APSIPA Transactions on Signal and Information Processing, 2024, 13, e38

Overview Paper

When Federated Learning Meets Medical Image Analysis: A Systematic Review with Challenges and Solutions

Tian Yang[1*], Xinhui Yu[1], Martin J. McKeown[2] and Z. Jane Wang[1]

[1] *Department of Electrical and Computer Engineering, University of British Columbia, Canada*
[2] *Department of Medicine, University of British Columbia, Canada*

ABSTRACT

Deep learning has been a powerful tool for medical image analysis, but large amount of high-quality labeled datasets are generally required to train deep learning models with satisfactory performance and generalization capability. In medical applications, collecting such large-scale datasets involves specific challenges: data annotation is time-consuming and expert-requisite, and privacy restrictions make it impractical for different institutions to share their own data to construct single large datasets. Federated learning (FL) is an effective method for addressing such concerns since it allows multiple institutions to collaboratively train deep learning models, without sharing individual data samples directly, in line with privacy protection requirements. However, there are numerous challenges when applying FL in medical image analysis, including data heterogeneity and low label quality, that may impede FL from being implemented effectively. This paper conducts a systematic literature review of the challenges and solutions when applying FL in medical image analysis. We present a novel taxonomy of FL-specific challenges in medical image analysis research and summarize representative solutions for these challenges. We anticipate

*Corresponding author: tianyang@ece.ubc.ca

Received 17 July 2024; revised 07 September 2024; accepted 07 September 2024
ISSN 2048-7703; DOI 10.1561/116.20240048

this review will be proved helpful for researchers to have better knowledge of challenges and existing solutions in related fields, and provide inspiration for developing more advanced solutions in the future.

Keywords: Federated learning, deep learning, medical image analysis

1 Introduction

Medical image analysis (MEDIA) is a significant area in global healthcare application and research, as it can be essential for medical diagnosis and guidance of treatment. As an effective and powerful approach that can perform reliable vision analysis, machine learning (ML), especially deep learning (DL) has been widely applied in medical image analysis areas in order to alleviate cumbersome manual work [165]. In order to achieve accurate and robust performance on vision tasks, deep learning techniques usually require large, diverse datasets to train models. However, this brings particular challenges when applying deep learning to medical image analysis tasks, as there are specific and strict restrictions on data privacy and security.

Medical imaging data are extremely sensitive, as they can directly contain identifiable subject information in the header file, and the images themselves may contain sensitive information about personal health information. As data collection and processing resources might be limited in individual medical centers, training a deep learning model with satisfactory performance and generalization ability may require combining data sources from multiple institutes to construct a large enough training dataset. This is often infeasible due to privacy considerations, since medical centers are often not allowed to share their own data samples according to privacy restriction policies, such as the Health Insurance Portability and Accountability Act (HIPAA) [49] in the United States and General Data Protection Regulation (GDPR) [108] in Europe. In order to help multiple medical institutes collaboratively train deep learning models without violating privacy restrictions, federated learning (FL) provides an alternative solution that can train machine learning models utilizing datasets from multiple sources without directly sharing the original data samples. Therefore, FL has become an attractive research and application topic in the medical image analysis area in recent years.

When applied to the medical image analysis area, FL has encountered specific challenges compared to general vision tasks. For example, typical FL algorithms require training datasets on different clients to be statistically independent and identically distributed (iid) [95], but medical image data usually cannot fulfill such requirements since data collected from different

centers can have diverse imaging protocols or focus on different types of diseases, leading to substantial data heterogeneity. Additionally, it can be impractical for each hospital to acquire large amounts of high-quality labeled medical images, as there is a high demand for qualified experts to assist in collecting such data. Moreover, federated deep learning techniques have a high demand for computational and network transmission resources, which can be difficult for medical centers with limited capability.

This review paper aims to conduct a thorough literature review of FL research works in the medical image analysis area, especially focusing on discussing specific research challenges in this area and summarizing solutions to such challenges. We present a novel taxonomy of involved research challenges, and conduct summaries and analysis for corresponding solutions. The remaining parts of this paper are organized as follows: The remaining content of Section 1 summarizes the differences between this work and other existing surveys, and the searching and analysis strategies of this survey. Section 2 introduces basic knowledge of FL and its application in medical image analysis. Section 3 presents our review and analysis results for the research challenges and solutions in existing medical imaging FL works. Section 4 proposes potential future directions for the application of FL in medical image analysis.

1.1 Related surveys

We studied the existing surveys since 2022 concerning FL applications in medical domains. Some of these surveys discuss general conditions of FL applied in medical-related areas, not focusing on medical image analysis [21, 22, 124, 7]. Among other surveys that pay attention to medical image analysis, some of them attach the most importance to the applications of FL in medical image analysis tasks, and only have brief discussions on specific challenges and their solutions in these domains [100, 113, 116, 73, 107]. Compared to several surveys that focus on discussing challenges [25, 109, 121, 76], the main difference of our survey is that we propose a hierarchical taxonomy of challenges and conduct a comprehensive analysis on solutions from different perspectives, while the discussions in previous surveys cover a smaller range and are less systematic.

A recent survey by Guan *et al.* [44] has a similar discussion as ours on FL in medical image analysis. Our survey differs from this work in the following aspects:

- *Broader and newer coverage:* Guan *et al.* [44] discussed 77 papers published up to October 2023, while our survey covers 130 papers published up to May 2024, where 105 papers focusing on research challenges are discussed in detail. As our survey has a broader and more up-to-date coverage of research papers, more advanced works are included and

some new topics involved with novel challenges are discussed here. For instance, multi-modality FL recently began to be applied in medical domains, which was mentioned as one future direction in Guan *et al.* [44], while this topic is discussed in detail in our paper; and we also include the topic of fairness, which has became a growing interest in medical FL applications.

- *Different perspectives:* Guan *et al.* [44] classified challenges in medical imaging FL into three categories according to different components of FL architecture: client-end, server-end, and client-server communication. Instead, we employ a different perspective according to the source and essence of different challenges: We first divide challenges into (1) data heterogeneity, (2) low label quality, (3) attack and defense, (4) communication burden, and (5) underexplored challenges. Then, in each category, we perform detailed descriptions and discussions. Such an organization leads to a systematic taxonomy of challenges and solutions, which is illustrated in Section 3. The number of papers covered in our survey per year is summarized in Table 1.

Table 1: Number of covered papers focusing on challenges per year.

Year	2020	2021	2022	2023	2024
Number of papers	2	10	14	57	22

1.2 Searching and analysis process

We conducted the following steps to collect and study research papers related to our topic:

- *Paper collection:* We collected the initial set of research papers by searching the following databases and search engines: (1) ACM Digital Library, (2) arXiv, (3) Elsevier, (4) Google Scholar, (5) IEEE Xplore, (6) PubMed, (7) SpringerLink, with the search term 'federated learning medical image analysis' 'federated learning medical imaging'. After this step, we built a raw corpus of 352 papers.

- *Corpus refinement:* To refine the raw paper corpus, we first removed duplicate papers among different datasets, and then filtered out those papers that did not focus on medical image analysis, such as the papers analyzing temporal medical data. After this step, we retained a collection of 130 papers. Additionally, as this paper aims to provide a comprehensive study on challenges and solutions in medical imaging FL research, we also discounted papers that just simply applied federated learning techniques

in medical image analysis tasks without emphasis on any challenges. After this step, we got a refined corpus of 105 papers.

- *Challenge analysis:* We assessed the papers in the refined corpus to analyze the research challenges they worked on. We summarized these challenges and proposed a taxonomy for common challenges in the medical imaging FL research. For each type of research challenge, we provided a brief formalization and summarized its existing solutions.

2 Background

2.1 Federated Learning

Federated learning (FL) is a type of distributed machine learning framework, where multiple institutes collaboratively train a machine learning model in a data-privacy-preserving pattern. Privacy protection is accomplished by keeping sensitive datasets locally at each institute (i.e., clients) and not sharing them with other entities, while collaborative model training is achieved by exchanging and aggregating model parameters or gradients, usually with the help of a central server that can communicate with the clients.. The most popular federated learning architecture, FedAvg, was first proposed by McMahan *et al.* [95]. Typical FedAvg can be formalized as follows, which is illustrated in Figure 1. Suppose there are multiple institutes considered as clients and a central server. Each client holds a dataset that cannot be shared with other entities. The central server is responsible for communicating with the clients and helping them collaboratively train a machine-learning model. The typical FedAvg setting assumes that each client shares the same machine-learning model with the server. The goal of federated learning is to obtain a global model whose performance is better than models trained locally at each client using their own dataset. In order to achieve this goal, federated learning typically performs several communication rounds for training, where each round executes the following steps:

- *Initialization*: A set of clients are selected to participate in this training round. The server sends its global model parameters to initialize client models.
- *Local training*: Each client trains the local model with its local dataset for several epochs.
- *Global aggregating*: Each client sends its new model parameters or accumulated gradients to the server, and the server aggregates these updates from all participating clients to obtain a new global model. For

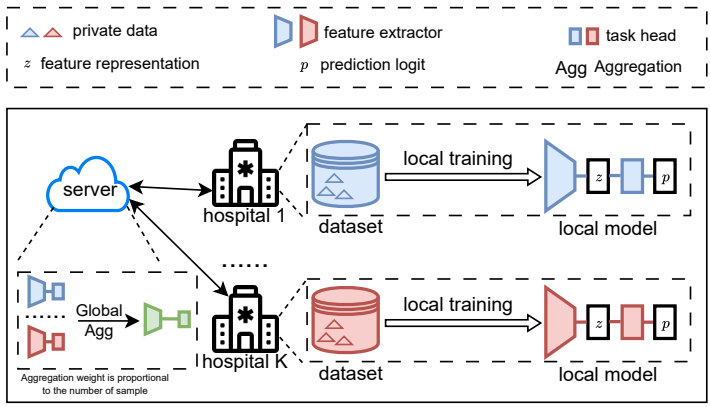

Figure 1: An illustration of the typical FedAvg architecture. During a local training stage, each client (hospital) trains local machine learning model with its own private data for several epochs. At each federated communication round, a central server receives parameters of local models from clients, aggregates the parameters to obtain a global model, and sends the global model to clients, which are used to re-initialize local models for a new local training stage. Such iteration is repeated until the global model converges or after specific number of communication rounds.

typical FedAvg, it aggregates client updates weighted by the number of samples on each client, namely using the following function:

$$w = \sum_{k=1}^{K} \frac{n_k}{n} w_k \tag{1}$$

where w is the aggregated global model parameters or gradients, K is the number of participating clients, n_k is the number of data samples on client k, $n = \sum_{i=1}^{K} n_i$ is the total number of data samples across all participating clients, w_k is the model update from client k at the end this round.

The above procedure will continue until the global model converges or after a specific number of communication rounds. During the entire training process, each client retains its training dataset locally to ensure that no sensitive information is transferred, thereby protecting privacy.

2.2 Federated Learning Taxonomy

According to Yang *et al.* [151], federated learning architecture can be divided into three categories according to the heterogeneity of sample space and feature

space across clients: Horizontal federated learning, vertical federated learning, and federated transfer learning. Figure 2 shows the differences of these three categories of FL.

- *Horizontal federated learning (HFL)*. HFL assumes that the datasets of the clients have different sample spaces but share the same feature space. This is the most common type of FL architecture in medical image analysis research. A typical setting of HFL in medical applications is that different hospitals collect the same type of data, such as chest X-ray images, brain MRIs, etc., from different groups of subjects. It is natural to apply FedAvg and its variant algorithms in HFL scenarios, where the entire training procedure can be regarded as training a strong model with an enriched global dataset by combining the datasets from different clients while protecting subject privacy by not directly exposing local data samples outside each client.

- *Vertical federated learning (VFL)*: VFL happens when the client datasets share the same sample space but have different feature spaces. Such a scenario is relatively rare in medical image analysis FL research. Representative scenarios of VFL in the medical image analysis area occur in multi-modality learning. For example, some work such as Yan *et al.* [148] explores such settings in MRI reconstruction tasks that different clients have different modalities of MRI data collected from the same group of subjects. Such vertical data samples can be used to align the training of these clients in order to achieve better global performance.

- *Federated Transfer Learning (FTL)*: Theoretically, FTL considers the scenario where the sample space and feature space are both different among clients, according to the definition in Yang *et al.* [151]. Such a setting requires extracting common latent knowledge from different clients and utilizing this common knowledge to facilitate client learning. Research strictly following such a setting is quite limited in medical image analysis areas. However, transfer learning methods that aim to share useful knowledge across clients while preserving privacy are indeed applied in some medical FL work, which can be viewed in later sections.

