Pattern Recognition 151 (2024) 110424

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Federated learning for medical image analysis: A survey

Hao Guan [a], Pew-Thian Yap [a], Andrea Bozoki [b], Mingxia Liu [a,*]

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

ARTICLE INFO

Keywords:
Federated learning
Machine learning
Medical image analysis
Data privacy

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/centers 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 have systematically gathered research papers on federated learning and its applications in medical image analysis published between 2017 and 2023. Our search and compilation were conducted using databases from IEEE Xplore, ACM Digital Library, Science Direct, Springer Link, Web of Science, Google Scholar, and PubMed. In this survey, we first introduce the background of federated learning for dealing with privacy protection and collaborative learning issues. 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.

1. Introduction

Medical image analysis has been greatly pushed forward by computer vision and machine learning [1–4]. The remarkable success of modern machine learning methods, *e.g.*, deep learning [5], can be attributed to the building and release of grand-scale natural image databases, such as ImageNet [6] and Microsoft Common Objects in Context (MS COCO) [7]. Unlike natural image analysis, the field of medical image analysis still faces the fundamental challenge of the "small-sample-size" problem [8,9].

Based on small sample data, it is difficult for us to estimate real data distributions, greatly hindering the building of robust and reliable machine 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/centers is intractable due to strict privacy protection policies such as Health Insurance Portability and Accountability Act (HIPAA) [10] and General Data Protection Regulation (GDPR) [11]. For example, the United States HIPAA has rigidly restricted the exchange of personal health data and images [10]. Thus, directly sharing and pooling medical images across different sites/centers 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 [12–14] 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, federated learning 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.

* Corresponding author.
E-mail address: mingxia_liu@med.unc.edu (M. Liu).

https://doi.org/10.1016/j.patcog.2024.110424
Received 16 October 2023; Received in revised form 5 March 2024; Accepted 9 March 2024
Available online 12 March 2024

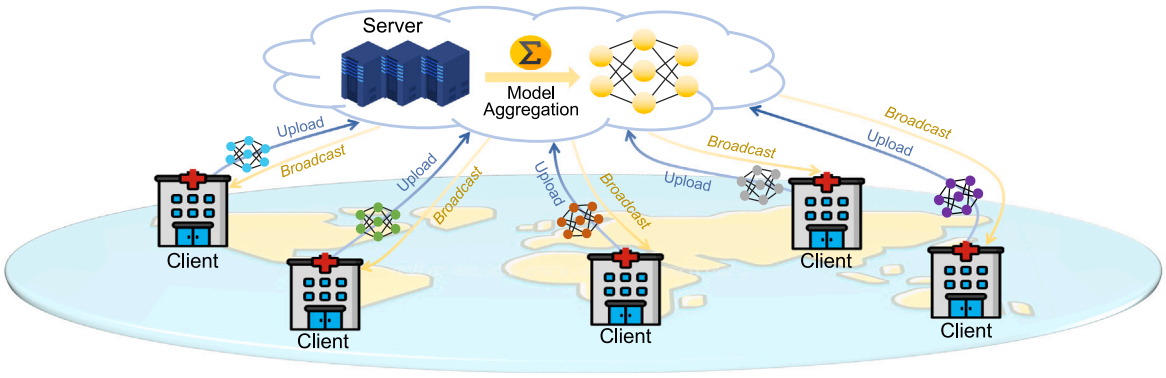

Fig. 1. Overview of federated learning (FL) for medical image analysis, including a server and multiple clients. Each selected client trains a model on its local dataset. The server collects the local models and calculates a global model that is broadcast to all the selected clients for deployment.

1.1. Related surveys

There have been several survey papers on federated learning [15–20], 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 [21–25]. However, some of them focus on electronic health records [21,22] or internet of medical things [26], without paying attention to medical imaging. And some survey papers cover very broad areas in medicine and healthcare applications [23,24], without detailed introduction on federated learning in medical image analysis. A recent survey also reviews the application of federated learning on medical image analysis [27]. Our survey paper reviews and discusses the most recent advances in federated learning for medical image analysis and has significant differences from the previous one in the following aspects.

- **Different Coverage.** The previous review was limited over the time period before December 2022. Our paper covers the papers from 1 January 2017 to 31 October 2023. With a broader coverage, most recent advances of the state-of-the-art models or methods in federated learning in medical image analysis (*e.g.*, transformer) have been included in our paper.
- **Software Platforms.** The previous survey does not include federated learning software platforms that have been applied to medical image analysis. Our paper emphasizes the implementation of federated learning techniques for medical image analysis. Specifically, we introduce some new and influential FL software platforms and benchmark medical imaging datasets for federated learning research in medical imaging.
- **Experimental Study.** The previous survey only conducts a survey and summary of published papers. As for our work, besides a summary of existing work, we also conduct an experimental study to evaluate typical federated learning methods for medical image analysis empirically. This could offer the readers a more intuitive understanding of this research topic.
- **Future Direction and New Arisen Problems.** Due to the inclusion of the most recent papers, our survey paper offers a more comprehensive summary of newly arisen research problems (*e.g.*, model generalizability for unseen clients, and FL for medical video analysis) and points out a broader range of future directions of federated learning for medical image analysis.
- **Different Perspective and Organization.** Different from previous surveys that are based on multiple research issues in federated learning, 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. This categorization can be more intuitive and clear to picture federated learning. 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 it helps provide more insights into different methods.

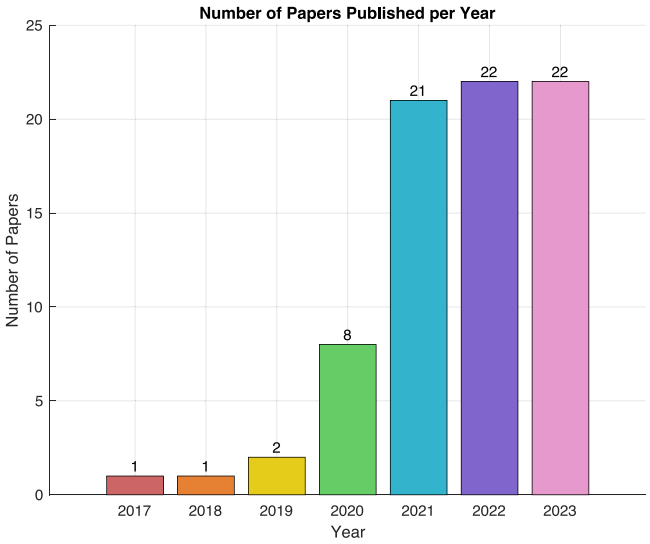

Fig. 2. Overview of the number of papers (in terms of published years) that have been collected for this survey on federated learning in medical image analysis.

1.2. Searching and analysis process

The first paper on federated learning was released in the year of 2016, thus the searching process for this survey ranges from 1st January 2017 to 31st October 2023 (see Fig. 2). There are three steps in conducting this survey paper. ***First***, we performed a literature search using the academic databases and engines, including (1) IEEE Xplore, (2) ACM Digital Library, (3) Science Direct, (4) Springer Link, (5) Web of Science, (6) Google Scholar and (7) PubMed. ***Second***, we refined the initial result from the digital libraries by removing duplicated papers and papers that do not have close relationship with medical imaging (*e.g.*, non-imaging healthcare data). ***Third***, we analyzed the refined papers, extracted the common research questions and technical solutions, and constructed this survey paper.

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

Patient data protection has become an important issue in the digital era. Using and selling patient data has many negative implications [28]. Thus, many governments have introduced tough new laws and regulations on privacy data protection, such as the CCPA in the United States [29] and GDPR in Europe [11]. Collecting, sharing, and processing of personal data are strictly constrained, and violating these laws and regulations may face high-cost penalties [30]. With these strict restrictions from laws, medical images, one of the most important privacy information, cannot be easily shared among different sites/centers. 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 in real-world practice.

2.1.2. Challenges of medical image analysis

The conventional approach to training machine learning models in medical image analysis involves utilizing data from a single site or center. However, this method is usually subject to limited sample size. It is common that there are very limited number of images in local datasets. This situation often arises due to the high costs associated with imaging and labeling procedures. Consequently, the datasets suffer from the "small-sample-size" problem [8,9]. This issue can severely impact the learning performance of machine learning models, leading to suboptimal results that lack statistical significance. Another significant concern is the inherent bias in the distribution of data collected from a specific site or center. These datasets may not accurately represent the true data distribution, thereby introducing bias into machine learning models. For instance, it is common to encounter unbalanced data in medical sites, where the number of healthy subjects significantly outweighs that of patients. Such imbalances can skew the model's predictions and compromise its effectiveness in real-world applications. In addition, medical image datasets collected from a specific site often reflect the characteristics and demographics of the local patient population. Consequently, models trained solely on such data may fail to capture the variability present in broader patient cohorts or diverse clinical settings. 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 $\{D_1, D_2, \ldots, D_N\}$, 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.

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* (FedAvg) algorithm proposed by McMahan et al. [12]. 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. We summarize a typical workflow of federated learning as follows.

