	0.783±0.025	0.901±0.049	0.897±0.013
	FedSGD	0.823±0.030	0.752±0.047	0.882±0.011	0.880±0.034
	FedProx	0.858±0.031	0.815±0.077	0.896±0.034	0.900±0.044
ADNI-2	ADNI1-tr	0.811	0.623	0.960	0.885
	ADNI2-tr	–	–	–	–
	Single	0.828±0.016	0.750±0.045	0.890±0.046	0.863±0.043
	Mix	0.872±0.012	0.843±0.048	0.898±0.038	0.910±0.013
	FedAvg	0.842±0.021	0.823±0.018	0.864±0.038	0.907±0.017
	FedSGD	0.844±0.039	0.819±0.066	0.871±0.056	0.908±0.040
	FedProx	0.856±0.045	0.845±0.072	0.861±0.047	0.908±0.036

6.3. Federated Learning Settings for Comparison

We compare 3 conventional machine learning and 3 popular FL methods in our study, with details given below.

(1) **Cross**. Training is conducted on one client dataset and then the trained model is directly tested on the data of the other client, as shown in Fig. 6 (a). Specifically, ADNI-1 is used as the training set (denoted as ADNI1-tr), then the trained model is tested on ADNI-2. ADNI-2 is used as the training set (denoted as ADNI2-tr), then the trained model is evaluated on ADNI-1.

(2) **Single**. Training and testing are conducted within each client dataset separately, as shown in Fig. 6 (b). In each client, 80% of the data is used for training while the other is used for testing.

(3) **Mix**. All the training data in each client are pooled together for training a model, then the trained model is evaluated on the test data of all the clients, as shown in Fig. 6 (c). Note this strategy needs to share data, and thus, could not preserve privacy.

(4) **FedAVG** (McMahan et al., 2017; Li et al., 2020e). Each client trains its own model, then their model weights (*e.g.*, the weight $\mathbf{w}$ of logistic regression) are aggregated to calculate a global model. The final trained global model is tested on all the test data in each client, as shown in Fig. 6 (d). The number of iterations for local model training is set to 10.

(5) **FedSGD** (McMahan et al., 2017). Each client trains a local model, then the gradients from each client are aggregated to calculate a global model. The global model is then applied to all the test data in each client for assessment, as shown in Fig. 6 (d). The number of iterations for local model training is set to 10.

(6) **FedProx** (Li et al., 2020c). Every client trains its own model with an additional proximal term (the coefficient μ is set to 0.1). Local training is conducted only once. The model weights of each client are aggregated to get a global model. The trained global model is then assessed on the test data in each client, as shown in Fig. 6 (d).

6.4. Result and Analysis

The classification result of different methods is shown in Table 2. In the "Cross" setting, the client dataset for training is denoted as "<client> (tr)". Since there is only one test dataset in this setting, no standard deviation is reported.

From Table 2, we can get the following observations. 1) The "mix" strategy has the best performance. This is because it combines all the training data of the clients together and the learning model can get access to the largest amount of data information than the other methods. 2) The "cross" strategy has the worst performance. This should be caused by the well-known "domain shift" problem. Since ADNI-1 and ADNI-2 have different scanning parameters and populations, then directly transferring a model may not achieve good classification results. 3) Federated learning methods achieve satisfactory performance. This can be explained by FL can leverage more data information than the baseline methods (*i.e.*, "cross" and "single") even without cross-site data sharing. 4) Among the FL methods, we find that aggregation of model weights (*i.e.*, FedAvg, FedProx) can be more advantageous than a fusion of the gradients of each client model (*i.e.*, FedSGD).

7. Discussion

7.1. Challenges of FL for Medical Image Analysis

7.1.1. Data Heterogeneity Among Clients

Data heterogeneity is widespread in real-world medical image sites. Such heterogeneity can hardly be avoided in practice due to the following factors. 1) Medical images from different sites/datasets are typically acquired by different scanners or scanning protocols. 2) Patients in different sites/hospitals have different distributions. The heterogeneous data distribution, *i.e.*, "domain shift" or "client shift", may cause significant degradation or biased performance of a federated learning system. How to alleviate the negative influence of data heterogeneity is one of the most important and challenging research problems for federated learning in medical imaging.