2.3 Application of FL in Medical Image Analysis

FL has been widely applied in various medical image analysis tasks. It is quite suitable for medical applications, as medical data usually contains sensitive information of patients which cannot be exposed to other entities. According to our literature survey, FL is mainly applied in medical image analysis tasks including classification, segmentation, and reconstruction. There are also rare

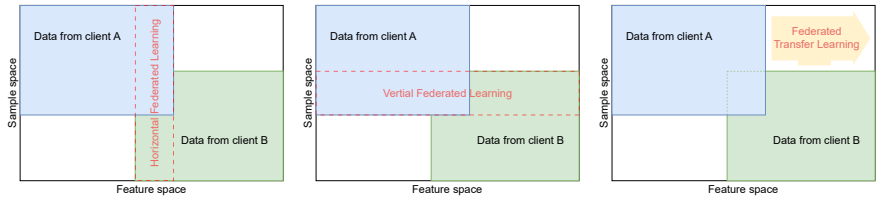

Figure 2: An illustration of differences of HFL, VFL and FTL. Adapted from Yang *et al.* [151].

works that focus on other tasks, such as image synthesis [129, 24] and object detection [153]. Following we introduce the application of FL in these tasks in detail:

- *Classification*. A classification problem can be formalized as follows: Suppose there is a set of N data samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in a sample space $\mathcal{X}$, and each sample $\mathbf{x}_i \in \mathcal{X}$ is associated with a label $y_i \in \mathcal{Y}$, where $\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$ is the label space denoting K different possible classes that the sample belongs to. The goal of the classification task is to learn such a mapping function $f : \mathcal{X} \to \mathcal{Y}$ that the output of this function $\hat{y} = f(\mathbf{x})$ given input $\mathbf{x}$ should be as close as possible to the true label y associated with $\mathbf{x}$. In practice, the label y is often expressed as the one-hot form, namely $y \in \{0, 1\}^K$, where the k-th element of y is 1 while other elements are 0's if sample $\mathbf{x}$ belongs to the k-th class. It can be extended to the multi-label classification problem, where each data sample can belong to multiple classes, namely y can have multiple 1's.

  Classification is one of the most common tasks in medical image analysis. A natural practice is to predict the type of disease given the image of a medical examination. Federated learning has been applied in medical image classification for various diseases. Additionally, its application is not limited to classifying types of diseases but also exists in other medical imaging-related tasks such as surgical phase recognition [65]. Table 2 summarizes common imaging techniques, imaging physical sites, and involved diseases in existing FL works in medical image analysis focusing on classification tasks. Some works used datasets with many classes of diseases, and we list the name and reference of these datasets used by these works instead of listing all classes in order to reduce redundancy.

- *Segmentation*. Image segmentation can be viewed as a special case of classification problem in computer vision, which needs to perform pixel-wise classification given an input image to indicate whether each pixel

Table 2: Example FL applications in medical image classification tasks.

Modality	Body part	Diseases and Papers
X-ray	breast	breast cancer: [27, 59]
	chest	tuberculosis: [81, 80]
		pneumonia: [146, 96, 125, 112, 17, 2, 139, 88, 16]
		ChestX-ray14 (dataset) [134]: [26, 123, 9, 92]
dermatoscopic	skin	pigmented lesions: [26, 157, 141, 155, 146, 39, 131, 155, 135, 31, 141, 5, 88, 92]
WSI	chest	lung cancer: [51]
	prostate	prostate cancer: [68]
	abdominal	colorectal cancer: [45]
	kidney	kidney cancer: [51]
CT	chest	pneumonia: [163, 16]
		lung cancer: [50]
	brain	intracranial hemorrhage: [141, 135]
	abdominal	liver tumor: [96, 168, 112, 41]
		gastric cancer: [34]
		kidney cyst/tumor/stone: [9]
MRI	breast	breast cancer: [27]
	brain	Alzheimer's disease: [158, 9, 71]
		brain tumor: [10]
endoscopy	gastrointestinal	HyperKvasir (dataset) [12]: [168, 31, 142]
fundus camera	eye	diabetic retinopathy: [155, 146, 130]
		glaucoma: [139]
microscope	blood cell	hematologic/oncologic disease: [168, 5, 112, 41]
	breast	breast tumor: [4]
	colon	colorectal cancer: [112, 41]

belongs to a specific segmentation area or background. Therefore, given an input image sample $\mathbf{X} \in \mathbb{R}^{H \times W}$, its associated label is a segmentation mask $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{H \times W}$, where each element in this mask map m_{ij} is a label vector indicating the category of the (i, j)-th pixel.

Medical image segmentation is an important problem in medical image analysis, which significantly helps clinicians identify areas of interest efficiently. Federated learning has been widely applied in the segmentation of images of various diseases obtained from different medical imaging technologies. Table 3 provides a summary of common imaging techniques and imaging physical sites in existing FL works focused on segmentation tasks in medical image analysis.

- *Reconstruction.* Reconstruction is a special topic in medical image analysis, which involves generating high-quality images from raw or incomplete data acquired through various medical imaging techniques, such as Magnetic Resonance Imaging (MRI) and Positron Emission

Table 3: Example FL applications in medical image segmentation tasks.

Modality	Body part	Papers
MRI	knee	[48]
	prostate	[157, 83, 58, 105, 167, 98]
	heart	[154, 94, 93, 52]
	brain	[104, 99, 52, 159]
	spine	[82, 8]
CT	chest	[84, 163, 28, 127]
	abdominal	[160, 55, 66]
X-ray	chest	[138, 74]
dermatoscopic	skin	[84, 140]
endoscopy	abdominal	[128, 167]
fundus camera	eye	[83, 105, 128]

Tomography (PET). The goal of image reconstruction is to produce clean and useful images for medical diagnosis that can accurately represent the internal structures of the subject body. Federated learning has also been applied in medical image reconstruction tasks. Most works focus on reconstruction of brain MRI images [37, 35, 91, 148, 32, 42, 36, 46]. Several works explore PET image denoising, which is a special form of medical image reconstruction [162, 118, 161].

2.3.1 Evaluation Metrics

Here, we summarize the evaluation metrics that are commonly used in the above works of FL in medical image analysis according to our literature study results:

- *Classification*: Typical evaluation metrics for classification tasks are based on these values:
  - TP = True Positives
  - TN = True Negatives
  - FP = False Positives
  - FN = False Negatives

  The most commonly used classification performance evaluation metrics is Accuracy (ACC):

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Some works apply the following Balanced Accuracy (BACC) and F1-Score (F1) in consideration of class imbalance:

$$\text{BACC} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$.
The AUC (Area Under the Receiver Operating Characteristic Curve) is also applied in some works in order to measure the model's classification ability to distinguish among classes:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\, dx \tag{5}$$

where TPR (true positive rate) $= \frac{TP}{TP+FN}$, FPR (false positive rate) $= \frac{FP}{FP+TN}$.

- *Segmentation*: According to our literature study, the following four evaluation metrics are commonly used in FL-based medical image segmentation tasks. Most works pick two from these metrics in their experiments. Generally, Dice similarity coefficient (Dice) and Intersection over Union (IoU) are used to measure the overlap between predicted and ground-truth segmentation masks, while Average Symmetric Surface Distance (ASSD) and Hausdorff Distance (HD) evaluate the boundary error between prediction and ground-truth.

  - *Dice*, measuring the similarity between two sets of data, which is used to measure the overlap between the predicted segmentation and the ground-truth mask:

    $$\text{Dice} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{6}$$

    where A is the set of pixels in the predicted segmentation while B is the set of pixels in the ground-truth mask, "$|\cdot|$" counts the number of elements in a set.
  - *IoU*, another metrics measuring the overlap between segmentation masks:

    $$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

  - *ASSD*, measuring the average distance between surface points of the predicted and ground-truth segmentations:

    $$\text{ASSD} = \frac{1}{|S_A| + |S_B|}\left(\sum_{a \in S_A} \min_{b \in S_B} d(a,b) + \sum_{b \in S_B} \min_{a \in S_A} d(b,a)\right) \tag{8}$$

where S_A and S_B are the sets of boundary points in the predicted and ground-truth segmentations, respectively, $d(a, b)$ is the distance between a point a in S_A and a point b in S_B.

– *HD*, measuring the maximum distance between surface points of the predicted and ground-truth segmentations:

$$\mathrm{HD} = \max\left\{\sup_{a\in S_A}\inf_{b\in S_B} d(a,b), \sup_{b\in S_B}\inf_{a\in S_A} d(b,a)\right\} \tag{9}$$

where sup denotes the least upper bound, and inf denotes the greatest lower bound. Sometimes the 95th Percentile of the Hausdorff Distance (HD95), a more robust version of HD, is applied instead of HD, which is less sensitive to outliers:

$$\mathrm{HD95} = P_{95}\left(\{d(a, S_B)|a \in S_A\} \cup \{d(b, S_A)|b \in S_B\}\right) \tag{10}$$

where P_{95} denotes the 95th percentile of the set of distances, $d(a, S_B)$ is the minimum distance from point a in S_A to any point in S_B, and $d(b, S_A)$ is the minimum distance from point b in S_B to any point in S_A.

- *Reconstruction*: The performance of medical image reconstruction is evaluated by measuring the quality of the reconstructed image compared with the original image. The following two metrics are used in federated medical image reconstruction works:

  – *Peak Signal-to-Noise Ratio (PSNR)*:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{11}$$

  where $\mathrm{MSE} = \frac{1}{WH}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\left[I(i,j) - K(i,j)\right]^2$ is the mean square error between the original image and the reconstructed image, MAX is the maximum possible pixel value of the image (255 for 8-bit images), where $I(i, j)$ and $K(i, j)$ are the pixel values at position (i, j) in the original image and the reconstructed image respectively.

  – *Structural Similarity Index (SSIM)*:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{12}$$

  where x and y are the original and reconstructed image pixel values; μ_x and μ_y are the mean intensities of x and y respectively; σ_x^2 and σ_y^2 are the variances of x and y respectively, σ_{xy} is the covariance of x and y; $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are two constants to avoid division by zero, where L is the range of the pixel values (255 for 8-bit images), K_1 and K_2 are constants typically set as 0.01 and 0.03 respectively.

3 Challenges and Solutions

3.1 Data Heterogeneity

Data heterogeneity is a fundamental yet practical challenge in the application of FL in medical image analysis. There are various types, and each impacts the performance and effectiveness of the task model trained in the FL paradigm. In this subsection, we first define the common types of data heterogeneity in MEDIA and then summarize the existing solutions aiming to reduce its negative effect.

3.1.1 Heterogeneity Types

- Feature distribution skew: Let K be the number of hospitals and $P_k(x)$ be the feature distribution of hospital k. Different hospitals may use various types of imaging equipment or protocols, leading to variations in the features present in the medical images. Therefore, $\{P_k(x)\}_{k=1}^{K}$ are different across hospitals.

- Label distribution skew: Let $P_k(y)$ be the label distribution of hospital k. Taking skin cancer detection as an example, some hospitals primarily serve populations with certain skin types and, therefore, have a high incidence of specific skin conditions. It results in the case where certain diagnoses are overpresented in some hospitals and underrepresented in others, which can be formulated as $P_k(y) \neq P_{k'}(y)$ $(k, k' \in \{1, 2...K\}$ and $k \neq k')$.

- Label concept shift: Let $P_k(x|y)$ be the conditional distribution of hospital k. Label concept shift occurs when the features associated with the same label vary across hospitals and it can be mathematically formulated as $P_k(x|y) \neq P_{k'}(x|y)$ $(k, k' \in \{1, 2...K\}$ and $k \neq k')$. For example, in one hospital, certain features in brain MRI scans such as slight swelling or minor anomalies are consistently correlated with a neurological disorder. In contrast, in another hospital, due to differences in patient demographics or disease prevalence, these features may not strongly predict the same disorder.

3.1.2 Generalized FL

Generalized FL (gFL) and personalized FL (pFL) are two promising directions to address the data heterogeneity. The primary goal of gFL is to develop a single, robust model that performs well across all participating clients and can be generalized to other unseen clients. The objective of gFL can be expressed as:

$$\min_{\theta}\{\mathcal{F}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{F}_k(\theta)\} \tag{13}$$

where θ represents the global model parameters and $\mathcal{F}_k(\theta)$ is the loss function defined by client k. This formulation seeks to minimize the average loss across all clients.

Based on our proposed taxonomy, most gFL approaches are grouped into four types: data-based, loss-based, architecture & training-based and aggregation method-based.