(1) **Client Selection and Initialization.** The server selects a set of clients that meet certain requirements. For example, a medical site/center might only check in to the server when it can correctly get access to the intranet of a federation with relatively good bandwidth. Some recent FL models dynamically select clients that meet certain requirements such as training efficiency [31] or anomaly score [32]. A global model is initialized on the central server. This model serves as the starting point for training across different medical sites/centers (*i.e.*, clients).

(2) **Local Training.** The global model is sent to all the participating medical sites/centers. Each site/center trains a machine learning model (*e.g.*, U-Net) on its local medical imaging data using certain optimization methods (*e.g.*, stochastic gradient descent). With the development of artificial intelligence, some recent work introduced more advanced models for client training such as vision transformer [33]. Since the data never leaves its original location, this process can enhance privacy and security.

(3) **Model Upload.** After local training, each medical site/center calculates the updates to the model (*e.g.*, gradients or model changes) and sends/uploads these updates back to the central server. Importantly, only model updates are shared, not the data itself.

(4) **Aggregation.** The central server aggregates all the updates uploaded by the clients, typically using certain algorithms that ensure a fair and effective combination of the different contributions. This aggregation results in an updated global model. While classic FL systems use equal weights for aggression, some recent models explore using more adaptive weighting strategies [34] to enhance training efficiency.

(5) **Broadcast.** During the broadcasting step, the server sends the updated model parameters or gradients to the clients, enabling them to perform local computations and contribute to the collaborative model training process. By efficiently distributing model updates, the broadcasting step facilitates synchronized model updates across the federated network while minimizing communication overhead. Research on this topic has focused on optimizing communication protocols and minimizing communication overhead [31,35] while ensuring efficient dissemination of model updates.

(6) **Iteration and Convergence.** The above steps are repeated for several iterations. With each round, the global model becomes more refined and accurate, as it learns from a diverse set of data sources. This process continues until the model reaches a satisfactory level of accuracy or a predefined number of iterations are completed. Recent research work focuses on improving the overall training efficiency and accelerate the convergence [36].

(7) **Deployment.** The final global model is then deployed for use in applications, maintaining the benefits of having learned from a wide and diverse set of data sources. In real-world practice, several factors or challenges need to be considered such as compatibility with existing hospital systems, integration challenges, and user adoption hurdles.

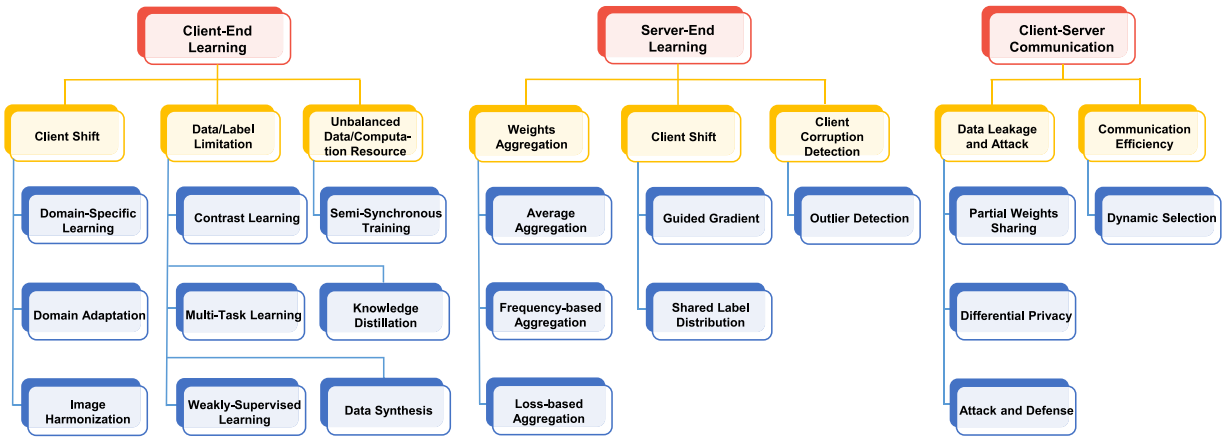

Fig. 3. Overview of federated learning (FL) methods for medical image analysis.

2.4. Types of federated learning

2.4.1. Horizontal federated learning

Horizontal federated learning [17], also known as homogeneous federated learning, is characterized by data distribution across different entities that share the same feature space but have different samples. In the context of medical image analysis, this can be thought of as different medical institutions holding medical imaging data (*e.g.*, MRIs, X-rays) of different patients.

Examples in Medical Image Analysis. Consider multiple hospitals across different regions participating in a study to improve the diagnosis of lung diseases using chest X-rays. Each hospital has its own set of patient data (images), but the features extracted from these images (*e.g.*, lung size, shape, and texture) are similar across all datasets. Horizontal FL allows these hospitals to collaboratively train a model to diagnose lung diseases more accurately without sharing the actual patient data.

2.4.2. Vertical federated learning

Vertical federated learning [17], or heterogeneous federated learning, occurs when different entities possess different feature sets for the same samples. In medical imaging, this translates to different institutions having different types of data (*e.g.*, omics data, demographic information, and imaging data) for the same set of patients.

Examples in Medical Image Analysis. Vertical FL is increasingly prevalent in medical imaging studies due to the multidisciplinary nature of healthcare data. For instance, a hospital might have imaging data, while a research lab could hold genomic data for the same set of patients. Through vertical FL, these diverse datasets can be utilized to create more comprehensive models for disease diagnosis and prognosis, without compromising patient privacy.

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/deep learning methods can be plugged and integrated into an FL framework. For example, a U-Net [37] can be used in each client for medical image segmentation and is trained in a federated manner. Federated learning is concerned with multiple issues such as data, machine learning models, privacy protection mechanisms, and communication architecture. As shown in Fig. 3, 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 (3) communication methods. In each group, different methods are clustered according to the specific research problems they aim to address which will be elaborated in the following sections.

3.2. Client-end learning

3.2.1. Client end: Domain shift among clients

Problem. This research addresses the challenge of significant cross-site data distribution variance in medical imaging, often resulting from varying scanning settings and diverse subject populations across different sites. The focus is on developing strategies to mitigate this variance's adverse impact on the accuracy and reliability of FL model training.

In practice, multi-site medical images may have significantly different data distributions, which is the well-known "domain shift" problem [3] (also referred to as "client shift" in an FL system). As shown in Fig. 4, images from different imaging sites have significantly different intensity distributions. When projected in the feature space, the domain shift may negatively influence machine learning performance. Thus certain techniques, *e.g.*, domain adaptation, are adopted to alleviate this issue by making the distribution differences smaller. 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. This method uses client data to locally fine-tune the global model and alleviate negative influence of client shift. This method is also known as customized/personalized FL [39,40]. Feng et al. [41] propose an encoder–decoder structure within an FL framework for magnetic resonance (MR) image reconstruction. A 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 [39,42]. Chakravarty et al. [43] propose a framework that combines a Convolutional 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. [44] 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. [45] propose to train a locally adapted model that accumulates both global gradients

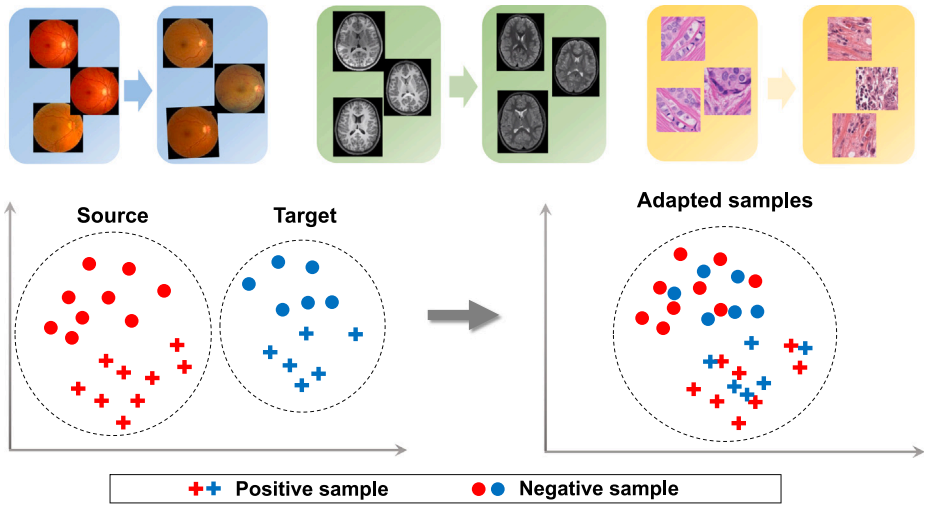

Fig. 4. Domain shift among different medical sites (domains). Domain adaptation aims to reduce domain differences and enhance machine learning performance across different sites.
Source: Image courtesy to Guan et al. [38].