7.1.2. Privacy Leakage/Poisoning Attacks

In classic FL, only the model parameters (*e.g.*, weights) are exchanged and updated without data sharing. This is considered an effective way of privacy protection. But further research reveals that FL still faces privacy and security risks, including privacy leakage (Geiping et al., 2020; Yin et al., 2021a; Zhu et al., 2019) and poisoning attacks (Lyu et al., 2022; Xia et al., 2023). These issues can happen at both the server end and the client end. Since an FL system contains the communication and interaction of many entities/parties, how to effectively protect individual privacy and data security is a very challenging problem.

7.2. Future Research Directions

7.2.1. Dealing with Client Shift

Domain shift between client datasets (client shift) has become a major concern of federated learning in medical image analysis. To tackle this problem, domain adaptation (Guan and Liu, 2022) has attracted extensive interest. Classic domain adaptation methods typically need access to both source and target domains which may violate the privacy protection restraint. Thus, developing more efficient federated domain adaptation methods will be a promising research direction. Another promising solution is personalized FL techniques (T Dinh et al., 2020; Tan et al., 2022) which utilize local data to further optimize a trained global model.

7.2.2. Multi-Modality Fusion for FL

Numerous imaging techniques/tools have been developed to create various visual representations of every subject, such as structural MRI, functional MRI, computed tomography (CT), and positron emission tomography (PET). Most existing FL studies only focus on images of a single modality in each client. How to leverage multi-modal imaging data in an FL system is an interesting problem with practical value. Currently, a few works make early steps on FL with multiple modalities (Qayyum et al., 2022). More research work is expected on this topic.

7.2.3. Model Generalizability for Unseen Clients

Most existing FL studies focus on model training and test within a fixed federation system. That is, a global model is trained on and applied to the same client datasets (inside clients). An interesting question is: when facing data from unseen sites which are outside of a federation (outside clients), how to guarantee the generalizability of an FL model? This is typically a domain generalization problem (Zhou et al., 2023; Wang et al., 2022) or a test-time adaptation problem (*i.e.*, using inference samples as a clue of the unseen distribution to facilitate adaptation) (He et al., 2021; Varsavsky et al., 2020). Currently, there are a few works that introduce domain generalization into federated learning (Jiang et al., 2023; Liu et al., 2021a). In the future, evaluating and enhancing the generalizability of a trained FL model to unseen sites or even unseen classes (*i.e.*, open-set recognition (Geng et al., 2020; Qin et al., 2022)) will be a promising research direction.

7.2.4. Weakly-Supervised Learning for FL

Weakly-supervised learning is a promising technique that handles data with incomplete, inexact, and inaccurate labels. These problems are common and widespread in medical imaging data. How to deal with these "imperfect" data (*e.g.*, learning from noisy labels (Karimi et al., 2020)) in an FL system is worthy of further exploration.

7.2.5. FL Security: Attack and Defense

Several existing FL systems have been shown to be vulnerable to inside or outside attacks, concerning system robustness and data privacy (Lyu et al., 2022). Further exploration of strong defense strategies in FL is helpful to enhance the security of FL systems. Another interesting question is: if an institution wants to withdraw from a federation, how to guarantee its data has been removed from the trained FL model? One solution is the data auditing technique (Huang et al., 2022a) which can also be used to check if a poisoned/suspicious dataset is used in FL training.

7.2.6. Blockchain and Decentralization of FL

Most existing FL methods on medical tasks employ a centralized paradigm which demands a trustworthy central server. This pattern gradually shows many disadvantages such as vulnerability to poisonous attacks and lack of credibility. Recently, blockchain has been identified as a potentially promising solution to this problem (Zhu et al., 2023a). Using blockchain can avoid the dependence on the central server which can be the bottleneck of the whole federation. Some work has made efforts on this point for medical image analysis through leveraging blockchain (Kumar et al., 2021b; Noman et al., 2023) or other decentralization method (Roy et al., 2019). Currently, very limited work has been performed in this direction for medical image analysis, thus, there is much room for future research.