(a) Data-based methods

Data-based methods serve as a strategic approach to enhance the diversity and quality of training data across different clients, operating at both input data and feature levels. To increase diversity at input data-level, several techniques have been explored in MEDIA, including image augmentation [17], synthetic minority over-sampling technique (SMOTE) [139], and generative adversarial networks [106, 16, 53, 117, 149]. In addition, style-transfer models are employed in FL to enhance the style diversity of local data, thereby improving the generalization ability of the global model [15]. FedDG [85] tries to alleviate data heterogeneity in frequency space. To be specific, based on the insight that the amplitude spectrum of an image denotes low-level distributions while the phase spectrum denotes high-level semantics, clients share their amplitude spectrum with other clients and replace some low-frequency components of local images with those from other clients so that enrich client distribution. Instead of generating data samples within known classes, FedOSS [168] focuses on an open set recognition problem that aims to correctly identify unseen new samples as unknown classes. It proposes a sample synthesis strategy that can push samples that are close to the decision boundary outside the boundary to get virtual samples of an unknown class. On the other hand, feature augmentation [5] focuses on enriching the feature set itself to enhance its representation ability. To better align with the global data distribution, Huang *et al.* [52] uses the batch-wise mean and standard deviation of features in each institute to abstractly represent the discrepancy of data, and models each feature statistic probabilistically via a Gaussian distribution. The clients then implement feature augmentation to match the global distribution of cross-instituted averaged mean and standard deviation.

(b) Loss-based methods

Loss-based methods focus on modifying the loss functions to accommodate for the inherent data heterogeneity among different clients. These methods generally involve introducing additional regularization terms into the traditional task loss to guide the training process towards more generalizable solutions instead of local data distribution. Considering that different layers capture

varying levels of semantic information, existing techniques generally apply regularization at feature or logit levels. These methods can be further subdivided into three types: feature representation alignment, feature distribution alignment, and logit alignment.

- *Feature representation alignment:* These works apply regularization terms in loss function that focus on aligning the feature representation, the output of the model's feature extractor. For example, FedDAvT [72] considers the scenario that multiple clients with labeled data (source domain) assist a server with unlabeled data (target domain), and it applies the L2-norm between the feature representations of source model. FEDMBP [39] applies the L2-norm between class-specific feature prototypes of the local model and the global model. In addition to the L2-norm loss, both KL loss [68] and contrastive loss [88, 153] are employed to align feature representations. For example, Yang *et al.* [153] utilizes contrastive loss to align the feature representation of data generated by the current local model with that of the global model, while simultaneously distancing it from the feature representation produced by the previous local model. FedDG [85] focuses on a different aspect instead: the contrastive loss is applied in order to align feature representations within the same class while distinguishing different classes, aiming to enhance boundary prediction in medical image segmentation tasks.

- *Feature distribution alignment:* Some works try to align client local feature distribution with global. For example, Gao *et al.* [40] proposes a regularization term that aligns feature distribution by the means and variances from the batch normalization layer of the deep learning model, as these statistics in BN layers represent the characteristic of data distribution [77]. FedDAvT [72] applies a regularization term that utilizes maximum mean discrepancy, which can measure the difference between two distributions.

- *Logit alignment:* Instead of the above works that focus on the intermediate output of network models, some works pay attention to the prediction of the model. For example, FedAD [42] considers the scenario where the server holds a publicly available dataset to guide local training. Clients make predictions with local models using their own local dataset and the public dataset, respectively, and an L2-norm between the prediction logits is used as a knowledge distillation loss. Similarly, RFLPV [131] regulates the logits of the local model to prevent excessive deviation from the global model, using KL divergence for this purpose.

(c) Architecture and training-based approaches

Architecture and training-based approaches play an important role in achieving generalized FL by either architectural design or the development of training strategies. For example, given that ensemble learning captures a broader range of patterns in the data than a single model, FedEL [142] extends this idea into FL by combining a shared feature extractor and a group of classifiers to perform disease diagnosis. Meanwhile, pre-trained foundation models [74] and quantum tensor network model [9] have been investigated, and they show satisfactory performance in addressing data heterogeneity. In terms of training strategy-based approaches, one common method is adversarial training with a focus on training a domain-invariant feature extractor. These approaches align the feature space across different hospitals [59]. In addition, the sharpness-aware minimization is utilized to simultaneously minimize loss value and loss sharpness, leading to a more generalized model [55]. Instead of making an effort on client training, DC-SFL [154] aims to alleviate global model drift on the server side by performing a one-step gradient descent with a weight correction loss formed by the L2-norm between the current and previous global model.

(d) Aggregation method-based approaches

Aggregation method-based approaches aim to enhance the overall model's performance by adaptively aggregating local models, taking into account the data distribution rather than using uniform weights. For example, Yue *et al.* [155] employs reinforcement learning (RL) to seek the optimal weights with the reward defined as the accuracy of the global model on the dataset at the server side. Inspired by the fact that client-specific models should contribute more to the global server, FedMAS [31] assigns larger weights to the local model which exhibits a large class-aware divergence between itself and the global model. In addition, other criteria, such as the gradient similarity [11], have been explored to generate a more generalized and robust global model.

3.1.3 Personalized FL

Considering that it is challenging to fit a single model to diverse data distributions of different clients, personalized FL is proposed as a promising direction to address the data heterogeneity issue. In contrast to gFL, the primary objective of pFL is to develop individual models that are tailored to the needs of each client with contributions from other clients. The objective of pFL can be expressed as:

$$\min_{\{\theta_1,\ldots\theta_K\}} \{\mathcal{F} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{F}_k(\theta_k)\} \tag{14}$$

where θ_k represents the local model parameters of client k.

Based on our proposed taxonomy, most pFL approaches are broadly categorized into three types: architecture-based, training-based and aggregation method-based. Even though the methods used to implement pFL may appear similar to those employ in gFL from the perspective of group names, they are guided by different principles due to their distinct ultimate objectives. Hence, these methods are different from those designed for gFL.

(a) Architecture-based approaches Architecture-based approaches aim to achieve personalization by decoupling the network into shared and private components. This allows the model to capture both global patterns and local nuances. For example, FLOP [150] decouples the classification network used for disease diagnosis into two parts: a feature extractor and a classifier. Clients share the feature extractor for federated averaging while keeping the classifier private. A similar decoupling idea is used by Wang *et al.* [130], but with a key difference: FLOP aggregates the shared part using averaging, whereas the latter uses uncertainty information for aggregation. UniFed [56] extends the idea of FedBN [75] into the medical image analysis, which addresses data heterogeneity by using batch normalization statistics calculated from each local client rather than the averaged version from all clients.

This concept is not limited to classification tasks. It is also utilized in medical image denoising [162], synthesis[24], reconstruction [91] and segmentation [128]. The key point is how to design shared-private components. To be specific, FedFTN [162] is proposed for multi-institutional low-count PET denoising. In this framework, all clients share a common denoising network, while each client designs and trains a feature transformation network using their local data. This approach modulates the feature outputs of the denoising network, enabling personalized denoising tailored to each institute's specific needs. Similarly, pFLSynth [24] designs a personalized block to modulate the statistics of generated feature maps to be institute-specific, inserting it after each convolutional block. Considering that each local client may focus on features in difference channels for magnetic image reconstruction, ACM-FedMRI [91] includes not only a shared image reconstruction network but also a client-specific hypernetwork. This hypernetwork is designed to guide channel selection, optimizing the features extracted for the reconstruction task. FedDP [128] applies a transformer for the medical image segmentation task, where they make queries personalized while keys are shared. The idea is based on the insight that query embeddings contain pixels' own feature information in local images, while key embeddings are related to support information from other pixels, so that such personalization design can help clients learn long-range relationships across data from all clients.

(b) Training-based approaches Training-based approaches focus on designing the learning process to develop personalized models that incorporate

knowledge from various client while preserving local expertise. For example, Per-FedAvg [33] is introduced in the PPPML-HMI [164] FL framework to handle data heterogeneity. Per-FedAvg is built on the top of model agnostic meta-learning (MAML) formulation to learn a good initial global model, which is updated with a few steps of gradient descent for stronger personalization. Several works focus on achieving personalization by preventing local model updating from forgetting previous local-specific knowledge after receiving a global model from the server. For example, IOP-FL [58] is proposed to fuse gradients from local model and global model when training locally on clients instead of solely using one of them. Chen *et al.* [20] designs a multi-step knowledge transfer strategy for local training in order to transfer global knowledge smoothly to clients, with the help of a deputy model receiving global model parameters served as a teacher to the local model.

(c) Aggregation method-based approaches Aggregation method-based approaches for achieving pFL focus on refining how data from multiple clients is integrated to create a global model that also maintains a level of personalization for each participant. These methods typically modify the standard aggregation process used in FedAvg to better address the unique data characteristics and needs of each client. For instance, FedAGA [41] and GRACE [157] capture inter-client relationships by evaluating similarities based on gradients of client models. It then generates personalized models for each client by aggregating all local models in a weighted manner, with weights proportional to the similarity values. Similarly, similarity information is used in FedLPPA [79] and pFedNet [159] for local model aggregation. This similarity-based weighting helps to mitigate the negative influence of models from other clients that are trained on data distributions significantly different from that of the target client. By doing so, it effectively supports the generation of personalized global models that are better tailored to each local client's specific data characteristics. From a frequency domain perspective, [20] uses a low-pass filter to filter out high-frequency components of model parameters when aggregating client models, and these high-frequency components are maintained locally when clients receive global models from the server. This strategy is based on the insight that low-frequency components of parameters are the basis for the network capability, while high-frequency components may contain client-specific knowledge. In addition, instead of manually designing aggregation weights as above, some works aim to assign a learnable weight set for each client to further enhance flexibility. For example, APPLE [90] maintains a learnable weight vector on each client. During local training, each client applies an aggregated model whose parameters are weighted sum from all client models so that the weights can be optimized by gradient descent. Different from previous approaches, HPFL [81] digs into a more detailed level, which applies a hyper-network to learn aggregation weights per layer from other client models.

3.1.4 Summary

We summarize the gFL and pFL approaches in Table 4. The setup and configurations for these approaches are shown in Figure 3. It distinguishes among five main types of approaches: Data-based, Loss-based, Architecture-based, Training-based, and Aggregation method-based, assessing their presence in both gFL and pFL, alongside their respective advantages and disadvantages.

Table 4: Comparisons of representative approaches in generalized FL and personalized FL.

Approach Type	**Present in gFL**	**Present in pFL**	**Advantages**	**Disadvantages**
Data-based [17, 139, 106, 16, 53, 117, 149, 15, 85, 168, 5, 52]	✓	-	Easy to implement; Diversity enhancement; Bias reduction	Increased demand for storage resources; Possibility of privacy leakage
Loss-based [72, 39, 68, 88, 153, 85, 40, 77, 42, 131]	✓	-	Easy integration with existing training procedures	Risk of over-regularization
Architecture-based [142, 74, 9, 59] (for gFL); [150, 130, 56, 162, 24, 91, 128] (for pFL)	✓	-	Fast adaption to the downstream tasks	Potential high development costs
	-	✓	Privacy enhancement	Need to determine the optimal privatization strategy
Training-based [55, 154] (for gFL); [164, 58, 20] (for pFL)	✓	✓	Customizable to clients' specific needs	Increased training cost
Aggregation method-based [155, 31, 11] (for gFL); [41, 157, 79, 159, 20, 90, 81] (for pFL)	✓	✓	Reduced burden at the client side	Need to design criteria for calculating weights

Data-based approaches are mainly employed in gFL. They aim to improve the generalization ability of local models by enhancing the diversity of the input data or features and are straightforward to implement. However, due to the increased volume of data, these methods demand more storage resources. In addition, clients may need to share information about local data, leading to privacy leakage.

Similar to data-based approaches, loss-based approaches are mainly used in gFL. These methods focus on regularizing the output of local and global models at the feature or logit levels, aiming to prevent the local model from deviating significantly from the global model, therefore improving the convergence rate of the global model. They are noted for easy integration with existing training procedures. However, there is a risk of over-regularization, which can lead to a neglect of local data in the learning process.