(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. [46] build an FL framework based on a 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. [47] 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. This method uses domain adaptation to reduce medical data distribution differences of clients. Domain adaptation is a sound machine learning technique that has been widely used in medical image analysis [3]. It aims to reduce domain shift (in terms of certain distances) among different medical image datasets and enhance the generalizability of a machine learning model. Many medical-related FL studies resort to domain adaptation for improved performance. Li et al. [48] use domain adaptation to align distribution differences of functional MRI data among clients. In their method, data in each client are added with noise to enhance privacy protection. A domain discriminator/classifier is trained on these data with noises to reduce domain shift. Dinsdale et al. [49] propose a domain adaptation-based FL framework to remove domain shift among clients caused by different scanners. In their framework, medical 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. [50] leverages batch normalization (BN) in a deep neural network to handle client (histopathology datasets) shift. Guo et al. [51] 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. This method typically uses image-to-image translation models to harmonize the medical images of different clients. After harmonization, all the medical images are expected to have similar styles, thus reducing domain shift. Qu et al. [52] propose a generative replay strategy to handle data heterogeneity among clients. They first train an auxiliary variational autoencoder (VAE) to generate medical images that 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. [53] employ cycle-GAN [54] to minimize the variations of medical images among clients. One client/site with low data complexity is selected as a reference, and then cycleGAN is used to harmonize medical images from other clients to the reference site. Jiang et al. [55] propose a frequency-based harmonization method to reduce client shift in medical images. Medical images are firstly transformed into frequency domain and phase components are kept locally, while the average amplitudes from each client are shared and then normalized to harmonize all the client medical images.

3.2.2. *Client end: Limited data and labels*

Problem. This research tackles the prevalent issue in medical imaging where datasets are frequently small-sized and deficient in label information. The focus is on developing strategies to mitigate their negative impact on FL model training and prevent biased results.

In real-world practice, there are often limited medical images in one client/site, and labeled medical 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 FL in medical image analysis.

(1) Contrast Learning. Contrastive learning [56–58] is a self-supervised learning method where models learn to distinguish between similar and dissimilar data points. A model trained with contrast learning can provide good initialization for further fine-tuning (with a few labeled data) on downstream tasks. [59,60] use contrast learning to pretrain the encoder of a U-Net in each client, then the global U-Net is fine-tuned with limited labeled medical images. In this way, the negative influence caused by the shortage of labeled medical images can be largely reduced. Similar strategies can be found in [61].

(2) Multi-Task Learning. Multi-task learning [62] is an effective learning paradigm for data augmentation. Smith et al. [63] propose a novel optimization framework, *i.e.*, MOCHA, which extends classic multi-task learning in the federated environment. Huang et al. [64] build a federated multi-task framework in which several related tasks, *i.e.*, attention-deficit/hyperactivity disorder (ADHD), autism spectrum

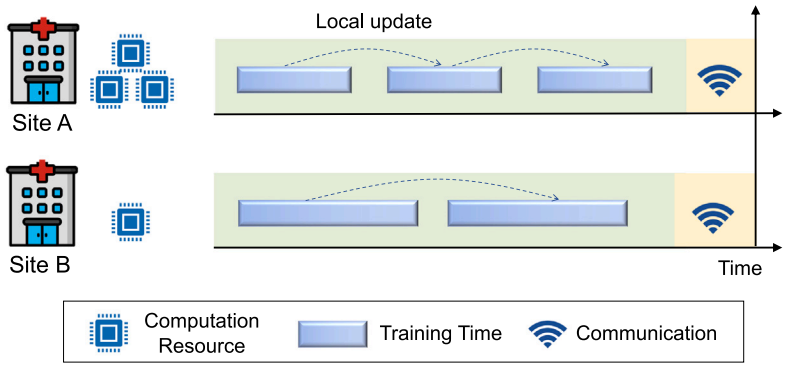

Fig. 5. Different local updates for clients with different computation and data resources.

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 [65] is an extensive group of methods that train a model under weak supervision, including (1) Incomplete supervision [66,67]; (2) Inexact supervision [68,69]; and (3) Inaccurate supervision [70,71]. Yang et al. [72] introduce semi-supervised learning into FL for chest CT segmentation. In their method, unlabeled CT images are leveraged to assist the federated training. For unlabeled CT images in a client, the global model assigns them pseudo labels. Meanwhile, it also outputs predictions on augmented data of the original unlabeled images. A consistency loss is utilized on these predictions to further adjust the global model weights. Lu et al. [73] 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. [74] 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.

(4) Knowledge Distillation. Knowledge distillation [75] is a process where a smaller, more efficient model (the "student") is trained to replicate the behavior of a larger, more complex model (the "teacher"). This is achieved by using the outputs of the teacher model as a guide for training the student model, effectively transferring the knowledge. Kumar et al. [76] leverage knowledge distillation for COVID-19 detection in 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 alleviates the demand for large data during the FL training process. He et al. [77] use knowledge distillation to address the problem of weakly-supervised learning in heterogeneous 3D MR knee images. Unlabeled data in the client is used to distill knowledge from the large-scale national data repository to improve the performance of the collaborative model.

(5) Data Synthesis. This method typically uses generative models (*e.g.*, GAN) to create synthesized medical images as data enhancement for federated learning. Zhu et al. [78] propose an FL framework with virtual sample synthesis for medical image analysis. Given an image $\mathbf{x}$ in the client, the authors first use Virtual Adversarial Training [79] to generate synthetic images that are similar to $\mathbf{x}$, and then use all the synthesized images for local model training. Chang et al. [80] present a novel GAN-based synthetic learning approach for extracting information from each client to generate a homogeneous dataset with entirely synthetic medical images for downstream applications. Peng et al. [81] propose a federated graph learning framework for brain disease prediction, where a Graph Convolutional Network (GCN) is used as the learning model in each client. 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: Heterogeneous environments (computation resource & data scale)

Problem. This research addresses the disparity in data volumes and computational resources, such as GPU availability, across various medical imaging sites/centers. The focus is on developing strategies to minimize its impact on federated training effectiveness.

In the standard FL algorithm (*i.e.*, FegAvg), each client model conducts a predefined number of training epochs (with equal batches or learning rates) before reaching a synchronization time point when it shares its model with the federation. Different medical sites, however, often have significantly different computation resources and amounts of data (images), which may lead to slow convergence of FL model optimization. For example, each medical site (client) is supposed to conduct 50 epochs' updates before model uploading. A site with advanced GPUs may take 1 s, while a site with weak computation utility may take 100 s. Thus, the stronger client will have to spend 99 s waiting for weight sharing. This will slow down the convergence of the overall federation. Aiming at handling this issue in medical imaging analysis tasks, Stripelis et al. [36] propose a Semi-Synchronous Training strategy in federated learning and apply it to brain age prediction. As shown in Fig. 5, 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. This research seeks effective strategies for aggregating client weights in a federated learning system, aiming to ensure consistent performance and avoid performance degradation following each client–server communication.

Weight aggregation in federated learning plays a crucial role by combining the model updates from multiple decentralized clients to form a single, improved global model. Chen et al. [82] propose a Progressive Fourier Aggregation strategy at the server end. Based on

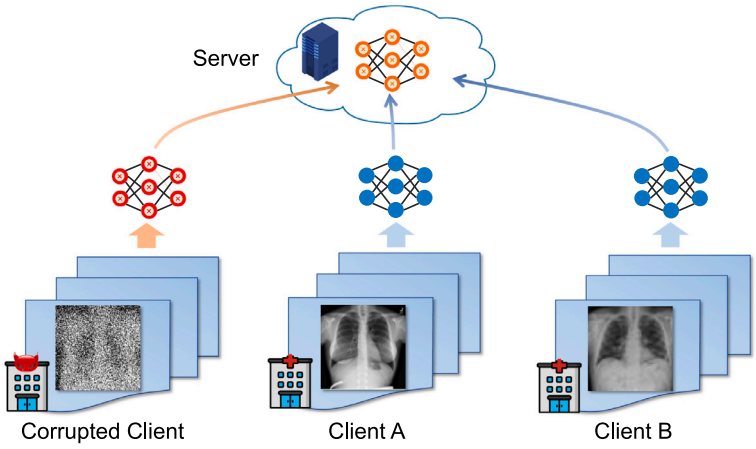

Fig. 6. Corrupted clients will lead to a corrupted global model, thus negatively influencing the entire federated learning system.

previous studies that low-frequency components of parameters form the basis of deep network capability [83], only the low-frequency components are aggregated to share knowledge learned from different clients, while the high-frequency parts are disregarded. Li et al. [34] consider the training loss of each client as the impact factor of the weight aggregation in FL for COVID-19 detection. The client with relatively bad performance caused by uneven image data will get a smaller weight for the global weight aggregation.

3.3.2. Sever end: Domain shift among clients

Problem. This research addresses the issue of domain shift among clients which can lead to non-convergence of federated models. The focus is on developing server-side solutions that effectively tackle these domain discrepancies, ensuring convergence and stability of the federated models.

The client shift can be handled on the server side during the global model optimization process. Hosseini et al. [84] argue that the image heterogeneity between different medical centers (clients) may lead to a biased global model, *i.e.*, a machine learning model that has good performance for some clients while exhibiting inferior performance for other clients. Thus, they propose a revised optimization objective (motivated by fair resource allocation approaches in wireless network research), to facilitate uniform model performance in histopathology image analysis 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. [85] 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 et al. [86] 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 model training in each client, a weighted cross-entropy loss is designed as the batch loss. The weight is computed based on the label distributions in each batch, concerning their label distributions across the entire federation.