7.2.7. FL for Medical Video Analysis

Most existing FL systems focus on combining cross-site medical images. As an extension of 2D/3D medical images, medical videos have been rarely explored. Some pioneering work has employed FL to effectively take advantage of medical video from multiple sites/datasets for surgical phase recognition (Kassem et al., 2022). In the future, FL systems consisting of medical videos for surgical or other applications will attract more research attention.

7.2.8. Large-Scale Medical Image Benchmark for FL

Most existing medical image databases for FL research only consist of relatively small datasets at each client. Some work just split a single large dataset (*e.g.*, CheXpert (Irvin et al., 2019)) into different parts which are simulated as different client datasets. There is a lack of large-scale federations which consist of various sites across the world. Only a few works have leveraged real-world datasets from multiple cities or countries. Li et al. (2022) collected chest X-ray images from different cities for the task of COVID-19 detection. Roth et al. (2020) leverage seven clinical institutions from across the world to build a federated learning model for breast density classification. Dayan et al. (2021) builds a large-scale federation through international cooperation. Building large-scale benchmarks (including publicly available medical imaging databases and state-of-the-art FL algorithms) through extensive international cooperation is very beneficial for real-world FL applications.

8. Conclusion

In this paper, we review the recent advances in federated learning (FL) for medical image analysis. We summarize existing FL methods from a system view and categorize them into client-end, server-end, and client-server communication methods. For each category, we provide a novel "question-answer" paradigm to elaborate on the motivation and mechanism of different FL methods in medical image analysis. We also introduce existing software tools/platforms and benchmark medical image datasets that have been used for federated learning. In addition, we conduct an experiment to empirically compare different FL methods on a popular benchmark imaging database (*i.e.*, ADNI). We further discuss current challenges, potential research opportunities, and future directions of FL-based medical image analysis.

We hope that this survey paper could provide researchers with a clear picture of the recent development of FL in medical image analysis and that more research efforts can be inspired and initiated in this exciting research field.

References

Adnan, M., Kalra, S., Cresswell, J.C., Taylor, G.W., Tizhoosh, H.R., 2022. Federated learning and differential privacy for medical image analysis. Scientific Reports 12, 1953.

Agbley, B.L.Y., Li, J.P., Haq, A.U., Bankas, E.K., Mawuli, C.B., Ahmad, S., Khan, S., Khan, A.R., 2023. Federated fusion of magnified histopathological images for breast tumor classification in the internet of medical things. IEEE Journal of Biomedical and Health Informatics .

Alkhunaizi, N., Kamzolov, D., Takávc, M., Nandakumar, K., 2022. Suppressing poisoning attacks on federated learning for medical imaging, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 673–683.

Andreux, M., du Terrail, J.O., Beguier, C., Tramel, E.W., 2020. Siloed federated learning for multi-centric histopathology datasets, in: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Springer. pp. 129–139.

Antunes, R.S., André da Costa, C., Küderle, A., Yari, I.A., Eskofier, B., 2022. Federated learning for healthcare: Systematic review and architecture proposal. ACM Transactions on Intelligent Systems and Technology (TIST) 13, 1–23.

Aouedi, O., Sacco, A., Piamrat, K., Marchetto, G., 2023. Handling privacy-sensitive medical data with federated learning: Challenges and future directions. IEEE Journal of Biomedical and Health Informatics 27, 790–803.

Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., et al., 2021. Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica 83, 242–256.

Bdair, T., Navab, N., Albarqouni, S., 2021. FedPerl: semi-supervised peer learning for skin lesion classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 336–346.

Bernard, O., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Transactions on Medical Imaging 37, 2514–2525.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konevcnỳ, J., Mazzocchi, S., McMahan, B., et al., 2019. Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems 1, 374–388.

Budrionis, A., Miara, M., Miara, P., Wilk, S., Bellika, J.G., 2021. Benchmarking PySyft federated learning framework on MIMIC-III dataset. IEEE Access 9, 116869–116878.

Bzdok, D., Eickenberg, M., Grisel, O., Thirion, B., Varoquaux, G., 2015. Semi-supervised factored logistic regression for high-dimensional neuroimaging data. Advances in Neural Information Processing Systems 28.

California Consumer Privacy Act (CCPA), 2018. CCPA. https://oag.ca.gov/privacy/ccpa .

Campello, V.M., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. IEEE Transactions on Medical Imaging 40, 3543–3554.