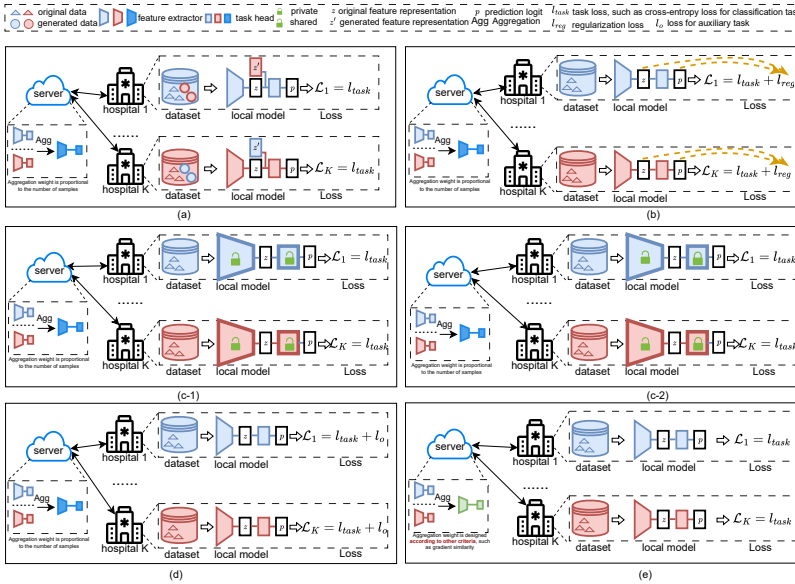

Figure 3: Setup and configurations for five types of generalized (gFL) and personalized FL (pFL) approaches. (a)Data-based approaches in gFL: Utilizes augmented input data and features, represented by circles and z', respectively. (b)Loss-based approaches in gFL: Implements regularization (denoted as l_{reg}) at the feature z or logit p levels. (c-1)Architecture-based approaches in gFL: Utilizes foundation models with increased parameter counts, represented as larger trapezoids and rectangles. (c-2)Architecture-based approaches in pFL: Features a design with shared-private layers, differing from (c-1) by having private task head parameters, symbolized by locked locks. (d)Training-based approaches in both gFL and pFL: Designs auxiliary tasks tailored to the specific needs of gFL and pFL, denoted as l_o. (e)Aggregation method-based approaches in both gFL and pFL. Employs diverse aggregation criteria designed at the server side to optimize learning outcomes.

Architecture-based approaches are utilized in both gFL and pFL, although they adopt different strategies for the two objectives. To be specific, in gFL, these approaches mainly rely on foundation models, which possess numerous parameters and are pre-trained using large volumes of data. While these models provide generalizability and can quickly adapt to downstream tasks, they necessitate greater computational resources and memory capacity from the clients for training. In contrast, pFL employs a shared-private layers design to achieve personalization. This design enhances privacy protection, as only a portion of the parameters are shared. However, it also presents challenges in determining the optimal privatization strategy.

Training-based approaches achieve generalization and personalization through the design of auxiliary tasks tailored to the specific requirements of gFL and pFL. In gFL, for example, an auxiliary task such as adversarial

training is employed to develop a domain-invariant feature extractor, enhancing the model's ability to generalize across diverse data domains. In contrast, in pFL, the focus shifts to meta-learning as the auxiliary task, which facilitates fast personalization to individual client needs. These training-based approaches are highly valued for their ability to be customized to specific client requirements. However, they inherently involve additional training costs, reflecting the complexity and resource-intensive nature of their implementation.

Lastly, aggregation method-based approaches are popular in both gFL and pFL, with different criteria tailored for each. Generally, in gFL, the criteria aim to enhance the overall model's performance. Conversely, in pFL, the focus is on generating personalized models for each client by evaluating the similarities between the client and others, thereby ensuring that each model is finely tuned to individual needs while absorbing knowledge from others. Since aggregation method-based approaches are executed server-side, it can reduce the computational burden on the client side. However, the success of these approaches greatly relies on the effectiveness of the designed criteria for calculating weights, which ultimately determines task performance.

Overall, each approach offers unique benefits suitable for specific contexts within federated learning frameworks but also comes with inherent challenges that must be carefully managed.

3.2 Low Label Quality

Low label quality is another dominant challenge when applying FL to medical imaging tasks. In the medical FL context, low label quality mostly refers to imperfect labeling in medical datasets. Such problems are not specific to FL scenarios, as they can occur in many kinds of deep learning tasks, but they bring specific challenges when applied to FL. In this subsection, we will summarize common types of low label quality challenges and demonstrate existing solutions to solve them according to our literature study.

3.2.1 Class Imbalance

- Problem description: Class imbalance, also referred to as long-tailed data distribution, presents a significant challenge in medical image classification tasks due to the scarcity of certain conditions or abnormalities. This imbalance can significantly compromise a model's ability to accurately detect these rare conditions, as deep learning algorithms tend to favor the majority class, overlooking the minority classes.

- Solutions: To address this issue, existing solutions have been developed from different aspects. For example, DSIFL [139] employs SMOTE to generate synthetic data of the minority class through linear interpolation

among instances within the minority class to enhance representation. At the feature level, FedAWA [155] fuses the features of majority classes and uses the fused features as well as the original features of the minority classes as the classifier input for a class balance. Furthermore, several solutions are proposed at the loss level. For example, FedAR [92] introduces a class-balanced cross-entropy loss where the examples are re-weighted according to the inverse of the effective number of samples per class. FCA [135] modifies the standard cross entropy loss by adding per-class margins, determined by class frequency, into the loss calculation to prioritize minority classes. Extending this concept, FedIIC [141] also adapts the cross entropy loss by considering not only class frequency but also the difficulties associated with each class. From the perspective of model aggregation, Abbas *et al.* [2] address class imbalance by incorporating the class imbalance ratio into the aggregation of local models.

3.2.2 Label Deficiency

Label deficiency means that not all samples in the dataset are equipped with complete and authentic labels. Specifically, the labels for some data samples might be missing or noisy. Based on our literature study, there are mainly four types of label deficiency scenarios in FL-based medical imaging research: (a) label deficiency within clients; (b) label deficiency across clients; (c) label deficiency in multi-label tasks; and (d) label deficiency in intensity.

(a) Label deficiency within client

- Problem description: The dataset on each client has labeled and unlabeled samples, and labeled samples are usually much fewer than unlabeled ones.
- Solutions: The common practice in existing works is to conduct self-supervised learning on all data samples, and then fine-tune on labeled samples. Contrastive learning is a state-of-the-art self-supervised learning method that is widely used in FL-based medical imaging tasks. Generally, contrastive learning aims to enforce the image encoder to learn to identify positive samples of an image while distinguishing its negative samples. In practice, positive samples are usually generated by augmenting the original image, while negative samples are other images, and the insight of contrastive learning is that a robust deep learning encoder should have the capability to generate augmentation-independent representations to one image. Given that MoCo, the commonly-used contrastive learning method employs a memory bank that allows for a greater number of

negative samples without increasing the batch size, several approaches [30, 144, 143] integrate it into FL to make full use of unlabeled data. MoCo first performs self-supervised learning with InfoNCE loss [102], then fine tunes on labeled samples. Additionally, when doing model aggregation on the server, FedMoco [30] assigns more weight to the client model that has larger representational similarity [69] with respect to previous communication round based on the insight that such model has learned more meaningful representations.

(b) Label deficiency across clients

- Problem Description: There are two types of clients: one has fully labeled datasets, while another has no labels. The labels of the labeled clients are usually assumed to be fully trustworthy.

- Solutions: The core challenge in such a scenario is to transfer the knowledge obtained from labeled clients to unlabeled clients. Most existing works apply semi-supervised learning techniques, using the global model to generate pseudo labels for unlabeled clients [99, 71, 112, 127, 105, 138, 152]. For example, Fat [99] feeds an augmented sample generated by mixup [156] from two unlabeled data samples to the local model, which is initialized by global model aggregated from labeled clients at each communication round, to get a prediction, and uses the pseudo label which is the mixup of the outputs of a momentum updated auxiliary model from these two samples to supervise the local model. To denoise pseudo labels generated by the global model which is aggregated from labeled clients, Qiu *et al.* [105] employ Monte Carlo Dropout [38] and use these refined pseudo labels to supervise the training on unlabeled clients. S2FA [152] considers a special scenario where all clients are labeled while the server has an unlabeled dataset. To enhance the quality of pseudo labels, the server picks out the prediction from local models with the most votes. The constructed data-pseudo label pair is used to train the server-side model, contributing to the global model aggregation.

  In addition, it is unfair to treat all clients solely according to the number of samples on the client as FedAvg when aggregating model parameters. In view of this, Saha *et al.* [112] assign larger weights to the clients whose model parameters are closer to that of the averaged global model parameter, and Wang *et al.* [127] evaluate the performance on the validation set, which is used for weights calculation. Considering that labeled data can provide meaningful task-specific information even if it has a smaller number of samples compared to unlabeled data, many works propose customized weighting strategies, where unlabeled clients are often assigned with lower weights [105, 138, 152].

Considering the challenges of learning class-specific discriminative knowledge from pseudo labels, some works [86, 138] leverage class-specific features in the form of the average feature maps from all samples within a class as assistive knowledge to aid in the model training on the unlabeled clients. To be specific, FedIRM [86] leverages client-invariant disease class relationship knowledge denoted by class-specific averaged feature representations at each client, and unlabeled clients will align its local disease knowledge with global knowledge from labeled clients. FedSemiSeg [138] utilizes class (foreground or background) prototype, which is the average feature map obtained from the local model with client images, to construct a contrastive loss as a regularization term.

(c) Label deficiency in multi-label tasks

- Problem description: Multi-label classification (MLC) is a special scenario in deep learning classification tasks. Instead of the common situation that each data sample belongs to one class, which uses an one-hot vector to denote classification labels, samples in MLC can belong to multiple classes so that the label for such samples is a vector containing multiple 1's. Label deficiency in MLC usually considers such a scenario where each data sample is labeled with partial classes, and the labels for other classes are missing, while the labeled classes vary for different samples. In FL-based medical imaging research, existing works generally consider the following setting: the dataset on each client has the same label space, but different clients have different label spaces. Federated learning is prone to getting stuck with local overfitting in such a setting.

- Solutions: To tackle this issue, Gao *et al.* [40] apply a weak label form [137] to unify multi-dimensional labels. To be specific, for all unlabeled classes, assign an equal value that is added to 1 as the label. The authors propose an unbiased loss function that makes the gradient of more probable unlabeled classes larger, avoiding forgetting the knowledge for unlabeled classes from the global model. They further introduce a regularization term calculated by the means and variances from the BN layer, in order to align feature statistics. Dong *et al.* [29] tackle this issue from the perspective of optimization and model aggregation. To be specific, the authors form a bi-level optimization problem where the classifier is optimized in inner optimization and the feature extractor in outer optimization, which can mitigate overfitting to local partially labeled data. For model aggregation, feature extractors are weighted by the number of samples, while classifiers are weighted by the number of samples per labeled class.

(d) Label deficiency in intensity

- Problem description: Such challenge refers to weakly-supervised learning, which usually occurs in medical image segmentation tasks. Labels are categorized into different levels of intensity: pixel (foreground areas labeled per pixel), bounding box (foreground areas labeled by bounding boxes), image (a general label for the whole image), and unlabeled. These label levels, except pixel level, are named weak labels, which can introduce noises when supervising segmentation learning tasks.

- Solutions: Solutions to address this challenge can be broadly divided into two groups: aggregation-based and pseudo label-based. Aggregation-based methods assign adaptive weights to clients by evaluating the quality of each client based on one specific criteria, aiming to diminish the adverse impact of clients with low-quality data. For example, FedAR [92] assesses the quality of each client based on their model performance on the validation set. Similarly, QA-SplitFed [62] evaluates client quality based on the upper bound of a 95% confidence interval for the mean loss value, with the aggregation weight being inversely proportional to this value. FedA^{3}I [140] designs a strategy to estimate the noise level in each client and adjust aggregation weights accordingly. To be specific, this method defines an inner and outer region, starting from the contour of the noisy segmentation mask to a distance, and assesses learning difficulty by dividing the cross-entropy loss between the two regions. The estimated noise is directly proportional to the difficulties calculated for the two regions. Moreover, based on the insight that weakly labeled clients have lower loss since learning from pseudo labels is easier, FedMix [136] proposed a model aggregation strategy that clients with lower loss initially get lower weights, which gradually increase as training goes. On the other hand, instead of assigning adaptive weights according to loss, FedDM [167] tries to make modifications on shared gradients in order to alleviate gradient conflict caused by noisy labels that vary on clients when aggregating client models on the server. They perform orthogonal decomposition on each client gradient and the parallel and opposite components are ignored during aggregation.