3.3.3. Sever end: Client corruption/anomaly detection

Problem. This research investigates how to shield a federated learning system from the impact of clients corrupted by noisy image labels or malicious attacks. The focus is on developing robust mechanisms that identify and avoid those adverse influences, maintaining the performance of the overall system.

Classic FL 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 images or the client is honest without malicious attack. In real-world practice (as shown in Fig. 6), however, a client may be trained with "dirty" medical images that have noisy labels, poor scanning qualities, or suffer from poisoning attacks from malicious parties. How to deal with this issue is critical for ensuring the safety of a medical federated learning system. Alkhunaizi et al. [32] propose a sever-end outlier detection method for medical images, 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 its effectiveness.

3.4. Client–server communication

3.4.1. Data leakage and attack

Problem. This research focuses on developing effective methods to prevent medical image data leakage and privacy violations during server-client interactions in federated learning.

Protection of privacy, *i.e.*, ensuring the medical image data of each client are not seen and accessed by other clients/sever, is the main concern of FL systems. Prior studies have shown that, even without inter-site data sharing, pixel-level images can be reconstructed or recovered by the leaked gradients of a machine learning model [87–89]. Therefore, it is critical to study advanced techniques to proactively avoid data leakage during communication between the server and multiple clients. Many studies have focused on this topic in recent years.

(1) Partial Weights Sharing. Yang et al. [35] argue 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 [90].

(2) Differential Privacy. Gradient information of a deep neural work may contain individual privacy that can be reconstructed by malicious parties. Differential privacy [91] could limit the certainty in inferring an individual's presence in the training dataset. And several recent studies [48,73,92] 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. [93] apply gradients attack [87] 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. [94] design a gradient inversion algorithm to estimate the running statistics (*i.e.*, mean and variance) of batch normalization layers to match the gradients 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

Problem. This research is dedicated to formulating strategies that optimize client–server communication, aiming to accelerate the convergence process and ensure more effective model training.

To improve the communication efficiency during FL training, Zhang et al. [31] propose a dynamic fusion-based FL approach for COVID-19 diagnosis. Their framework dynamically selects the participating clients for weight fusion according to the performance of local client models, and conducts model aggregation based on participating clients' training time. If a client does not upload its 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 for medical image analysis. 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 [95][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 8000 stars and 1900 forks on GitHub,[2] which shows its popularity. Budrionis et al. [96] 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. The 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. [97] 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) [93] 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. [93] 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. [98] use Fed-BioMed to conduct multi-center analysis for structural brain imaging data (MRI) across different datasets and verify its effectiveness.

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 [15,99].

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, as shown in Table 1.

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 [48,100]. 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 [32,43], where each subset is treated as a client dataset.

5.2. Brain images

5.2.1. ADNI

The Alzheimer's Disease Neuroimaging Initiative (ADNI) [101,102] 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 1900+ subjects and 59 centers are provided for analysis and research.

[3] https://github.com/securefederatedai/openfl.
[4] https://www.fets.ai.
[5] https://fedbiomed.gitlabpages.inria.fr.

5.2.2. ABIDE

Autism Brain Imaging Data Exchange (ABIDE) initiative [103] 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) [104] is a benchmark dataset for brain tumor segmentation. BraTS is updated regularly for the Brain Tumor Segmentation Challenge.[6] 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) [105] 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 São Paulo (São Paulo, Brazil).

5.2.5. UK Biobank

UK Biobank [106] is a large-scale, influential biomedical database and research resource containing genetic and health data from half a million participants. As for imaging data, it has four imaging centers and contains valuable brain scans and cardiac MRI information. As a large-scale database with multiple imaging centers, UK Biobank can contribute to varied research areas in medical image analysis, such as federated learning and domain adaptation.

5.2.6. IXI

IXI Dataset[7] 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 [107] 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)[8] 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 [108] 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) [109][9] is a publicly-available database of chest X-ray images, containing 3616 COVID-19 positive cases, 10,192 normal controls, 6012 lung opacity (non-COVID infection), and 1345 viral pneumonia cases.

5.3.4. COVIDx

The COVIDx dataset [110] 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, 8066 normal cases, and 5538 non-COVID-19 pneumonia cases.

5.3.5. ACDC

Automatic Cardiac Diagnosis Challenge (ACDC) [111] is a large publicly available and fully annotated dataset for cardiac MRI assessment. This dataset consists of 150 patients who are divided into 5 categories in terms of well-defined characteristics based on physiological parameters.

5.3.6. M&M

Multi-Center, Multi-Vendor, and Multi-Disease Cardiac Segmentation (M&Ms) Challenge [112][10] 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) [113][11] 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 [114][12] 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)[13] 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) [115] 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) [116][14] is a large-scale landmark cancer genomics database. Whole-slide images for normal controls and cancers are provided for histology and microscopy research.

[6] https://www.med.upenn.edu/cbica/brats.
[7] https://brain-development.org/ixi-dataset.
[8] https://www.kaggle.com/datasets/nih-chest-xrays/data.
[9] https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.
[10] https://www.ub.edu/mnms.
[11] https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000.
[12] https://challenge.isic-archive.com/data.
[13] https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data.
[14] https://www.cancer.gov/ccg/research/genome-sequencing/tcga.

5.5.4. Knee: fastMRI

The fastMRI [117,118][15] is a large-scale dataset for medical image reconstruction using machine learning approaches. This dataset contains more than 1500 knee MRIs (1.5 and 3 T) and DICOM images from 10,000 clinical knee MRIs (1.5 and 3 T).

5.5.5. MedMNIST

MedMNIST [119] is a dataset for medical image classification. Similar to the MNIST dataset,[16] 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 (see Table 1).

6. Experiment

To evaluate federated learning (FL) performance in medical image analysis, we compared various FL approaches using the ADNI dataset [101,102]. This dataset comprises two subsets: ADNI-1, featuring 1.5T T1-weighted MRIs from 428 subjects (199 with Alzheimer's Disease (AD) and 229 normal controls (NCs)), and ADNI-2, with 3.0T T1-weighted MRIs from 360 subjects (159 AD patients and 201 NCs). Duplicate subjects across ADNI-1 and ADNI-2 were excluded for independence. Analysis focused on 90 brain regions-of-interest (ROIs) based on the AAL atlas [162], using mean gray matter volumes as features. For each experiment, 80% of the data was used for training and 20% for testing, with this split performed five times to ensure reliability. Results include mean and standard deviation values to account for variability.

6.1. Experimental setup

The task here is AD vs. NC classification. We use four metrics for performance evaluation, including classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the ROC curve (AUC). Logistic Regression is used as the machine learning model for each FL setting, which has been widely used in medical imaging analysis [163–166]. 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. 7(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. 7(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. 7(c). Note this strategy needs to share data, and thus, could not preserve privacy.

(4) **FedAVG** [12,167]. 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. 7(d). The number of iterations for local model training is set to 10.

(5) **FedSGD** [12]. Each client trains a local model, and 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. 7(d). The number of iterations for local model training is set to 10.

[15] https://fastmri.med.nyu.edu.

[16] http://yann.lecun.com/exdb/mnist.

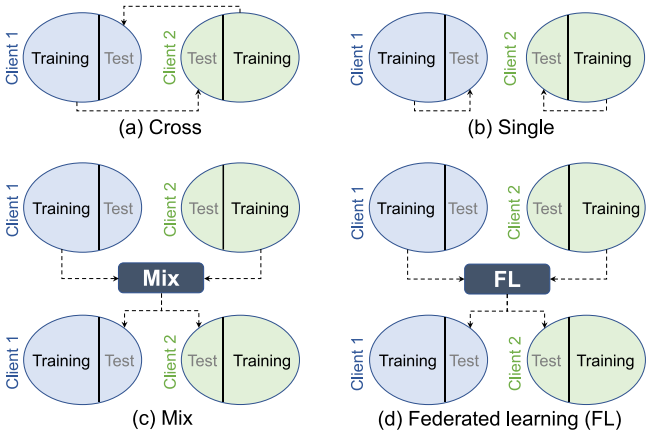

Fig. 7. Different settings for performance comparison.

(6) **FedProx** [168]. 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. 7(d).

6.2. 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 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, 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 federated learning 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.

Table 1
Benchmark datasets for federated learning in medical image analysis.