Cancer Genome Atlas Research Network, Weinstein, J., Collisson, E., Mills, G., Shaw, K., Ozenberger, B., Ellrott, K., Shmulevich, I., Sande, C., Stuart, J., 2013. The cancer genome atlas pan-cancer analysis project. Nature Genetics 45, 1113–1120.

Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition 77, 329–353.

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H., 2022. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Medical Image Analysis 75, 1–15.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in Neural Information Processing Systems 33, 12546–12558.

Chakravarty, A., Kar, A., Sethuraman, R., Sheet, D., 2021. Federated learning for site aware chest radiograph screening, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1077–1081.

Chen, Z., Yang, C., Zhu, M., Peng, Z., Yuan, Y., 2022. Personalized retrogress-resilient federated learning toward imbalanced medical data. IEEE Transactions on Medical Imaging 41, 3663–3674.

Chen, Z., Zhu, M., Yang, C., Yuan, Y., 2021. Personalized retrogress-resilient framework for real-world medical federated learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 347–356.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis 54, 280–296.

Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al., 2020. Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8, 132665–132676.

Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.S., et al., 2021. Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine 27, 1735–1743.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 248–255.

Di Martino, A., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry 19, 659–667.

Dinsdale, N.K., Jenkinson, M., Namburete, A.I., 2022. Fedharmony: Unlearning scanner bias with distributed data, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 695–704.

Divya, R., Shantha Selva Kumari, R., 2021. Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification. Neural Computing and Applications 33, 8435–8444.

Dong, N., Kampffmeyer, M., Voiculescu, I., 2022. Learning underrepresented classes from decentralized partially labeled medical images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 67–76.

Dong, N., Voiculescu, I., 2021. Federated contrastive learning for decentralized unlabeled medical images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 378–387.

Dwork, C., Roth, A., et al., 2014. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science 9, 211–407.

Elmas, G., Dar, S.U., Korkmaz, Y., Ceyani, E., Susam, B., Ozbey, M., Avestimehr, S., Cukur, T., 2022. Federated learning of generative image priors for MRI reconstruction. IEEE Transactions on Medical Imaging .

Fan, Z., Su, J., Gao, K., Hu, D., Zeng, L.L., 2021. A federated deep learning framework for 3D brain MRI images, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–6.

Feki, I., Ammar, S., Kessentini, Y., Muhammad, K., 2021. Federated learning for COVID-19 screening from chest X-ray images. Applied Soft Computing 106, 107330.

Feng, C.M., Yan, Y., Wang, S., Xu, Y., Shao, L., Fu, H., 2022. Specificity-preserving federated learning for MR image reconstruction. IEEE Transactions on Medical Imaging .

Flanders, A.E., et al., 2020. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiology: Artificial Intelligence 2, 1–8.

Foley, P., Sheller, M.J., Edwards, B., Pati, S., Riviera, W., Sharma, M., Moorthy, P.N., Wang, S.h., Martin, J., Mirhaji, P., et al., 2022. OpenFL: The open federated learning library. Physics in Medicine & Biology 67, 214001.

Frénay, B., Verleysen, M., 2013. Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems 25, 845–869.

Geiping, J., Bauermeister, H., Dröge, H., Moeller, M., 2020. Inverting gradients-how easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems 33, 16937–16947.

General Data Protection Regulation, 2019. GDPR. `https://gdpr-info.eu/` .

Geng, C., Huang, S.j., Chen, S., 2020. Recent advances in open set recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 3614–3631.

Guan, H., Liu, M., 2022. Domain adaptation for medical image analysis: A survey. IEEE Transactions on Biomedical Engineering 69, 1173–1185.

Guo, P., Wang, P., Zhou, J., Jiang, S., Patel, V.M., 2021. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2423–2432.

Gürler, Z., Rekik, I., 2022. Federated brain graph evolution prediction using decentralized connectivity datasets with temporally-varying acquisitions. IEEE Transactions on Medical Imaging .

Hatamizadeh, A., Yin, H., Molchanov, P., Myronenko, A., Li, W., Dogra, P., Feng, A., Flores, M.G., Kautz, J., Xu, D., et al., 2023. Do gradient inversion attacks make federated learning unsafe? IEEE Transactions on Medical Imaging .