  Pseudo-label-based methods stand out for their universality and flexibility, utilizing model predictions to generate pseudo-labels. To optimize pseudo-label generation, FedLPPA [79] designs an encoder-dual decoder architecture for medical image segmentation. One decoder is global, aggregating information from clients in an average manner, and the other is personalized, which incorporates useful information from similar clients. The pseudo label is the combination of outputs from both decoders. Bai *et al.* [8] design a strategy to identify noisy labels by comparing the pre-

diction distribution of the global model and the original label, therefore correcting noisy labels for segmentation model training. FedMix [136] treats weak labels as refining factors for pseudo labels generated by a global model and applies cross-pseudo supervision [18], which trains two models that use their refined pseudo labels to co-supervise each other on weakly labeled clients. Meanwhile, only reliable samples are used for training, with reliability evaluated based on the prediction consistency of the pseudo labels. Instead of regarding weak labels as assistant knowledge, FedDM [167] considers weak labels as noisy ground-truth labels. It leverages the prediction consistency and inconsistency of models sent from other clients to help each weakly labeled client to pick out clean labels and filter out noise labels. For the multi-organ segmentation tasks, the datasets are generally partially labeled. In view of this, the works [66, 55] propose a simple yet effective strategy using a pretrained organ-specific segmentation model to generate pseudo labels for local training. To overcome the impact of label noise on local training, FedGP [19] constructs the purified graph with reliable samples with small loss values and gradually adds more purified samples selected with the output confidence and prediction consensus, and then uses it to generate reliable pseudo labels using topological knowledge.

3.2.3 Summary

In this subsection, we explore the challenges associated with low label quality encountered in the application of FL in MEDIA. Each challenge is associated with specific assumptions, tasks, objectives, and tailored solutions, which are summarized in Table 5.

Class imbalance is a challenge not only in FL paradigm but also in centralized learning. This issue often arises due to the scarcity of certain conditions or abnormalities, leading to some diseases being underrepresented. To improve the model's ability to learn from these underrepresented classes, techniques such as synthetic data generation and resampling are commonly employed. In addition, many studies focus on modifying the cross-entropy loss, the commonly-used loss in classification tasks, to ensure fair representation of all classes.

Considering that data annotation is both time-consuming and labor-intensive, the challenge *Label deficiency within client* assumes that each client possesses a small amount of labeled data alongside a substantial volume of unlabeled data. To maximize the utility of unlabeled data, contrastive learning, a self-supervised learning method, is employed to learn good feature representations. The model is further fine-tuned on labeled data to enhance its performance on specific tasks.

The challenge *Label deficiency across clients* is FL-specific in the MEDIA sector. It considers scenarios where some clients have labeled data while others do not. To facilitate the transfer of knowledge from labeled to unlabeled clients, pseudo labeling, a simple but effective technique for semi-supervised learning, is employed. To enhance the quality of these pseudo labels, the above-mentioned proposed FL frameworks contribute to refining the model that generates these labels, ensuring their accuracy and reliability.

The challenge *Label deficiency in multi-label tasks* involves tackling incomplete annotations for data expected to have multiple labels. To accurately predict the full set of relevant labels for each instance, strategies include designing specialized loss functions and modifying aggregation weights to cater to the complex nature of multi-label data.

Compared to classification tasks, obtaining exact labels for segmentation tasks is more challenging. Annotations can vary across different levels, such as pixel-level, bounding box-level, or even image-level, with all levels except pixel-level considered noisy. To mitigate the adverse effects of this challenge *Label deficiency in intensity*, two promising strategies are proposed. The first strategy aims to diminish the negative impact of noisy clients by assigning them lower weights. The second involves correcting noisy labels through the use of pseudo-labeling, thereby improving label accuracy.

Overall, Table 5 highlights the tailored approaches necessary to address various data quality challenges in FL, emphasizing the need for specialized strategies to ensure robust model training and generalization across diverse and unevenly distributed datasets.

3.3 Attack and Defense

3.3.1 Inference attack

- Attack description: An inference attack in FL is a type of security where an adversary seeks to infer sensitive information about the training data used by participants without directly accessing it. Common types include model inversion attacks and membership inference attacks. To be specific, a model inversion attack aims to reconstruct participant-specific data, while a membership inference attack is to determine whether a particular data record was used in the training dataset.

- Defense: Existing solutions employ cryptographic protocols to defend against inference attacks. The most commonly used cryptographic techniques are Differential Privacy (DP) and Homomorphic Encryption (HE):

  – *Differential Privacy* is a data sharing framework that aims to protect individual private information when sharing statistical information

Table 5: Summary of challenges related to low label quality and corresponding solutions.

Challenge	Assumption	Task	Objectives	Solutions
Class imbalance	Labeled data with uneven class distribution	classification	To enhance model's ability to learn from underrepresented classes	Synthetic data generation [139]; Resampling [155]; Loss function modification [92, 135, 141]; Aggregation weights design [2]
Label deficiency within client	Each client has labeled and unlabeled data	classification; segmentation	To make full use of the unlabeled data	Contrastive learning [30, 144, 143]
Label deficiency across clients	Labeled clients and unlabeled clients	classification; segmentation	To transfer knowledge from labeled clients to unlabeled clients	Pseudo labeling [99, 71, 112, 127, 105, 138, 152]; Class-specific knowledge alignment [86, 138]
Label deficiency in multi-label tasks	Incomplete annotations for data expected to have multiple labels	classification	To accurately predict the full set of relevant labels for each data	Loss function design [40]; Aggregation weights design [29]
Label deficiency in intensity	Noisy labels caused by annotations at different levels	segmentation	To reduce negative effect caused by noisy clients **OR** To correct noisy labels	Aggregation weights design [92, 62, 140, 136, 167]; Pseudo labeling [79, 8, 136, 18, 167, 66, 55, 19]

of datasets. By applying DP, when one individual sample within the dataset is modified, the shared statistical information should not generate such change from which attackers are able to identify private information about this individual sample. In the medical FL domain, the common practice to realize DP is to add Gaussian noise to gradients [101, 3, 63, 118, 110] and apply DP-SGD algorithm [1]. To be specific, Kalra *et al.* [63] and Riedel *et al.* [110] discuss the problem that DP directly applied to neural networks containing batch normalization (BN) layers can violate the privacy requirements of DP-SGD, as it makes gradients depend on batch samples. In view of this, they applied specially designed layers to replace BN layers.

– *Homomorphic Encryption* [126] is a cryptographic protocol that allows direct operations on encrypted data without requiring decryption first. It reduces computational demands and time needed for operations while safeguarding data privacy. It is suitable to

apply HE in FL scenarios, as it can enable information aggregation to be performed correctly in a privacy-preserving manner. For example, DC-SFL [154] applied a widely-used HE framework Paillier cryptosystem, which has additive homomorphism, to encrypt model parameters before aggregating on the server.

In addition to the commonly-used cryptographic protocols, the certificateless ring signature [132] is utilized to obscure the source of parameter updates to resist source inference attack, an extension of membership inference attack. In addition, RFLPV [131] utilizes the masking technique to obscure the original gradient, hereby enhancing privacy protection against inference attacks. Based on Split Learning, p-FeSTA [103] introduces a feature-space permutation strategy that randomly shuffles client intermediate feature patches before sending them to the server in order to avoid the risk of reverting original images from these intermediate features. Instead of making effort to encrypt or obscure sensitive information during federated communication, some works aim to share less sensitive or non-sensitive information rather than local data samples, model weights or gradients. For example, to address the data leakage issue, FLOP [150] implements a strategy where only a part of the model, specifically the feature extractor of the task model, is shared across clients to achieve collaboration. FedAD [42] proposes a knowledge distillation-based FL framework where the server holds an unlabeled public dataset and clients have labeled private datasets. The clients first train locally based on their own datasets to initialize local models, and then make predictions with the unlabeled public samples. The prediction logits are then sent to the server, serving as teachers to help the server train a model that can perform well on its unlabeled dataset.

3.3.2 Poisoning attack

- Attack description: A poisoning attack is a case where an attacker manipulates data or model updates submitted by one or more compromised clients to corrupt the global model, resulting in poor performance. Considering that the state-of-the-art aggregation methods rely on the distance between malicious and benign client model parameters to defend against poisoning attacks, Joshi *et al.* [61] introduce an attack strategy that aims to maximize the objective loss function while ensuring that the Euclidean distance between the malicious and benign parameters is kept marginal.

- Defense: The current defense strategies mainly design robust aggregation techniques to mitigate the influence of potentially malicious updates.

These techniques heavily rely on the Euclidean and cosine distances between parameters from different clients to differentiate between malicious and benign clients [61]. Based on this assessment, the server assigns lower weights to parameters from identified malicious clients, effectively reducing their influence on the model.

3.3.3 Backdoor attack

- Attack Description: A backdoor attack is a sophisticated form of sabotage in FL where one or more compromised clients embed a trigger in the training data in order to activate altered model behaviors under specific conditions while maintaining normal performance on standard tasks. Since the attacked global model appears legitimate under routine evaluation, it makes the attack stealthy and challenging to detect.

- Defense: Similar to the defenses against poisoning attacks, robust aggregation techniques are also utilized to protect against backdoor attacks. For instance, FedDetect [60] necessitates that each client reports the loss of task model. Utilizing this information, the server conducts outlier detection to identify malicious clients. The aggregation weights are then adjusted to be inversely proportional to the number of red flags each client receives.

3.3.4 Byzantine attack

- Attack description: Byzantine attack refers to a situation where some participants or the server act in a malicious or unreliable manner to disrupt the overall training of the task model. The server, in this case, acts as a Byzantine server, deliberately manipulating the aggregation process or the results it sends back to different participants.

- Defense: Considering the potential risks posed by a malicious server that could alter model parameters and falsify the aggregation, Moulahi *et al.* [97] employ blockchain technology to secure model parameters against the byzantine attack at the server side. To be specific, Smart Contract (SC), instead of the server, is responsible for model aggregation and distribution. In addition, blockchain technology is employed by Singh *et al.* [119] and Jatain *et al.* [54] to create a more tractable, immutable, and transparent FL environment. These implementations highlight blockchain's role in enhancing the integrity and reliability of FL systems.

3.3.5 Summary

In this subsection, we focus on the attacks and corresponding defenses in FL-based medical image analysis. Figure 4 provides an overview of different types of attacks, visually depicting the objectives of each attack and identifying the parties involved in initiating these attacks. To defend against these attacks, cryptographic protocols and technologies such as blockchain have been integrated into FL. In addition, some works have introduced strategies from the perspective of FL training pipeline, such as designing robust aggregation methods. These defenses are summarized in Figure 5. Despite advancements in defensive strategies, research on defense mechanisms remains relatively underdeveloped compared to the extensive solutions proposed for addressing other challenges such as data heterogeneity and low label quality. With AI technologies becoming increasingly advanced, their potential for conducting attacks pose serious security threats. Therefore, effective and robust defense mechanisms are essential to ensure FL's security and privacy in real-world applications.

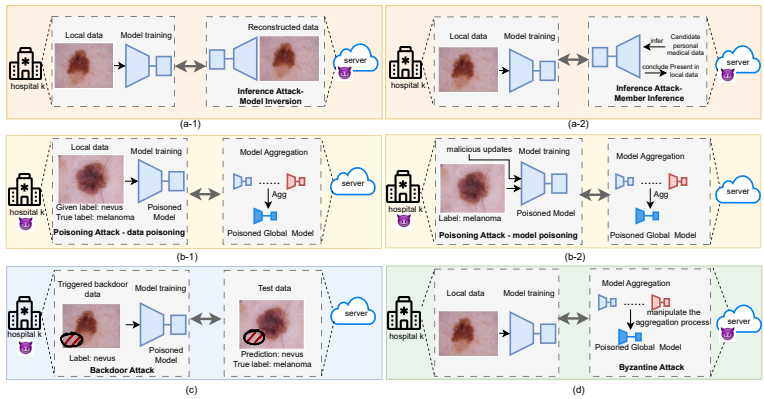

Figure 4: Overview of attacks: (a-1) Inference Attack-Model Inversion Attack: An attacker utilizes screenshots of FL model parameters to reconstruct local data. (a-2) Inference Attack-Member Inference Attack: An attacker uses FL model parameters to identify personal medical data within local datasets. (b-1) Poisoning Attack-Data Poisoning Attack: An attacker introduces tainted and infected data during the training process. (b-2) Poisoning Attack-Model Poisoning Attack: An attacker manipulates local model gradients to compromise the training process. (c): Backdoor Attack: An attacker embeds a trigger in the training data to activate altered model behaviors under specific conditions while maintaining normal performance on standard tasks. For instance, an attacker adds a sticker to all training data labeled nevus. The resulting global model will classify any data featuring this sticker as belonging to the nevus category. (d): Byzantine Attack: The attacker acts in a malicious or unreliable manner, aiming to disrupt the overall training process.