Reference	Task	Dataset	Modality	Learning model
Brain				
Peng et al. (2022) [81]	ASD, AD classification	ABIDE [103], ADNI [101]	fMRI	GCN
Gürler et al. (2022) [120]	Brain connectivity prediction	OASIS [121]	MRI	GNN
Islam et al. (2022) [122]	Brain tumor classification	UK Data Service [123]	MRI	CNN
Dinsdale et al. (2022) [49]	Age prediction	ABIDE [103]	MRI	CNN (VGG)
Jiang et al. (2022) [124]	Intracranial hemorrhage diagnosis	RSNA [105]	CT	CNN (DenseNet)
Stripelis et al. (2021) [36]	Brain age prediction	UK Biobank [106]	MRI	CNN
Liu et al. (2021) [125]	Intracranial hemorrhage diagnosis	RSNA [105]	CT	CNN (DenseNet)
Fan et al. (2021) [85]	ASD classification	ABIDE [103]	MRI	CNN
Li et al. (2020)[48]	ASD classification	ABIDE [103]	fMRI	MLP
Sheller et al. (2019) [126]	Brain tumor segmentation	BraTS [104]	MRI	U-Net
Li et al. (2019) [90]	Brain tumor segmentation	BraTS [104]	MRI	CNN
Chest				
Hatamizadeh et al. (2023) [94]	Image generation (attack)	COVID CXR [109] ChestX-ray14 [108]	Chest X-ray	CNN (ResNet)
Yan et al. (2023) [33]	Classification	COVID-FL [33]	Chest X-ray	Transformer
Alkhunaizi et al. (2022) [32]	Classification	CheXpert [107]	Chest X-ray	CNN
Dong et al. (2022) [127]	Classification	ChestX-ray14 [108]	Chest X-ray	CNN
Chakravarty et al. (2021) [43]	Classification	CheXpert [107]	Chest X-ray	CNN, GNN
Lung				
Yang et al. (2021) [35]	COVID-19 diagnosis	COVIDx [110]	Chest X-ray	CNN
Feki et al. (2021) [128]	COVID-19 diagnosis	Local dataset	Chest X-ray	CNN
Kumar et al. (2021) [76]	COVID-19 diagnosis	COVID-19 CXR [109]	Chest X-ray	CNN
Dong et al. (2021) [61]	COVID-19 diagnosis	COVID-19 CXR [109]	Chest X-ray	CNN
Yang et al. (2021) [72]	Segmentation	Local dataset	CT	CNN
Heart				
Linardos et al. (2022) [129]	Cardiac diagnosis	ACDC [111], M&M [112]	MRI	CNN
Qi et al. (2022) [124]	Cardiac segmentation	M&M [112], Emidec [130]	MRI	U-Net
Li et al. (2021) [131]	Cardiac image synthesis	Local dataset	CT	GAN
Wu et al. (2021) [59]	Cardiac segmentation	ACDC [111]	MRI	U-Net
Breast				
Agbley et al. (2023) [132]	Breast tumor classification	BreakHis [133]	Pathology	CNN
Wicaksana et al. (2022) [134]	Breast tumor segmentation	BUS [135], BUSIS [136], UDIAT [137]	Ultrasound	U-Net
Skin				
Yan et al. (2023) [33]	Skin lesion classification	ISIC [114]	Dermoscopy	Transformer
Wicaksana et al. (2022) [39]	Skin lesion classification	HAM10000 [113]	Dermoscopy	CNN
Alkhunaizi et al. (2022) [32]	Skin lesion classification	HAM10000 [113]	Dermoscopy	CNN
Jiang et al. (2022) [124]	Skin lesion classification	HAM10000 [113]	Dermoscopy	CNN (DenseNet)
Liu et al. (2021) [125]	Skin lesion classification	HAM10000 [113]	Dermoscopy	CNN (DenseNet)
Bdair et al. (2021) [138]	Skin lesion classification	HAM10000 [113] ISIC19 [139] Derm7pt [140] PAD-UFES [141]	Dermoscopy	CNN (EfficientNet)
Chen et al. (2021) [142]	Skin lesion classification	HAM10000 [113] ISIC17 [143]	Dermoscopy	CNN (VGG)
Eye				
Yan et al. (2023) [33]	Diabetic classification	Retina [144]	Color retinal image	Transformer
Qiu et al. (2023) [145]	Fundus segmentation	RIM-ONE [146]	Color retinal image	CNN (MobileNet)
Wang et al. (2023) [147]	Fundus segmentation	RIF [148]	Color retinal image	Transformer
Qu et al. (2022) [52]	Diabetic classification	Retina [144]	Color retinal image	VAE, CNN
Abdomen				
Zhu et al. (2023) [149]	Prostate segmentation	PROMISE12 [115] NCI-ISBI 2013 [150]	MRI	U-Net
Qiu et al. (2023) [145]	Prostate segmentation	PROMISE12 [115]	MRI	CNN (MobileNet)
Xu et al. (2023) [151]	Tumor segmentation	LiTS [152]	CT	U-Net
Wicaksana et al. (2022) [39]	Cancer classification	ProstateX [153]	MRI	CNN
Luo et al. (2022) [86]	Cancer classification	OrganMNIST [119]	CT	CNN
Liu et al. (2022) [154]	Polyp detection	GLRC [155]	Colonoscopy	CNN
Yan et al. (2021) [53]	Cancer classification	ProstateX [153]	MRI	GAN
Roth et al. (2021) [156]	Prostate segmentation	MSD-Prostate [157] PROMISE12 [115] ProstateX [153] NCI-ISBI 2013 [150]	MRI	U-Net

(*continued on next page*)

Table 1 (*continued*).

Reference	Task	Dataset	Modality	Learning model
Histology				
Hosseini et al. (2023) [84]	Cancer classification	TCGA [116]	Pathology	CNN (DenseNet)
du Terrail et al. (2023) [158]	Cancer classification	Local dataset	Pathology	CNN
Lu et al. (2022) [73]	Cancer classification	TCGA [116]	Pathology	CNN
Adnan et al. (2022) [159]	Cancer classification	TCGA [116]	Pathology	CNN (DenseNet)
Luo et al. (2022) [86]	Cancer classification	PathMNIST [119]	Pathology	CNN
Wagner et al. (2022) [47]	Image harmonization	PESO [160]	Pathology	GAN
Ke et al. (2021) [46]	Image harmonization	TCGA [116]	Pathology	GAN
Others				
Feng et al. (2022) [41]	MRI reconstruction	fastMRI [117] BraTS [104]	MRI	U-Net
Elmas et al. (2022) [161]	MRI reconstruction	fastMRI, BraTS, IXI	MRI	GAN
Guo et al. (2021) [51]	MRI reconstruction	fastMRI, BraTS, IXI	MRI	U-Net

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

7.1.2. Privacy leakage/poisoning attacks

In classic FL, only the model parameters 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 [87–89] and poisoning attacks [169,170]. 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.

7.1.3. Technological limitations

While a majority of research is centered on algorithm design for various medical applications, the practical implementation of FL systems encounters significant technological hurdles. For instance, certain FL algorithms demand substantial computational resources, posing challenges for the underlying hardware infrastructure. Furthermore, addressing communication costs, optimizing network resource allocation, and ensuring synchronization are all formidable obstacles when striving to construct a robust and functional FL system in real life.

7.1.4. Long-term viability of FL-based medical image analysis

Federated Learning is not just a novel machine learning algorithm; it represents a dynamic and systematic approach to engineering. It is imperative to focus on the long-term viability of FL systems when applied to medical image analysis. This involves addressing critical issues such as scalability, sustainability, and evolving regulations. In real-world scenarios, unforeseen challenges emerge, such as clients leaving or joining the training process, as well as unexpected technical and connectivity issues. Effectively managing an FL system for robust and stable long-term medical image analysis is a complex endeavor.

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 [3] has attracted extensive interest. Classic domain adaptation methods typically need access to both source and target domains which may violate the privacy protection restraint in FL. Thus, developing more efficient federated domain adaptation methods will be a promising research direction. Another promising solution is personalized FL techniques [40,171] which utilize local data to further optimize a trained global model.

7.2.2. Multi-modality fusion for federated learning

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. 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 multi-modal medical data [172]. More research work is expected on this topic in the future.

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 (internal clients). An interesting question is: When facing data from unseen sites that are outside of a federation (outside clients), how to guarantee the generalizability of an FL model? This is typically a domain generalization problem [173,174] or a test-time adaptation problem (*i.e.*, using inference samples as a clue of the unseen distribution to facilitate adaptation) [175,176]. Currently, there are a few works that introduce domain generalization into federated learning [45,148]. 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 [177,178]) will be a promising research direction.

7.2.4. Weakly-supervised learning for federated learning

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 [179]) in an FL system is worthy of further exploration.

7.2.5. Federated learning security: Attack and defense

Several existing FL systems have been shown to be vulnerable to internal or external attacks, concerning system robustness and data privacy [169]. 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 [180] which can also be used to check if a poisoned/suspicious dataset is used in FL training.

7.2.6. Blockchain and decentralization of federated learning

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 [181]. 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 [182,183] or other decentralization methods [184]. 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. Federated learning 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 [74]. 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 federated learning

Most existing medical image databases for FL research only consist of relatively small datasets for each client. Some work just split a single large dataset (*e.g.*, CheXpert [107]) into different parts which are simulated as different client datasets. There is a lack of large-scale federations that include various sites across the world. Only a few works have leveraged real-world datasets from multiple cities or countries. Li et al. [34] collect chest X-ray images from different cities for COVID-19 detection. Roth et al. [185] leverage seven clinical institutions from across the world to build a federated learning model for breast density classification. Dayan et al. [100] build 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 beneficial for FL applications in medicine.