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738.

He, Y., Carass, A., Zuo, L., Dewey, B.E., Prince, J.L., 2021. Autoencoder based self-supervised test-time adaptation for medical image analysis. Medical Image Analysis 72, 102136.

Hosseini, S.M., Sikaroudi, M., Babaie, M., Tizhoosh, H., 2023. Proportionally fair hospital collaborations in federated learning of histopathology images. IEEE Transactions on Medical Imaging .

Huang, Y., Huang, C.Y., Li, X., Li, K., 2022a. A dataset auditing method for collaboratively trained machine learning models. IEEE Transactions on Medical Imaging .

Huang, Z.A., Hu, Y., Liu, R., Xue, X., Zhu, Z., Song, L., Tan, K.C., 2022b. Federated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans. IEEE Transactions on Biomedical Engineering .

Irvin, J., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 590–597.

Islam, M., Reza, M.T., Kaosar, M., Parvez, M.Z., 2022. Effectiveness of federated learning and CNN ensemble architectures for identifying brain tumors using MRI images. Neural Processing Letters , 1–31.

Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging 27, 685–691.

Jiang, M., Wang, Z., Dou, Q., 2022. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1087–1095.

Jiang, M., Yang, H., Cheng, C., Dou, Q., 2023. Iop-fl: Inside-outside personalization for federated medical image segmentation. IEEE Transactions on Medical Imaging .

Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al., 2021. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning 14, 1–210.

Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima Jr, I., Mancuso, J., Jungmann, F., Steinborn, M.M., et al., 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence 3, 473–484.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis 65, 101759.

Kassem, H., Alapatt, D., Mascagni, P., AI4SafeChole, C., Karargyris, A., Padoy, N., 2022. Federated cycling (fedcy): Semi-supervised federated learning of surgical phases. IEEE Transactions on Medical Imaging .

Ke, J., Shen, Y., Lu, Y., 2021. Style normalization in histology with federated learning, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 953–956.

Kholod, I., Yanaki, E., Fomichev, D., Shalugin, E., Novikova, E., Filippov, E., Nordlund, M., 2020. Open-source federated learning frameworks for IoT: A comparative review and analysis. Sensors 21, 167.

Knoll, F., Zbontar, J., Sriram, A., Muckley, M.J., Bruno, M., Defazio, A., Parente, M., Geras, K.J., Katsnelson, J., Chandarana, H., et al., 2020. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. Radiology: Artificial Intelligence 2, e190007.

Kouw, W.M., Loog, M., 2019. A review of domain adaptation without target labels. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 766–785.

Kumar, A., Purohit, V., Bharti, V., Singh, R., Singh, S.K., 2021a. Medisecfed: private and secure medical image classification in the presence of malicious clients. IEEE Transactions on Industrial Informatics 18, 5648–5657.

Kumar, R., Khan, A.A., Kumar, J., Golilarz, N.A., Zhang, S., Ting, Y., Zheng, C., Wang, W., et al., 2021b. Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. IEEE Sensors Journal 21, 16301–16314.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Li, D., Kar, A., Ravikumar, N., Frangi, A.F., Fidler, S., 2020a. Federated simulation for medical imaging, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 159–168.

Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., He, B., 2021. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering , 1–20.

Li, T., Sahu, A.K., Talwalkar, A., Smith, V., 2020b. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine 37, 50–60.

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

Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., et al., 2019. Privacy-preserving federated brain tumour segmentation, in: Machine Learning in Medical Imaging:

10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10, Springer. pp. 133–141.

Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S., 2020d. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Medical Image Analysis 65, 1–14.

Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z., 2020e. On the convergence of fedavg on non-iid data, in: Proceedings of International Conference on Learning Representations, pp. 1–12.

Li, Y., Wang, N., Shi, J., Hou, X., Liu, J., 2018. Adaptive batch normalization for practical domain adaptation. Pattern Recognition 80, 109–117.

Li, Z., Xu, X., Cao, X., Liu, W., Zhang, Y., Chen, D., Dai, H., 2022. Integrated cnn and federated learning for COVID-19 detection on chest X-ray images. IEEE/ACM Transactions on Computational Biology and Bioinformatics .