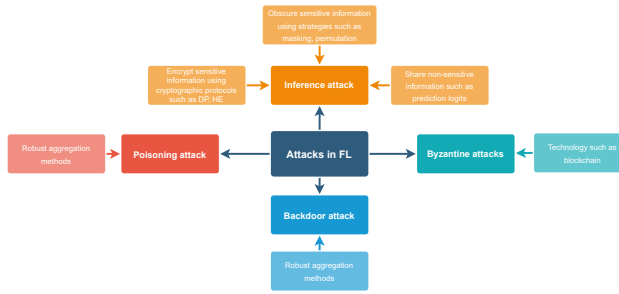

Figure 5: Summary of defenses against attacks in FL.

3.4 Communication Burden

The communication burden in FL is a significant challenge that arises from the need to frequently exchange model updates between numerous participating clients and a central server. This challenge is particularly pronounced in medical image analysis, where the commonly used task models, such as those based on Transformers, tend to have a large number of parameters. This is a consequence of the complexity and high dimensionality inherent in medical image data. Such extensive models significantly exacerbate the communication overhead, making the reduction of this burden even more crucial in the field of MEDIA. Existing works devoted to alleviating the communication burden associated with this process can be roughly divided into the following three categories: parameter reduction, alternative information transmission and efficient client participation.

3.4.1 Parameter reduction

These approaches aim to reduce the volume of gradients or model parameters that need to be transmitted between clients and the server. Techniques such as clipping [17] and gradient compression [159] are utilized to reduce the size of gradients, thereby saving bandwidth and accelerating the communication process. Instead of transmitting all parameters, some studies opt to send a subset of the task model's parameters. For example, FedFMS [87] employs SAM (Segment Anything Model) with adapters for medical image segmentation. It fine-tunes the adapter parameters and then transmits these parameters to the server, significantly reducing the volume of data sent. Shen *et al.* [117] and Dalmaz *et al.* [24] design an adversarial framework comprising a generator and discriminator to perform staining normalization and MRI synthesis tasks, respectively. They apply the same idea that clients collaboratively train only the generator while training the discriminator locally, optimizing the use of network resources. For efficient medical image analysis, Mu *et al.* [98] apply

causal learning that can determine the causal relationships of model parameters, and only those having strong causal relationships are used for aggregation, which is applicable in both classification and segmentation tasks. FedPR [36] reduces the number of shareable parameters by applying a pre-trained model backbone on clients before the federated communication phase, and only the parameters of classification heads that need to be fine-tuned will be transmitted. In terms of split learning, He *et al.* [47] design a model partitioning algorithm based on Lyapunov optimization that can dynamically decide the model split point according to real-time transmission rate and long-term energy consumption.

3.4.2 Alternative information transmission

Instead of sharing model parameters, this direction focuses on transmitting alternative information, which encapsulates critical model insights in a more compact form. For example, federated knowledge distillation [166] transmits logits to achieve knowledge sharing. To be specific, it operates under the assumption that there is a publicly accessible dataset that all clients can use. Each client submits the prediction logits from this dataset to the server, which then aggregates these logits to form a teacher logit. This teacher logit is used to guide the local model training, ensuring efficient and effective learning. Moreover, to tackle the challenges of missing modalities in multi-modal MRI reconstruction, Fed-PMG [147] introduces a pseudo-modality generation mechanism. This method involves sharing the distribution information of the amplitude spectrum in the frequency space among clients. To minimize communication costs typically associated with transmitting the original amplitude spectrum of all images, a clustering approach is used to project the set of amplitude spectra into finite cluster centroids, which are then shared among the clients. A similar idea is also applied in FedSemiSeg [138], which utilizes clustered class-specific prototypes to construct a contrastive loss as a regularization term.

3.4.3 Efficient client participation

This direction aims to optimize which clients participate and how often they communicate. For example, FedACS [43] proposes a client selection strategy, which selects a growing number of clients during FL rounds, where the priority of clients is based on the training loss. It is based on the insight that higher loss means higher data difficulty, further indicating more latent information that can make more contribution to the global model. FedISCA [64] discusses the 'One-Shot Federated Learning', where clients contribute to the global model with a single update, greatly minimizing the communication overhead.

3.4.4 Summary

In this subsection, we examine the critical need for strategies to reduce the communication burden in the application of FL in MEDIA. We provide a thorough review of existing approaches aimed at addressing this challenge, as summarized in Table 6. In addition, we analyze and compare the main idea, advantages, and disadvantages of each approach, offering insights into their practical applications and potential limitations.

Table 6: Comparisons of representative approaches for reducing the communication burden.

Approach	Main Idea	Advantages	Disadvantages
Parameter reduction [17, 159, 87, 117, 24, 98, 36, 47]	Reduce the size of transmitted parameters	Faster convergence	Potential information loss
Alternative information transmission [166, 147, 138]	Transmit derived data	Privacy protection	Potential information loss
Efficient client participation [43, 64]	Reduce the number of participating clients and communication rounds	Scalability	Risk of bias

The approach *Parameter reduction* focuses on reducing the size of parameters that need to be transmitted. The core concept is to streamline communication by either compressing the model or selectively transmitting only the most crucial parameters, thereby facilitating faster convergence. However, it risks potential loss of important information, which might lead to suboptimal model performance.

Instead of sending raw model parameters like the typical FL framework as FedAvg, *alternative information transmission* involves transmitting derived data such as logits or features extracted from the model. This method enhances privacy protection by reducing the granularity of the data shared between the server and clients, preventing potential leakage of sensitive information. However, similar to parameter reduction, it also faces the challenge of potential information loss, as the derived data might not capture all the nuances of the original input.

The last approach *Efficient client participation* aims to scale the system more effectively by reducing the number of communication rounds or selectively participating clients. While this method greatly enhances scalability and reduces resource consumption, it introduces the risk of bias. The selected clients might not represent the overall data distribution, potentially leading to a biased global model.

Overall, each of these strategies offers a viable solution to the challenge of communication burden in FL environments. However, they must be chosen and

implemented carefully, considering the specific requirements and constraints of the deployment context to effectively balance performance with communication efficiency.

3.5 Underexplored challenges

3.5.1 Fairness

Fairness in FL is crucial for ensuring equitable treatment for all clients involved in the collaborative training process. In the context of FL, there are two commonly studied fairness in medical image analysis, including performance fairness and collaboration fairness, each with targeted solutions to uphold specific aspects of fairness.

- Performance fairness ensures that the federated learning model performs uniformly well across all participating nodes and that model updates do not disproportionately benefit some nodes while disadvantaging others. To achieve performance fairness, Prop-FFL [51] proposes a novel optimization objective that includes two terms: one focused on reducing the training loss and the other focused on promoting fairness. The fairness term specifically aims to adjust the model parameters to ensure that all hospitals have a similar training loss, thereby creating a more equitable learning environment across different institutions.

- Collaboration fairness ensures that each client's contribution to the federated model is recognized and value appropriately. Contributions can be in the form of data volume, data quality or data diversity. The most important key to achieving collaboration fairness is to estimate client contribution. Shapley value (SV) [115], a classic approach to quantify the contribution of participants in cooperative game theory, evaluates the contribution of each client as the difference they make when added to every possible subset of clients, which is defined as follows,

$$\phi_i(v) = \sum_{S \subset N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}(v(S \cup i) - v(S)) \tag{15}$$

  where N is the set of all clients and S is a subset of N that does not include player i. $|N|$ is the total number of clients and $|S|$ is the number of clients in subset S. $v(S)$ is the value of coalition S. Considering that SV computation is computationally expensive, SaFE [70] proposes an efficient SV computation technique. To be specific, each client employs a simple logistic regression model as a proxy to approximate the original model. Instead of relying on Monte Carlo techniques, which are typically used for such computations, SaFE utilizes an ensembling approach to estimate SV, enhancing both efficiency and accuracy. Different from

approaches that use Shapley Values (SV), FedCE [57] estimates client contribution in both gradient and data spaces. In the gradient space, FedCE evaluates the differences in gradient direction between one client and all the other clients. Meanwhile, in the data space, it measures the prediction error on a client's data using the global model excluding the client's own parameters. A larger difference in gradient directions and a higher error value indicate a greater contribution from the client.

3.5.2 Model Heterogeneity

- Description: In the medical imaging field, the increasing use of foundation models such as SAM (Segment Anything) [67] reflects a shift towards more complex architectures that are designed for superior task performance. These models, characterized by their extensive parameter sets, require significant computational resources for development. However, some clinics with limited budgets, struggle to deploy, maintain, and train such advanced models. This disparity in capabilities necessitates the use of varied model architectures, leading to model heterogeneity.

- Solutions: Due to model heterogeneity, traditional FL methods, which rely on exchanging model parameters to facilitate collaborative learning, are not effectively applicable. The solutions to this issue can be divided into 2 groups: public data-based and proxy model-based. The public data-based approaches depend on a publicly accessible dataset to enable knowledge sharing among clients. For example, in FHFL [16], clients train their local models using private data and submit the prediction logits of the public data to the central server. This server is responsible for aggregating these logits and then broadcasting them back to the clients. The clients then utilize these aggregated logits as 'teacher logits' to distill knowledge from other clients to themselves, enhancing their own models' learning and performance. Considering that the ensembled logit provides little insight into the underlying structure knowledge of other clients, FedAD [42] combines local information at both the logit and feature-levels to facilitate knowledge transfer. This approach ensures a more comprehensive exchange of expertise, enhancing the overall learning process.

  Proxy model-based approaches require each client maintaining two models: a private model and a publicly shared model. The private models can have heterogeneous architectures, while the shared models have a uniform architecture. For example, ProxyFL [63] employs a proxy model to facilitate efficient information exchange while accommodating model heterogeneity. In addition, it implements deep mutual learning to train the two models on the client side. Although this method is

straightforward to implement, it introduces an extra training burden due to the incorporation of the proxy model.

3.5.3 Multi-Modality

Multi-modality learning is an attractive area in deep learning. The 'modality' here indicates a specific type of data or source of information, including visual images, language text, audio records, etc. Multi-modality learning aims to effectively combine or fuse different sources of information in order to acquire a more comprehensive understanding of given data and achieve better performance on specific tasks that involve different types of source data. However, research in applying multi-modality learning in medical FL is limited. Typical multi-modality learning models assign one encoder for each modality and apply one decoder that fuses the encoded embeddings from different modalities, which will be used in downstream tasks. Most existing works simply apply typical multi-modality learning methods in FL architecture [14, 13, 122, 89, 111]. Few works focus on exploring and solving multi-modality-specific challenges in medical FL scenario.

Existing works mainly consider such a challenge in multi-modality medical FL, that different clients hold data of partial and different modalities. To be specific, assuming that there is a global modality set, each client has a dataset possessing a subset of modalities, and different clients hold different subsets. The core challenge is the heterogeneity and domain shift caused by the missing of modalities on each client when trying to train a federated model based on the global modality set. Several works pay attention to solving the challenge in such a setting with MRI data, a representative multi-modality data source in medical imaging where different modalities focus on different body elements and can provide complementary diagnosis information. For example, Dai *et al.* [23] consider the scenario where the server has a dataset that has all modalities while the data on clients only has a single and different modality. Aiming to learn personalized models suitable for each client together with a global model on the server while dealing with modality heterogeneity, it assigns one encoder for each modality on both clients and server, which will participate in parameter aggregation, and the server and each client maintain a personalized decoder. In order to alleviate heterogeneity caused by missing modalities, clients utilize an attention mechanism to calibrate local decoded features with class-specific prototypes extracted from the fusion decoder on a server that contains information on all modalities. On the other hand, Yan *et al.* [148] view the challenge from a different perspective, which leverages vertical datasets on all clients obtained from the same group of subjects to alleviate domain shift caused by partial modality. To be specific, each client possessing partial modalities has a horizontal dataset and a vertical dataset,

whereas the vertical datasets on different clients have different modalities but are collected from the same group of subjects. It first decouples modality-specific and modality-invariant features on each client, and then introduces a regularization term maximizing the similarity of modality-invariant features of vertical samples among all clients in order to enhance consistency.