7.2.9. Model interpretability

In clinical practice, one of the most significant hurdles in adopting machine learning and AI lies in the "black box" nature of certain machine learning systems, such as deep learning [186]. Even as FL emerges as a promising machine learning prototype, it encounters similar challenges. Thus, the issue of model interpretability remains a critical factor to address for the seamless integration of FL into clinical practice. While some researchers have made an early effort towards this topic [187], more explorations are expected in the future.

7.2.10. Real-world implementation and practical issues

The majority of research on FL in medical imaging has primarily focused on algorithm development and simulation. FL methods, while promising, can encounter difficulties during real-world implementation, such as compatibility with existing hospital systems, integration challenges, and user adoption hurdles. Addressing these practical considerations is crucial for advancing the application of FL in medical image analysis.

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 benchmark medical image datasets that have been used for federated learning. In addition, we conduct an experiment to empirically compare representative FL methods on a popular benchmark imaging database (*i.e.*, ADNI). We further discuss current challenges, potential research opportunities, and future directions of FL in medical image analysis. We hope that this survey paper will 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.

CRediT authorship contribution statement

Hao Guan: Writing – original draft, Visualization, Software, Data curation. **Pew-Thian Yap:** Writing – review & editing. **Andrea Bozoki:** Writing – review & editing. **Mingxia Liu:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported in part by NIH, United States grants AG073297 and EB035160. Part of the data used in this work were obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI). The investigators within ADNI contributed to the design and implementation of ADNI and provided data but did not participate in the analysis or writing of this article.

References

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

[2] V. Cheplygina, M. de Bruijne, J.P. Pluim, Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. 54 (2019) 280–296.

[3] H. Guan, M. Liu, Domain adaptation for medical image analysis: A survey, IEEE Trans. Biomed. Eng. 69 (3) (2022) 1173–1185.

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

[5] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

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

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

[8] S.J. Raudys, A.K. Jain, et al., Small sample size effects in statistical pattern recognition: Recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (3) (1991) 252–264.

[9] A. Vabalas, E. Gowen, E. Poliakoff, A.J. Casson, Machine learning algorithm validation with a limited sample size, PLOS ONE 14 (11) (2019) 1–20.

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

[11] General Data Protection Regulation, GDPR, 2019, https://gdpr-info.eu/.

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

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

[14] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, Found. Trends® Mach. Learn. 14 (1–2) (2021) 1–210.

[15] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, B. He, A survey on federated learning systems: Vision, hype and reality for data privacy and protection, IEEE Trans. Knowl. Data Eng. (2021) 1–20.

[16] T. Li, A.K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Process. Mag. 37 (3) (2020) 50–60.

[17] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Trans. Intell. Syst. Technol. 10 (2) (2019) 1–19.

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

[19] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, Y. Gao, A survey on federated learning, Knowl.-Based Syst. 216 (2021) 106775.

[20] X. Yin, Y. Zhu, J. Hu, A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions, ACM Comput. Surv. 54 (6) (2021) 1–36.

[21] R.S. Antunes, C. André da Costa, A. Küderle, I.A. Yari, B. Eskofier, Federated learning for healthcare: Systematic review and architecture proposal, ACM Trans. Intell. Syst. Technol. 13 (4) (2022) 1–23.

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

[23] D.C. Nguyen, Q.-V. Pham, P.N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, W.-J. Hwang, Federated learning for smart healthcare: A survey, ACM Comput. Surv. 55 (3) (2022) 1–37.

[24] B. Pfitzner, N. Steckhan, B. Arnrich, Federated learning in a medical context: A systematic literature review, ACM Trans. Internet Technol. (TOIT) 21 (2) (2021) 1–31.

[25] N. Rieke, J. Hancox, W. Li, F. Milletari, H.R. Roth, S. Albarqouni, S. Bakas, M.N. Galtier, B.A. Landman, K. Maier-Hein, et al., The future of digital health with federated learning, NPJ Digit. Med. 3 (1) (2020) 119.

[26] O. Aouedi, A. Sacco, K. Piamrat, G. Marchetto, Handling privacy-sensitive medical data with federated learning: Challenges and future directions, IEEE J. Biomed. Health Inf. 27 (2) (2023) 790–803.

[27] M.F. Sohan, A. Basalamah, A systematic review on federated learning in medical image analysis, IEEE Access 11 (2023) 28628–28644.

[28] V. Chiruvella, A.K. Guddati, et al., Ethical issues in patient data ownership, Interact. J. Med. Res. 10 (2) (2021) e22269.

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

[30] A. Satariano, Google is fined $57 million under Europe's data privacy law, N.Y. Times 21 (2019).

[31] W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S.K. Lo, F.-Y. Wang, Dynamic-fusion-based federated learning for COVID-19 detection, IEEE Internet Things J. 8 (21) (2021) 15884–15891.

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

[33] R. Yan, L. Qu, Q. Wei, S.-C. Huang, L. Shen, D. Rubin, L. Xing, Y. Zhou, Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging, IEEE Trans. Med. Imaging (2023).

[34] Z. Li, X. Xu, X. Cao, W. Liu, Y. Zhang, D. Chen, H. Dai, Integrated CNN and federated learning for COVID-19 detection on chest X-ray images, IEEE/ACM Trans. Comput. Biol. Bioinform. (2022).

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

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

[37] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, October 5-9, 2015, Springer, 2015, pp. 234–241.

[38] H. Guan, M. Liu, DomainATM: Domain adaptation toolbox for medical data analysis, Neuroimage 268 (2023) 1–12.

[39] J. Wicaksana, Z. Yan, X. Yang, Y. Liu, L. Fan, K.-T. Cheng, Customized federated learning for multi-source decentralized medical image classification, IEEE J. Biomed. Health Inf. 26 (11) (2022) 5596–5607.

[40] C. T Dinh, N. Tran, J. Nguyen, Personalized federated learning with moreau envelopes, Adv. Neural Inf. Process. Syst. 33 (2020) 21394–21405.

[41] C.-M. Feng, Y. Yan, S. Wang, Y. Xu, L. Shao, H. Fu, Specificity-preserving federated learning for MR image reconstruction, IEEE Trans. Med. Imaging 42 (7) (2023).

[42] M. Zhang, L. Qu, P. Singh, J. Kalpathy-Cramer, D.L. Rubin, SplitAVG: A heterogeneity-aware federated deep learning method for medical imaging, IEEE J. Biomed. Health Inf. 26 (9) (2022) 4635–4644.

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

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

[45] M. Jiang, H. Yang, C. Cheng, Q. Dou, IOP-FL: Inside-outside personalization for federated medical image segmentation, IEEE Trans. Med. Imaging 42 (7) (2023).

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

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

[48] X. Li, Y. Gu, N. Dvornek, L.H. Staib, P. Ventola, J.S. Duncan, Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results, Med. Image Anal. 65 (2020) 1–14.

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

[50] M. Andreux, J.O. du Terrail, C. Beguier, E.W. Tramel, Siloed federated learning for multi-centric histopathology datasets, in: MICCAI Workshop on Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (DART), Lima, Peru, October 4–8, 2020, Springer, 2020, pp. 129–139.

[51] P. Guo, P. Wang, J. Zhou, S. Jiang, V.M. Patel, 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, 2021, pp. 2423–2432.

[52] L. Qu, N. Balachandar, M. Zhang, D. Rubin, Handling data heterogeneity with generative replay in collaborative learning for medical imaging, Med. Image Anal. 78 (2022) 102424.

[53] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, K.-T. Cheng, Variation-aware federated learning with multi-source decentralized medical image data, IEEE J. Biomed. Health Inf. 25 (7) (2020) 2615–2628.

[54] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[55] M. Jiang, Z. Wang, Q. Dou, Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 1087–1095.

[56] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu, Contrastive learning of global and local features for medical image segmentation with limited annotations, Adv. Neural Inf. Process. Syst. 33 (2020) 12546–12558.

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

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

[59] Y. Wu, D. Zeng, Z. Wang, Y. Shi, J. Hu, Distributed contrastive learning for medical image segmentation, Med. Image Anal. 81 (2022) 102564.

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

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

[62] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. 34 (12) (2021) 5586–5609.

[63] V. Smith, C.-K. Chiang, M. Sanjabi, A.S. Talwalkar, Federated multi-task learning, Adv. Neural Inf. Process. Syst. 30 (2017).

[64] Z.-A. Huang, Y. Hu, R. Liu, X. Xue, Z. Zhu, L. Song, K.C. Tan, Federated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans, IEEE Trans. Biomed. Eng. (2022).

[65] Z.-H. Zhou, A brief introduction to weakly supervised learning, Natl. Sci. Rev. 5 (1) (2018) 44–53.

[66] X. Yang, Z. Song, I. King, Z. Xu, A survey on deep semi-supervised learning, IEEE Trans. Knowl. Data Eng. (2022).

[67] J.E. Van Engelen, H.H. Hoos, A survey on semi-supervised learning, Mach. Learn. 109 (2) (2020) 373–440.

[68] G. Quellec, G. Cazuguel, B. Cochener, M. Lamard, Multiple-instance learning for medical image and video analysis, IEEE Rev. Biomed. Eng. 10 (2017) 213–234.

[69] M.-A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: A survey of problem characteristics and applications, Pattern Recognit. 77 (2018) 329–353.