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: European Conference on Computer Vision, Springer. pp. 740–755.

Linardos, A., Kushibar, K., Walsh, S., Gkontra, P., Lekadir, K., 2022. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. Scientific Reports 12, 3551.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88.

Litjens, G., Toth, R., Van De Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Medical Image Analysis 18, 359–373.

Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A., 2021a. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1013–1023.

Liu, Q., Yang, H., Dou, Q., Heng, P.A., 2021b. Federated semi-supervised medical image classification via inter-client relation matching, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 325–335.

Liu, X., Li, W., Yuan, Y., 2022. Intervention & interaction federated abnormality detection with noisy clients, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 309–319.

Liu, Z., Xu, J., Peng, X., Xiong, R., 2018. Frequency-domain dynamic pruning for convolutional neural networks. Advances in Neural Information Processing Systems 31.

Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F., Chen, T.Y., Mahmood, F., 2022. Federated learning for computational pathology on gigapixel whole slide images. Medical Image Analysis 76, 1–13.

Luo, J., Wu, S., 2022. Fedsld: Federated learning with shared label distribution for medical image classification, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q., Philip, S.Y., 2022. Privacy and robustness in federated learning: Attacks and defenses. IEEE Transactions on Neural Networks and Learning Systems .

Malekzadeh, M., Hasircioglu, B., Mital, N., Katarya, K., Ozfatura, M.E., Gündüz, D., 2021. Dopamine: Differentially private federated learning on medical data, in: The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-21), pp. 1–9.

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR. pp. 1273–1282.

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, 1993–2024.

Miller, K.L., et al., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nature Neuroscience 19, 1523–1536.

Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707–6717.

Miyato, T., Maeda, S.i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 1979–1993.

Muckley, M.J., Riemenschneider, B., Radmanesh, A., Kim, S., Jeong, G., Ko, J., Jun, Y., Shin, H., Hwang, D., Mostapha, M., et al., 2021. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. IEEE Transactions on Medical Imaging 40, 2306–2317.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clinics 15, 869–877.

Nguyen, D.C., Pham, Q.V., Pathirana, P.N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., Hwang, W.J., 2022. Federated learning for smart healthcare: A survey. ACM Computing Surveys (CSUR) 55, 1–37.

Noman, A.A., Rahaman, M., Pranto, T.H., Rahman, R.M., 2023. Blockchain for medical collaboration: A federated learning-based approach for multi-class respiratory disease classification. Healthcare Analytics , 100135.

Peng, L., Wang, N., Dvornek, N., Zhu, X., Li, X., 2022. Fedni: Federated graph learning with network inpainting for population-based disease prediction. IEEE Transactions on Medical Imaging .

Pfitzner, B., Steckhan, N., Arnrich, B., 2021. Federated learning in a medical context: A systematic literature review. ACM Transactions on Internet Technology (TOIT) 21, 1–31.

Qayyum, A., Ahmad, K., Ahsan, M.A., Al-Fuqaha, A., Qadir, J., 2022. Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge. IEEE Open Journal of the Computer Society 3, 172–184.

Qi, X., Yang, G., He, Y., Liu, W., Islam, A., Li, S., 2022. Contrastive re-localization and history distillation in federated CMR segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 256–265.

Qin, Z., Yang, L., Gao, F., Hu, Q., Shen, C., 2022. Uncertainty-aware aggregation for federated open set domain adaptation. IEEE Transactions on Neural Networks and Learning Systems .

Qiu, L., Cheng, J., Gao, H., Xiong, W., Ren, H., 2023. Federated semi-supervised learning for medical image segmentation via pseudo-label denoising. IEEE Journal of Biomedical and Health Informatics .

Qu, L., Balachandar, N., Zhang, M., Rubin, D., 2022. Handling data heterogeneity with generative replay in collaborative learning for medical imaging. Medical Image Analysis 78, 102424.

Quellec, G., Cazuguel, G., Cochener, B., Lamard, M., 2017. Multiple-instance learning for medical image and video analysis. IEEE Reviews in Biomedical Engineering 10, 213–234.

Rahman, K.J., Ahmed, F., Akhter, N., Hasan, M., Amin, R., Aziz, K.E., Islam, A.M., Mukta, M.S.H., Islam, A.N., 2021. Challenges, applications and design aspects of federated learning: A survey. IEEE Access 9, 124682–124700.