Additionally, as a special case, Borazjani *et al.* [11] focus on the heterogeneity of convergence speed for different modalities. Concentrating on cancer stage classification tasks with mRNA sequences, histopathological images, and textual clinical information data, it finds that the convergence speed for training encoders of different modalities is different, which may hinder the convergence of global training. It proposes a learning rate coefficient item based on local training loss to adjust the learning rate for different modalities, leading to a faster and more fluent training process in federated scheme.

4 Future Directions

4.1 Incorporation of gFL and pFL

As illustrated in the discussion of data heterogeneity challenges, most existing works focus on either generalized or personalized federated learning settings. However, in medical areas, both these two scenarios should be attached with importance when applying FL in practice. For example, medical institutes participating in FL training may have different computing capabilities and various aspects of interest in the medical data, resulting in personalized models that fulfill individual client requirements; it is also important to obtain a model with enough generalization ability, enabling a model trained by a limited group of institutes to facilitate medical diagnosis in a broader range of institutes. Only limited existing works take such a scenario into consideration [145, 58, 157]. A valuable future research direction could be better formalization and exploration of combining gFL and pFL in medical image analysis.

4.2 Extreme lack of labels scenarios

According to our summary of label deficiency challenges, existing works exploring semi-supervised scenarios are based on either of the following two settings: label deficiency within the client, where each client has a set of labeled data samples and more unlabeled samples, and label deficiency across clients, where some clients have fully labeled datasets while others have no labels. A more extreme but practical scenario is a combination of these two settings, namely, only a small subset of clients have a limited number of labeled samples, which is closer to real-world applications due to the lack of expert labeling resources in medical institutes. Further exploration in such extreme setting could help FL better applied in realistic scenarios.

4.3 Multi-modality FL

As mentioned earlier, multi-modality FL is an attractive research and application topic, but existing relative works were limited to simply applying multi-modality learning in the FL scheme, and few of them dived deeper into exploring specific challenges in this area. Furthermore, these few works mainly focus on MRI reconstruction tasks where different MRI modalities are all vision data. Despite the developing trend of fusing completely diverse modalities such as vision and language data [133], only one work [11] explored specific challenges in combining such diverse modalities in medical imaging FL scenario. Therefore, a promising direction of future research could be finding more domain-specific challenges and exploring effective solutions in multi-modality medical imaging FL tasks.

4.4 Explainability

Due to the 'black-box' nature of deep learning models, model interpretability and explainability have been the major concerns in deep learning applications. Such a challenge is of rather great significance in medical areas, as the reliability of medical diagnosis can directly affect the health and even life of patients, so physicians may have a high demand for deep learning-aided diagnosis tools to produce explainable and trustworthy results. Federated learning scenarios bring additional challenges in explainability, as institutional heterogeneity may bring difficulty for the federated model to achieve high interpretability on specific clients, and the application of privacy-preserving techniques such as differential privacy can reduce the transparency of the model. According to our survey, there are currently few works about explainability in medical imaging FL. Mu *et al.* [98] tried to introduce a causal reasoning learning technique to make the model explainable. Siniosoglou *et al.* [120] and Ambesange *et al.* [6] applied visualization techniques in order to interpret the learned knowledge of the deep learning model. In order to avoid the 'black-box' nature of neural network based deep learning models, Li *et al.* [78] applied Gradient Boosting Decision Tree (GPDT) model in FL framework, utilizing the explainability of decision tree models to achieve interpretability. A future research direction can be exploring FL-specific challenges in medical image analysis areas, making FL techniques more practical in real-world medical applications.

4.5 Incorporation with Large Models

Recently, large models have seen rapid development in research and applications. These novel and powerful models can provide FL with promising opportunities, e.g., they have the potential to offer promising solutions to the challenges in FL research in medical image analysis. Following are three possible directions in incorporating large models with FL in medical image analysis:

First, large vision foundation models, such as Segment Anything Model (SAM) [67], could be attractive to address data heterogeneity challenge of FL in medical image analysis. These models are pre-trained on massive and diverse datasets and thus have strong generalization capability, which can capture visual features with wider range and higher effectiveness. When applied to FL scenarios, these models can make feature extraction more consistent and robust across diverse datasets at different clients. Therefore, data heterogeneity across clients is less concerned, and the model fine-tuned by the federated learning process could generate more accurate and reliable outcomes.

Another exciting direction is to apply advanced Artificial Intelligence-Generated Content (AIGC) models to address the data heterogeneity challenge of FL in medical image analysis [114]. A major direction is to utilize image generative models to produce high-fidelity synthetic medical images, as an alternative tool in data-based methods in Section 3, which can significantly alleviate data deficiency and heterogeneity concerns in medical applications.

Additionally, as large language models (LLMs) have been successfully applied in various areas, especially those with strong multi-modal capability, including GPT-4o, Gemini, etc., specific medical tasks such as multi-modality-based diagnosis and medical report generation can benefit from these models. Specially, large vision-language models provide a magnetic solution to perform medical image-to-report generation tasks in FL scenarios. When deployed at different institutions and collaboratively fine-tuned in a federated learning framework, such models could significantly improve the quality and consistency of reports generated from medical images, while not violating privacy preserving rules.

5 Conclusion

In this paper, we conduct a survey on the challenges and existing solutions when applying federated learning techniques in medical image analysis. We first provide a summary of the background of FL applications in medical image analysis areas, including the formalization of common tasks and corresponding evaluation metrics. We then present and explain a novel taxonomy of challenges, which contains data heterogeneity, low label quality, attack and defense, communication burden, and several underexplored challenges, where each category is further divided according to the nature of the challenges. For each challenge, we summarize existing solutions in an organized, categorical manner, which systematically and intuitively demonstrates current research insight on these challenges. Furthermore, we discuss some possible future research directions in related fields. We expect that this survey could provide researchers with intuitive and comprehensive understanding in related areas, together with inspirations for future research.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy", in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, 308–18.

[2] Q. Abbas, K. M. Malik, A. K. J. Saudagar, and M. B. Khan, "Context-aggregator: An approach of loss-and class imbalance-aware aggregation in federated learning", *Computers in Biology and Medicine*, 163, 2023, 107167.

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

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

[5] F. Almalik, N. Alkhunaizi, I. Almakky, and K. Nandakumar, "FeSViBS: Federated Split Learning of Vision Transformer with Block Sampling", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 350–60.

[6] S. Ambesange, B. Annappa, and S. G. Koolagudi, "Simulating federated transfer learning for lung segmentation using modified UNet model", *Procedia Computer Science*, 218, 2023, 1485–96.

[7] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated learning for healthcare: Systematic review and architecture proposal", *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4), 2022, 1–23.

[8] L. Bai, D. Wang, H. Wang, M. Barnett, M. Cabezas, W. Cai, F. Calamante, K. Kyle, D. Liu, L. Ly, *et al.*, "Improving multiple sclerosis lesion segmentation across clinical sites: A federated learning approach with noise-resilient training", *Artificial Intelligence in Medicine*, 152, 2024, 102872.

[9] A. S. Bhatia and D. E. B. Neira, "Federated Hierarchical Tensor Networks: a Collaborative Learning Quantum AI-Driven Framework for Healthcare", *arXiv preprint arXiv:2405.07735*, 2024.

[10] L. Bhatia and S. Samet, "A decentralized data evaluation framework in federated learning", *Blockchain: Research and Applications*, 4(4), 2023, 100152.

[11] K. Borazjani, N. Khosravan, L. Ying, and S. Hosseinalipour, "Multi-Modal Federated Learning for Cancer Staging over Non-IID Datasets with Unbalanced Modalities", *arXiv preprint arXiv: 2401.03609*, 2024.

[12] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, *et al.*, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy", *Scientific data*, 7(1), 2020, 283.

[13] B. Casella, W. Riviera, M. Aldinucci, and G. Menegaz, "MERGE: A model for multi-input biomedical federated learning", *Patterns*, 4(11), 2023.

[14] J. Chen and R. Pan, "Medical report generation based on multimodal federated learning", *Computerized Medical Imaging and Graphics*, 113, 2024, 102342.

[15] J. Chen, M. Jiang, Q. Dou, and Q. Chen, "Federated domain generalization for image recognition via cross-client style transfer", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, 361–70.

[16] K. Chen, X. Zhang, X. Zhou, B. Mi, Y. Xiao, L. Zhou, Z. Wu, L. Wu, and X. Wang, "Privacy preserving federated learning for full heterogeneity", *ISA transactions*, 141, 2023, 73–83.

[17] S. Chen, Z. Jie, G. Wang, K.-C. Li, J. Yang, and X. Liu, "A new federated learning-based wireless communication and client scheduling solution for combating COVID-19", *Computer Communications*, 206, 2023, 101–9.

[18] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, 2613–22.

[19] Z. Chen, W. Li, X. Xing, and Y. Yuan, "Medical federated learning with joint graph purification for noisy label learning", *Medical Image Analysis*, 90, 2023, 102976.

[20] Z. Chen, C. Yang, M. Zhu, Z. Peng, and Y. Yuan, "Personalized retrogress-resilient federated learning toward imbalanced medical data", *IEEE Transactions on Medical Imaging*, 41(12), 2022, 3663–74.

[21] G. Choi, W. C. Cha, S. U. Lee, and S.-Y. Shin, "Survey of Medical Applications of Federated Learning", *Healthcare Informatics Research*, 30(1), 2024, 3–15.

[22] M. G. Crowson, D. Moukheiber, A. R. Arévalo, B. D. Lam, S. Mantena, A. Rana, D. Goss, D. W. Bates, and L. A. Celi, "A systematic review of federated learning applications for biomedical data", *PLOS Digital Health*, 1(5), 2022, e0000033.

[23] Q. Dai, D. Wei, H. Liu, J. Sun, L. Wang, and Y. Zheng, "Federated Modality-Specific Encoders and Multimodal Anchors for Personalized Brain Tumor Segmentation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 2, 2024, 1445–53.

[24] O. Dalmaz, M. U. Mirza, G. Elmas, M. Ozbey, S. U. Dar, E. Ceyani, K. K. Oguz, S. Avestimehr, and T. Çukur, "One model to unite them all: Personalized federated learning of multi-contrast MRI synthesis", *Medical Image Analysis*, 2024, 103121.

[25] E. Darzidehkalani, M. Ghasemi-Rad, and P. Van Ooijen, "Federated learning in medical imaging: part II: methods, challenges, and considerations", *Journal of the American College of Radiology*, 19(8), 2022, 975–82.

[26] Z. Deng, L. Luo, and H. Chen, "Scale Federated Learning for Label Set Mismatch in Medical Image Classification", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 118–27.

[27] G. Dharani Devi, R. V, and J. Jeyalakshmi, "Privacy-Preserving Breast Cancer Classification: A Federated Transfer Learning Approach", *Journal of Imaging Informatics in Medicine*, 2024, 1–17.

[28] W. Ding, M. Abdel-Basset, H. Hawash, and W. Pedrycz, "MIC-Net: A deep network for cross-site segmentation of COVID-19 infection in the fog-assisted IoMT", *Information Sciences*, 623, 2023, 20–39.

[29] N. Dong, M. Kampffmeyer, I. Voiculescu, and E. Xing, "Federated partially supervised learning with limited decentralized medical images", *IEEE Transactions on Medical Imaging*, 2022.

[30] N. Dong and I. Voiculescu, "Federated contrastive learning for decentralized unlabeled medical images", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, 378–87.

[31] M. Elbatel, H. Wang, R. Mart, H. Fu, and X. Li, "Federated model aggregation via self-supervised priors for highly imbalanced medical image classification", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 334–46.

[32] G. Elmas, S. U. Dar, Y. Korkmaz, E. Ceyani, B. Susam, M. Ozbey, S. Avestimehr, and T. Çukur, "Federated learning of generative image priors for MRI reconstruction", *IEEE Transactions on Medical Imaging*, 42(7), 2022, 1996–2009.

[33] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach", *Advances in Neural Information Processing Systems*, 33, 2020, 3557–68.

[34] B. Feng, J. Shi, L. Huang, Z. Yang, S.-T. Feng, J. Li, Q. Chen, H. Xue, X. Chen, C. Wan, *et al.*, "Robustly federated learning model for identifying high-risk patients with postoperative gastric cancer recurrence", *Nature Communications*, 15(1), 2024, 742.

[35] C.-M. Feng, H. Fu, S. Yuan, and Y. Xu, "Multi-contrast MRI super-resolution via a multi-stage integration network", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, Springer, 2021, 140–9.

[36] C.-M. Feng, B. Li, X. Xu, Y. Liu, H. Fu, and W. Zuo, "Learning federated visual prompt in null space for mri reconstruction", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 8064–73.