[70] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, IEEE Trans. Neural Netw. Learn. Syst. 34 (11) (2023).

[71] B. Frénay, M. Verleysen, Classification in the presence of label noise: A survey, IEEE Trans. Neural Netw. Learn. Syst. 25 (5) (2013) 845–869.

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

[73] M.Y. Lu, R.J. Chen, D. Kong, J. Lipkova, R. Singh, D.F. Williamson, T.Y. Chen, F. Mahmood, Federated learning for computational pathology on gigapixel whole slide images, Med. Image Anal. 76 (2022) 1–13.

[74] H. Kassem, D. Alapatt, P. Mascagni, C. AI4SafeChole, A. Karargyris, N. Padoy, Federated cycling (FedCy): Semi-supervised Federated Learning of surgical phases, IEEE Trans. Med. Imaging 42 (7) (2023).

[75] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.

[76] A. Kumar, V. Purohit, V. Bharti, R. Singh, S.K. Singh, Medisecfed: private and secure medical image classification in the presence of malicious clients, IEEE Trans. Ind. Inform. 18 (8) (2021) 5648–5657.

[77] X. He, et al., Dealing with heterogeneous 3D MR knee images: A federated few-shot learning method with dual knowledge distillation, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, IEEE, 2023, pp. 1–5.

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

[79] T. Miyato, S.-i. Maeda, M. Koyama, S. Ishii, Virtual adversarial training: a regularization method for supervised and semi-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2018) 1979–1993.

[80] Q. Chang, Z. Yan, M. Zhou, H. Qu, X. He, H. Zhang, L. Baskaran, S. Al'Aref, H. Li, S. Zhang, et al., Mining multi-center heterogeneous medical data with distributed synthetic learning, Nature Commun. 14 (1) (2023) 5510.

[81] L. Peng, N. Wang, N. Dvornek, X. Zhu, X. Li, Fedni: Federated graph learning with network inpainting for population-based disease prediction, IEEE Trans. Med. Imaging 42 (7) (2023).

[82] Z. Chen, C. Yang, M. Zhu, Z. Peng, Y. Yuan, Personalized retrogress-resilient federated learning toward imbalanced medical data, IEEE Trans. Med. Imaging 41 (12) (2022) 3663–3674.

[83] Z. Liu, J. Xu, X. Peng, R. Xiong, Frequency-domain dynamic pruning for convolutional neural networks, Adv. Neural Inf. Process. Syst. 31 (2018).

[84] S.M. Hosseini, M. Sikaroudi, M. Babaie, H. Tizhoosh, Proportionally fair hospital collaborations in federated learning of histopathology images, IEEE Trans. Med. Imaging 42 (7) (2023).

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

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

[87] J. Geiping, H. Bauermeister, H. Dröge, M. Moeller, Inverting gradients-how easy is it to break privacy in federated learning? Adv. Neural Inf. Process. Syst. 33 (2020) 16937–16947.

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

[89] L. Zhu, Z. Liu, S. Han, Deep leakage from gradients, Adv. Neural Inf. Process. Syst. 32 (2019).

[90] W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M.J. Cardoso, et al., 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, 2019, pp. 133–141.

[91] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy, Found. Trends® Theor. Comput. Sci. 9 (3–4) (2014) 211–407.

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

[93] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima Jr., J. Mancuso, F. Jungmann, M.-M. Steinborn, et al., End-to-end privacy preserving deep learning on multi-institutional medical imaging, Nat. Mach. Intell. 3 (6) (2021) 473–484.

[94] A. Hatamizadeh, H. Yin, P. Molchanov, A. Myronenko, W. Li, P. Dogra, A. Feng, M.G. Flores, J. Kautz, D. Xu, et al., Do gradient inversion attacks make federated learning unsafe? IEEE Trans. Med. Imaging (2023).

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

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

[97] P. Foley, M.J. Sheller, B. Edwards, S. Pati, W. Riviera, M. Sharma, P.N. Moorthy, S.-h. Wang, J. Martin, P. Mirhaji, et al., OpenFL: The open federated learning library, Phys. Med. Biol. 67 (21) (2022) 214001.

[98] S. Silva, A. Altmann, B. Gutman, M. Lorenzi, Fed-biomed: 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, 2020, pp. 201–210.

[99] I. Kholod, E. Yanaki, D. Fomichev, E. Shalugin, E. Novikova, E. Filippov, M. Nordlund, Open-source federated learning frameworks for IoT: A comparative review and analysis, Sensors 21 (1) (2020) 167.

[100] I. Dayan, H.R. Roth, A. Zhong, A. Harouni, A. Gentili, A.Z. Abidin, A. Liu, A.B. Costa, B.J. Wood, C.-S. Tsai, et al., Federated learning for predicting clinical outcomes in patients with COVID-19, Nature Med. 27 (10) (2021) 1735–1743.

[101] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, L. Beckett, The Alzheimer's disease neuroimaging initiative, Neuroimaging Clin. 15 (4) (2005) 869–877.

[102] C.R. Jack Jr., M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J. L. Whitwell, C. Ward, et al., The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Reson. Imaging 27 (4) (2008) 685–691.

[103] A. Di Martino, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, Mol. Psychiatry 19 (6) (2014) 659–667.

[104] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[105] A.E. Flanders, et al., Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge, Radiol.: Artif. Intell. 2 (3) (2020) 1–8.

[106] K.L. Miller, et al., Multimodal population brain imaging in the UK Biobank prospective epidemiological study, Nature Neurosci. 19 (11) (2016) 1523–1536.

[107] J. Irvin, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 590–597.

[108] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, 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, 2017, pp. 2097–2106.

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

[110] L. Wang, Z.Q. Lin, A. Wong, Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images, Sci. Rep. 10 (1) (2020) 1–12.

[111] O. Bernard, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11) (2018) 2514–2525.

[112] V.M. Campello, et al., Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge, IEEE Trans. Med. Imaging 40 (12) (2021) 3543–3554.

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

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

[115] G. Litjens, R. Toth, W. Van De Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, et al., Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge, Med. Image Anal. 18 (2) (2014) 359–373.

[116] Cancer Genome Atlas Research Network, J. Weinstein, E. Collisson, G. Mills, K. Shaw, B. Ozenberger, K. Ellrott, I. Shmulevich, C. Sande, J. Stuart, The cancer genome atlas pan-cancer analysis project, Nature Genet. 45 (10) (2013) 1113–1120.

[117] F. Knoll, J. Zbontar, A. Sriram, M.J. Muckley, M. Bruno, A. Defazio, M. Parente, K.J. Geras, J. Katsnelson, H. Chandarana, et al., fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning, Radiol.: Artif. Intell. 2 (1) (2020) e190007.

[118] M.J. Muckley, B. Riemenschneider, A. Radmanesh, S. Kim, G. Jeong, J. Ko, Y. Jun, H. Shin, D. Hwang, M. Mostapha, et al., Results of the 2020 fastMRI challenge for machine learning MR image reconstruction, IEEE Trans. Med. Imaging 40 (9) (2021) 2306–2317.

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

[120] Z. Gürler, I. Rekik, Federated brain graph evolution prediction using decentralized connectivity datasets with temporally-varying acquisitions, IEEE Trans. Med. Imaging 42 (7) (2023).

[121] P.J. LaMontagne, T.L. Benzinger, J.C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A.G. Vlassenko, et al., OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease, MedRxiv (2019) 1–37.

[122] M. Islam, M.T. Reza, M. Kaosar, M.Z. Parvez, Effectiveness of federated learning and CNN ensemble architectures for identifying brain tumors using MRI images, Neural Process. Lett. (2022) 1–31.

[123] C.R. Pernet, K.J. Gorgolewski, D. Job, D. Rodriguez, I. Whittle, J. Wardlaw, A structural and functional magnetic resonance imaging dataset of brain tumour patients, Sci. Data 3 (1) (2016) 1–6.

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

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

[126] M.J. Sheller, G.A. Reina, B. Edwards, J. Martin, S. Bakas, 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, 2019, pp. 92–104.

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

[128] I. Feki, S. Ammar, Y. Kessentini, K. Muhammad, Federated learning for COVID-19 screening from Chest X-ray images, Appl. Soft Comput. 106 (2021) 107330.

[129] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra, K. Lekadir, Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease, Sci. Rep. 12 (1) (2022) 3551.

[130] A. Lalande, et al., Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI, Data 5 (4) (2020) 89.

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

[132] B.L.Y. Agbley, J.P. Li, A.U. Haq, E.K. Bankas, C.B. Mawuli, S. Ahmad, S. Khan, A.R. Khan, Federated fusion of magnified histopathological images for breast tumor classification in the internet of medical things, IEEE J. Biomed. Health Inf. (2023).

[133] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, Breast cancer histopathological image classification using convolutional neural networks, in: 2016 International Joint Conference on Neural Networks, IJCNN, IEEE, 2016, pp. 2560–2567.

[134] J. Wicaksana, Z. Yan, D. Zhang, X. Huang, H. Wu, X. Yang, K.-T. Cheng, FedMix: Mixed supervised federated learning for medical image segmentation, IEEE Trans. Med. Imaging (2023).