Rajendran, S., Obeid, J.S., Binol, H., Foley, K., Zhang, W., Austin, P., Brakefield, J., Gurcan, M.N., Topaloglu, U., 2021. Cloud-based federated learning implementation across medical centers. JCO Clinical Cancer Informatics 5, 1–11.

Raudys, S.J., Jain, A.K., et al., 1991. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 252–264.

van Ravesteijn, V.F., van Wijk, C., Vos, F.M., Truyen, R., Peters, J.F., Stoker, J., van Vliet, L.J., 2009. Computer-aided detection of polyps in CT colonography using logistic regression. IEEE Transactions on Medical Imaging 29, 120–131.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al., 2020. The future of digital health with federated learning. NPJ digital medicine 3, 119.

Roth, H.R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., Gupta, S., Qu, L., Ihsani, A., Bizzo, B.C., et al., 2020. Federated learning for breast density classification: A real-world implementation, in: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Springer. pp. 181–191.

Roth, H.R., et al., 2021. Federated whole prostate segmentation in MRI with personalized neural architectures, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 357–366.

Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C., 2019. Braintorrent: A peer-to-peer environment for decentralized federated learning. arXiv:1905.06731 .

Satariano, A., 2019. Google is fined $57 million under Europes data privacy law. The New York Times 21.

Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S., 2019. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Springer. pp. 92–104.

Silva, S., Altmann, A., Gutman, B., Lorenzi, M., 2020. Fedbiomed: A general open-source frontend framework for federated learning in healthcare, in: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Springer. pp. 201–210.

Smith, V., Chiang, C.K., Sanjabi, M., Talwalkar, A.S., 2017. Federated multi-task learning. Advances in Neural Information Processing Systems 30.

Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G., 2022. Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems .

Stripelis, D., Ambite, J.L., Lam, P., Thompson, P., 2021. Scaling neuroscience research using federated learning, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1191–1195.

T Dinh, C., Tran, N., Nguyen, J., 2020. Personalized federated learning with moreau envelopes. Advances in Neural Information Processing Systems 33, 21394–21405.

Tan, A.Z., Yu, H., Cui, L., Yang, Q., 2022. Towards personalized federated learning. IEEE Transactions on Neural Networks and Learning Systems .

du Terrail, J.O., Leopold, A., Joly, C., Béguier, C., Andreux, M., Maussion, C., Schmauch, B., Tramel, E.W., Bendjebbar, E., Zaslavskiy, M., et al., 2023. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nature Medicine 29, 135–146.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5, 1–9.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–289.

US Department of Health and Human Services, 2020. HIPAA. `https://www.hhs.gov/hipaa/index.html` .

Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. PLOS ONE 14, 1–20.

Van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Machine Learning 109, 373–440.

Varsavsky, T., Orbes-Arteaga, M., Sudre, C.H., Graham, M.S., Nachev, P., Cardoso, M.J., 2020. Test-time unsupervised domain adaptation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, October 4–8, 2020, Springer. pp. 428–436.

Wachinger, C., et al., 2016. Domain adaptation for Alzheimer's disease diagnostics. NeuroImage 139, 470–479.

Wagner, N., Fuchs, M., Tolkach, Y., Mukhopadhyay, A., 2022. Federated stain normalization for computational pathology, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 14–23.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P., 2022. Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering .

Wang, L., Lin, Z.Q., Wong, A., 2020. Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Scientific Reports 10, 1–12.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106.

Wicaksana, J., Yan, Z., Yang, X., Liu, Y., Fan, L., Cheng, K.T., 2022a. Customized federated learning for multi-source decentralized medical image classification. IEEE Journal of Biomedical and Health Informatics 26, 5596–5607.

Wicaksana, J., Yan, Z., Zhang, D., Huang, X., Wu, H., Yang, X., Cheng, K.T., 2022b. Fedmix: Mixed supervised federated learning for medical image segmentation. IEEE Transactions on Medical Imaging .

Wilson, G., Cook, D.J., 2020. A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) 11, 1–46.

Wu, Y., Zeng, D., Wang, Z., Shi, Y., Hu, J., 2021. Federated contrastive learning for volumetric medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 367–377.

Wu, Y., Zeng, D., Wang, Z., Shi, Y., Hu, J., 2022. Distributed contrastive learning for medical image segmentation. Medical Image Analysis 81, 102564.

Xia, G., Chen, J., Yu, C., Ma, J., 2023. Poisoning attacks in federated learning: A survey. IEEE Access 11, 10708–10722.

Xiao, J., Yu, L., Zhou, Z., Bai, Y., Xing, L., Yuille, A., Zhou, Y., 2022. Catenorm: Categorical normalization for robust medical image segmentation, in: Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings, Springer. pp. 129–146.

Xu, A., Li, W., Guo, P., Yang, D., Roth, H.R., Hatamizadeh, A., Zhao, C., Xu, D., Huang, H., Xu, Z., 2022. Closing the generalization gap of cross-silo federated medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20866–20875.

Xu, X., Deng, H.H., Gateno, J., Yan, P., 2023. Federated multi-organ segmentation with inconsistent labels. IEEE Transactions on Medical Imaging .

Yan, R., Qu, L., Wei, Q., Huang, S.C., Shen, L., Rubin, D., Xing, L., Zhou, Y., 2023. Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. IEEE Transactions on Medical Imaging .

Yan, Z., Wicaksana, J., Wang, Z., Yang, X., Cheng, K.T., 2020. Variation-aware federated learning with multi-source decentralized medical image data. IEEE Journal of Biomedical and Health Informatics 25, 2615–2628.

Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H.R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., et al., 2021a. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Medical Image Analysis 70, 101992.

Yang, J., Shi, R., Ni, B., 2021b. MedMNIST classification decathlon: A lightweight automl benchmark for medical image analysis, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 191–195.

Yang, Q., Liu, Y., Chen, T., Tong, Y., 2019. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 10, 1–19.

Yang, Q., Zhang, J., Hao, W., Spell, G.P., Carin, L., 2021c. Flop: Federated learning on medical datasets using partial networks, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 3845–3853.

Yang, X., Song, Z., King, I., Xu, Z., 2022. A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering .

Yin, H., Mallya, A., Vahdat, A., Alvarez, J.M., Kautz, J., Molchanov, P., 2021a. See through gradients: Image batch recovery via gradinversion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16337–16346.

Yin, X., Zhu, Y., Hu, J., 2021b. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys (CSUR) 54, 1–36.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y., 2021a. A survey on federated learning. Knowledge-Based Systems 216, 106775.

Zhang, M., Qu, L., Singh, P., Kalpathy-Cramer, J., Rubin, D.L., 2022. SplitAVG: A heterogeneity-aware federated deep learning method for medical imaging. IEEE Journal of Biomedical and Health Informatics 26, 4635–4644.

Zhang, W., Zhou, T., Lu, Q., Wang, X., Zhu, C., Sun, H., Wang, Z., Lo, S.K., Wang, F.Y., 2021b. Dynamic-fusion-based federated learning for COVID-19 detection. IEEE Internet of Things Journal 8, 15884–15891.

Zhang, Y., Yang, Q., 2021. A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering 34, 5586–5609.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C., 2023. Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 4396–4415.

Zhou, Z.H., 2018. A brief introduction to weakly supervised learning. National Science Review 5, 44–53.

Zhu, J., Cao, J., Saxena, D., Jiang, S., Ferradi, H., 2023a. Blockchain-empowered federated learning: Challenges, solutions, and future directions. ACM Computing Surveys 55, 1–31.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.

Zhu, L., Liu, Z., Han, S., 2019. Deep leakage from gradients. Advances in Neural Information Processing Systems 32.

Zhu, M., Chen, Z., Yuan, Y., 2023b. FedDM: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting. IEEE Transactions on Medical Imaging .

Zhu, W., Luo, J., 2022. Federated medical image analysis with virtual sample synthesis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 728–738.

Ziller, A., Trask, A., Lopardo, A., Szymkow, B., Wagner, B., Bluemke, E., Nounahon, J.M., Passerat-Palmbach, J., Prakash, K., Rose, N., et al., 2021. Pysyft: A library for easy federated learning. Federated Learning Systems: Towards Next-Generation AI , 111–139.