[135] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2020) 1–5.

[136] Y. Zhang, M. Xian, H.-D. Cheng, B. Shareef, J. Ding, F. Xu, K. Huang, B. Zhang, C. Ning, Y. Wang, BUSIS: A benchmark for breast ultrasound image segmentation, in: Healthcare, Vol. 10, MDPI, 2022, pp. 1–16.

[137] M.H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A.K. Davison, R. Marti, Automated breast ultrasound lesions detection using convolutional neural networks, IEEE J. Biomed. Health Inform. 22 (4) (2017) 1218–1226.

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

[139] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC), 2019, arXiv preprint arXiv:1902.03368.

[140] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, Seven-point checklist and skin lesion classification using multitask multimodal neural nets, IEEE J. Biomed. Health Inf. 23 (2) (2018) 538–546.

[141] A.G. Pacheco, G.R. Lima, A.S. Salomão, B. Krohling, I.P. Biral, G.G. de Angelo, F.C. Alves Jr., J.G. Esgario, A.C. Simora, P.B. Castro, et al., PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones, Data Brief 32 (2020) 1–10.

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

[143] N.C. Codella, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC), in: 2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018, IEEE, 2018, pp. 168–172.

[144] E. Dugas, J. Jorge, W. Cukierski, Diabetic retinopathy detection, 2015, URL https://kaggle.com/competitions/diabetic-retinopathy-detection.

[145] L. Qiu, J. Cheng, H. Gao, W. Xiong, H. Ren, Federated semi-supervised learning for medical image segmentation via pseudo-label denoising, IEEE J. Biomed. Health Inf. (2023).

[146] F. Fumero, S. Alayón, J.L. Sanchez, J. Sigut, M. Gonzalez-Hernandez, RIM-ONE: An open retinal image database for optic nerve evaluation, in: 24th International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2011, pp. 1–6.

[147] J. Wang, Y. Jin, D. Stoyanov, L. Wang, FedDP: Dual personalization in federated medical image segmentation, IEEE Trans. Med. Imaging (2023).

[148] Q. Liu, C. Chen, J. Qin, Q. Dou, P.-A. Heng, 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, 2021, pp. 1013–1023.

[149] M. Zhu, Z. Chen, Y. Yuan, FedDM: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting, IEEE Trans. Med. Imaging (2023).

[150] NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures, 2021, https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=21267207.

[151] X. Xu, H.H. Deng, J. Gateno, P. Yan, Federated multi-organ segmentation with inconsistent labels, IEEE Trans. Med. Imaging (2023).

[152] P. Bilic, P. Christ, H.B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G.E.H. Mamani, G. Chartrand, et al., The liver tumor segmentation benchmark (lits), Med. Image Anal. 84 (2023) 1–24.

[153] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, H. Huisman, Computer-aided detection of prostate cancer in MRI, IEEE Trans. Med. Imaging 33 (5) (2014) 1083–1092.

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

[155] K. Li, et al., Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations, PLOS ONE 16 (8) (2021) 1–26.

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

[157] A.L. Simpson, et al., A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019, arXiv preprint arXiv:1902.09063.

[158] J.O. du Terrail, A. Leopold, C. Joly, C. Béguier, M. Andreux, C. Maussion, B. Schmauch, E.W. Tramel, E. Bendjebbar, M. Zaslavskiy, et al., Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer, Nature Med. 29 (1) (2023) 135–146.

[159] M. Adnan, S. Kalra, J.C. Cresswell, G.W. Taylor, H.R. Tizhoosh, Federated learning and differential privacy for medical image analysis, Sci. Rep. 12 (1) (2022) 1953.

[160] W. Bulten, et al., Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard, Sci. Rep. 9 (1) (2019) 1–10.
[161] G. Elmas, S.U. Dar, Y. Korkmaz, E. Ceyani, B. Susam, M. Ozbey, S. Avestimehr, T. Cukur, Federated learning of generative image priors for MRI reconstruction, IEEE Trans. Med. Imaging 42 (7) (2023).
[162] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, NeuroImage 15 (1) (2002) 273–289.
[163] R. Divya, R. Shantha Selva Kumari, Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification, Neural Comput. Appl. 33 (14) (2021) 8435–8444.
[164] D. Bzdok, M. Eickenberg, O. Grisel, B. Thirion, G. Varoquaux, Semi-supervised factored logistic regression for high-dimensional neuroimaging data, Adv. Neural Inf. Process. Syst. 28 (2015).
[165] C. Wachinger, et al., Domain adaptation for Alzheimer's disease diagnostics, NeuroImage 139 (2016) 470–479.
[166] V.F. van Ravesteijn, C. van Wijk, F.M. Vos, R. Truyen, J.F. Peters, J. Stoker, L.J. van Vliet, Computer-aided detection of polyps in CT colonography using logistic regression, IEEE Trans. Med. Imaging 29 (1) (2009) 120–131.
[167] X. Li, K. Huang, W. Yang, S. Wang, Z. Zhang, On the convergence of FedAvg on non-IID data, in: Proceedings of International Conference on Learning Representations, 2020, pp. 1–12.
[168] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: Proceedings of Machine Learning and Systems, Vol. 2, 2020, pp. 429–450.
[169] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, S.Y. Philip, Privacy and robustness in federated learning: Attacks and defenses, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[170] G. Xia, J. Chen, C. Yu, J. Ma, Poisoning attacks in federated learning: A survey, IEEE Access 11 (2023) 10708–10722.
[171] A.Z. Tan, H. Yu, L. Cui, Q. Yang, Towards personalized federated learning, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[172] A. Qayyum, K. Ahmad, M.A. Ahsan, A. Al-Fuqaha, J. Qadir, Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge, IEEE Open J. Comput. Soc. 3 (2022) 172–184.
[173] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, C.C. Loy, Domain generalization: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2023) 4396–4415.
[174] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, P. Yu, Generalizing to unseen domains: A survey on domain generalization, IEEE Trans. Knowl. Data Eng. 35 (8) (2023).
[175] Y. He, A. Carass, L. Zuo, B.E. Dewey, J.L. Prince, Autoencoder based self-supervised test-time adaptation for medical image analysis, Med. Image Anal. 72 (2021) 102136.
[176] T. Varsavsky, M. Orbes-Arteaga, C.H. Sudre, M.S. Graham, P. Nachev, M.J. Cardoso, Test-time unsupervised domain adaptation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, October 4–8, 2020, Springer, 2020, pp. 428–436.
[177] C. Geng, S.-j. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3614–3631.
[178] Z. Qin, L. Yang, F. Gao, Q. Hu, C. Shen, Uncertainty-aware aggregation for federated open set domain adaptation, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[179] D. Karimi, H. Dou, S.K. Warfield, A. Gholipour, Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, Med. Image Anal. 65 (2020) 1–19.
[180] Y. Huang, C.-Y. Huang, X. Li, K. Li, A dataset auditing method for collaboratively trained machine learning models, IEEE Trans. Med. Imaging 42 (7) (2023).
[181] J. Zhu, J. Cao, D. Saxena, S. Jiang, H. Ferradi, Blockchain-empowered federated learning: Challenges, solutions, and future directions, ACM Comput. Surv. 55 (11) (2023) 1–31.
[182] R. Kumar, A.A. Khan, J. Kumar, N.A. Golilarz, S. Zhang, Y. Ting, C. Zheng, W. Wang, et al., Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging, IEEE Sens. J. 21 (14) (2021) 16301–16314.
[183] A.A. Noman, M. Rahaman, T.H. Pranto, R.M. Rahman, Blockchain for medical collaboration: A federated learning-based approach for multi-class respiratory disease classification, Healthc. Anal. (2023) 1–16.
[184] A.G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, C. Wachinger, Braintorrent: A peer-to-peer environment for decentralized federated learning, 2019, arXiv:1905.06731.
[185] H.R. Roth, K. Chang, P. Singh, N. Neumark, W. Li, V. Gupta, S. Gupta, L. Qu, A. Ihsani, B.C. Bizzo, et al., Federated learning for breast density classification: A real-world implementation, in: MICCAI Workshop on Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, DART, Springer, 2020, pp. 181–191.
[186] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. 32 (24) (2020) 18069–18083.
[187] A. Li, R. Liu, M. Hu, L.A. Tuan, H. Yu, Towards interpretable federated learning, 2023, arXiv preprint arXiv:2302.13473.

Hao Guan is a postdoc research associate of the Department of Radiology and the Biomedical Research Imaging Center at the University of North Carolina at Chapel Hill, USA. His research field is in machine learning & AI for medicine.

Pew-Thian Yap is a professor of the Department of Radiology and the Director of the Image Analysis Core of the Biomedical Research Imaging Center at the University of North Carolina at Chapel Hill.

Andrea Bozoki is a professor of the Department of Neurology at the University of North Carolina at Chapel Hill. She specializes in memory and cognitive disorders.

Mingxia Liu is with the Department of Radiology and the Biomedical Research Imaging Center at the University of North Carolina at Chapel Hill. Her research interests include machine learning and medical data analysis.