[37] C.-M. Feng, Y. Yan, S. Wang, Y. Xu, L. Shao, and H. Fu, "Specificity-preserving federated learning for MR image reconstruction", *IEEE Transactions on Medical Imaging*, 42(7), 2022, 2010–21.

[38] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning", in *international conference on machine learning*, PMLR, 2016, 1050–9.

[39] T. Gao, X. Liu, Y. Yang, and G. Wang, "FEDMBP: Multi-Branch Prototype Federated Learning on Heterogeneous Data", in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, 2180–4.

[40] Z. Gao, F. Wu, W. Gao, and X. Zhuang, "A new framework of swarm learning consolidating knowledge from multi-center non-iid data for medical image segmentation", *IEEE Transactions on Medical Imaging*, 2022.

[41] J. Ge, G. Xu, J. Lu, C. Xu, Q. Z. Sheng, and X. Zheng, "FedAGA: A federated learning framework for enhanced inter-client relationship learning", *Knowledge-Based Systems*, 286, 2024, 111399.

[42] X. Gong, L. Song, R. Vedula, A. Sharma, M. Zheng, B. Planche, A. Innanje, T. Chen, J. Yuan, D. Doermann, *et al.*, "Federated learning with privacy-preserving ensemble attention distillation", *IEEE transactions on medical imaging*, 2022.

[43] Y. Gu, Q. Hu, X. Wang, Z. Zhou, and S. Lu, "FedACS: An efficient federated learning method among multiple medical institutions with adaptive client sampling", in *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, 2021, 1–6.

[44] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey", *Pattern Recognition*, 2024, 110424.

[45] G. N. Gunesli, M. Bilal, S. E. A. Raza, and N. M. Rajpoot, "A Federated Learning Approach to Tumor Detection in Colon Histology Images", *Journal of Medical Systems*, 47(1), 2023, 99.

[46] P. Guo, P. Wang, J. Zhou, S. Jiang, and 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, 2423–32.

[47] P. He, C. Lan, A. K. Bashir, D. Wu, R. Wang, R. Kharel, and K. Yu, "Low-Latency Federated Learning via Dynamic Model Partitioning for Healthcare IoT", *IEEE Journal of Biomedical and Health Informatics*, 2023.

[48] X. He, C. Tan, B. Liu, L. Si, W. Yao, L. Zhao, D. Liu, Q. Zhangli, Q. Chang, K. Li, *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, 1–5.

[49] U. D. of Health, H. Services, *et al.*, "HIPAA Privacy Rule and public health guidance from CDC and the US Department of Health and Human Services.", *MMWR: Morbidity & Mortality Weekly Report*, 52(17), 2003.

[50] A. Heidari, D. Javaheri, S. Toumaj, N. J. Navimipour, M. Rezaei, and M. Unal, "A new lung cancer detection method based on the chest CT images using Federated Learning and blockchain systems", *Artificial Intelligence in Medicine*, 141, 2023, 102572.

[51] S. M. Hosseini, M. Sikaroudi, M. Babaie, and H. Tizhoosh, "Proportionally fair hospital collaborations in federated learning of histopathology images", *IEEE transactions on medical imaging*, 2023.

[52] Y. Huang, W. Xie, M. Li, M. Cheng, J. Wu, W. Wang, J. You, and X. Liu, "Vicinal feature statistics augmentation for federated 3d medical volume segmentation", in *International Conference on Information Processing in Medical Imaging*, Springer, 2023, 360–71.

[53] M. Irfan, K. M. Malik, and K. Muhammad, "Federated fusion learning with attention mechanism for multi-client medical image analysis", *Information Fusion*, 108, 2024, 102364.

[54] D. Jatain, V. Singh, and N. Dahiya, "Blockchain Base Community Cluster-Federated Learning for Secure Aggregation of Healthcare Data", *Procedia Computer Science*, 215, 2022, 752–62.

[55] L. Jiang, L. Y. Ma, T. Y. Zeng, and S. H. Ying, "UFPS: A unified framework for partially annotated federated segmentation in heterogeneous data distribution", *Patterns*, 5(2), 2024.

[56] M. Jiang, X. Li, X. Zhang, M. Kamp, and Q. Dou, "Unifed: A unified framework for federated learning on non-iid image features", *arXiv preprint arXiv:2110.09974*, 2021.

[57] M. Jiang, H. R. Roth, W. Li, D. Yang, C. Zhao, V. Nath, D. Xu, Q. Dou, and Z. Xu, "Fair federated medical image segmentation via client contribution estimation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 16302–11.

[58] M. Jiang, H. Yang, C. Cheng, and Q. Dou, "IOP-FL: inside-outside Personalization for Federated medical image Segmentation", *IEEE Transactions on Medical Imaging*, 2023.

[59] A. Jiménez-Sánchez, M. Tardy, M. A. G. Ballester, D. Mateus, and G. Piella, "Memory-aware curriculum federated learning for breast cancer classification", *Computer Methods and Programs in Biomedicine*, 229, 2023, 107318.

[60] R. Jin and X. Li, "Backdoor attack and defense in federated generative adversarial network-based medical image synthesis", *Medical Image Analysis*, 90, 2023, 102965.

[61] I. Joshi, P. Upadhya, G. K. Nayak, P. Schüffler, and N. Navab, "DISBELIEVE: Distance Between Client Models Is Very Essential for Effective Local Model Poisoning Attacks", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 297–310.

[62] Z. H. Kafshgari, C. Shiranthika, P. Saeedi, and I. V. Bajić, "Quality-adaptive split-federated learning for segmenting medical images with inaccurate annotations", in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2023, 1–5.

[63] S. Kalra, J. Wen, J. C. Cresswell, M. Volkovs, and H. R. Tizhoosh, "Decentralized federated learning through proxy model sharing", *Nature communications*, 14(1), 2023, 2899.

[64] M. Kang, P. Chikontwe, S. Kim, K. H. Jin, E. Adeli, K. M. Pohl, and S. H. Park, "One-Shot Federated Learning on Medical Data Using Knowledge Distillation with Image Synthesis and Client Model Adaptation", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 521–31.

[65] H. Kassem, D. Alapatt, P. Mascagni, A. Karargyris, and N. Padoy, "Federated cycling (FedCy): Semi-supervised Federated Learning of surgical phases", *IEEE transactions on medical imaging*, 42(7), 2022, 1920–31.

[66] S. Kim, H. Park, M. Kang, K. H. Jin, E. Adeli, K. M. Pohl, and S. H. Park, "Federated learning with knowledge distillation for multi-organ segmentation with partially labeled datasets", *Medical Image Analysis*, 95, 2024, 103156.

[67] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 4015–26.

[68] F. Kong, X. Wang, J. Xiang, S. Yang, X. Wang, M. Yue, J. Zhang, J. Zhao, X. Han, Y. Dong, *et al.*, "Federated attention consistent learning models for prostate cancer diagnosis and Gleason grading", *Computational and Structural Biotechnology Journal*, 23, 2024, 1439–49.

[69] N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience", *Frontiers in systems neuroscience*, 2, 2008, 249.

[70] S. Kumar, A. Lakshminarayanan, K. Chang, F. Guretno, I. H. Mien, J. Kalpathy-Cramer, P. Krishnaswamy, and P. Singh, "Towards more efficient data valuation in healthcare federated learning using ensembling", in *International Workshop on Distributed, Collaborative, and Federated Learning*, Springer, 2022, 119–29.

[71] B. Lei, Y. Liang, J. Xie, Y. Wu, E. Liang, Y. Liu, P. Yang, T. Wang, C. Liu, J. Du, *et al.*, "Hybrid federated learning with brain-region attention network for multi-center Alzheimer's disease detection", *Pattern Recognition*, 153, 2024, 110423.

[72] B. Lei, Y. Zhu, E. Liang, P. Yang, S. Chen, H. Hu, H. Xie, Z. Wei, F. Hao, X. Song, *et al.*, "Federated Domain Adaptation via Transformer for Multi-site Alzheimer's Disease Diagnosis", *IEEE Transactions on Medical Imaging*, 2023.

[73] H. Li, C. Li, J. Wang, A. Yang, Z. Ma, Z. Zhang, and D. Hua, "Review on security of federated learning and its application in healthcare", *Future Generation Computer Systems*, 144, 2023, 271–90.

[74] M. Li and G. Yang, "Where to Begin? From Random to Foundation Model Instructed Initialization in Federated Learning for Medical Image Segmentation", *arXiv preprint arXiv:2311.15463*, 2023.

[75] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization", *arXiv preprint arXiv:2102.07623*, 2021.

[76] X. Li, L. Peng, Y. Wang, and W. Zhang, "Open Challenges and Opportunities in Federated Foundation Models Towards Biomedical Healthcare", *arXiv preprint arXiv:2405.06784*, 2024.

[77] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation", *Pattern Recognition*, 80, 2018, 109–17.

[78] Y. Li, Y. Feng, and Q. Qian, "FDPBoost: Federated differential privacy gradient boosting decision trees", *Journal of Information Security and Applications*, 74, 2023, 103468.

[79] L. Lin, Y. Liu, J. Wu, P. Cheng, Z. Cai, K. K. Wong, and X. Tang, "FedLPPA: Learning Personalized Prompt and Aggregation for Federated Weakly-supervised Medical Image Segmentation", *arXiv preprint arXiv:2402.17502*, 2024.

[80] C. Liu, Y. Luo, Y. Xu, and B. Du, "Foundation models matter: federated learning for multi-center tuberculosis diagnosis via adaptive regularization and model-contrastive learning", *World Wide Web*, 27(3), 2024, 1–17.

[81] C. Liu, Y. Luo, Y. Xu, and B. Du, "HPFL: hyper-network guided personalized federated learning for multi-center tuberculosis chest x-ray diagnosis", *Multimedia Tools and Applications*, 2023, 1–15.

[82] D. Liu, M. Cabezas, D. Wang, Z. Tang, L. Bai, G. Zhan, Y. Luo, K. Kyle, L. Ly, J. Yu, *et al.*, "Multiple sclerosis lesion segmentation: revisiting weighting mechanisms for federated learning", *Frontiers in Neuroscience*, 17, 2023, 1167612.

[83] F. Liu and F. Yang, "Federated Semi-supervised Medical Image Segmentation Based on Asynchronous Transmission", in *International Conference on Intelligent Computing*, Springer, 2023, 55–66.

[84] F. Liu and F. Yang, "Medical Image Segmentation Based on Federated Distillation Optimization Learning on Non-IID Data", in *International Conference on Intelligent Computing*, Springer, 2023, 347–58.

[85] Q. Liu, C. Chen, J. Qin, Q. Dou, and 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, 1013–23.

[86] Q. Liu, H. Yang, Q. Dou, and P.-A. Heng, "Federated semi-supervised medical image classification via inter-client relation matching", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, Springer, 2021, 325–35.

[87] Y. Liu, G. Luo, and Y. Zhu, "FedFMS: Exploring Federated Foundation Models for Medical Image Segmentation", *arXiv preprint arXiv:2403.05408*, 2024.

[88] Z. Liu, F. Wu, Y. Wang, M. Yang, and X. Pan, "FedCL: Federated contrastive learning for multi-center medical image classification", *Pattern Recognition*, 143, 2023, 109739.

[89] S. Lu, Z. Liu, T. Liu, and W. Zhou, "Scaling-up medical vision-and-language representation learning with federated learning", *Engineering Applications of Artificial Intelligence*, 126, 2023, 107037.

[90] J. Luo and S. Wu, "Adapt to adaptation: Learning personalization for cross-silo federated learning", in *IJCAI: proceedings of the conference*, Vol. 2022, NIH Public Access, 2022, 2166.

[91] J. Lyu, Y. Tian, Q. Cai, C. Wang, and J. Qin, "Adaptive channel-modulated personalized federated learning for magnetic resonance image reconstruction", *Computers in Biology and Medicine*, 165, 2023, 107330.

[92] B. Ma, Y. Feng, G. Chen, C. Li, and Y. Xia, "Federated adaptive reweighting for medical image classification", *Pattern Recognition*, 144, 2023, 109880.

[93] M. Manthe, S. Duffner, and C. Lartizien, "Whole brain radiomics for clustered federated personalization in brain tumor segmentation", in *Medical Imaging with Deep Learning*, PMLR, 2024, 957–77.

[94] M. Mazher, I. Razzak, A. Qayyum, M. Tanveer, S. Beier, T. Khan, and S. A. Niederer, "Self-supervised spatial–temporal transformer fusion based federated framework for 4D cardiovascular image segmentation", *Information Fusion*, 106, 2024, 102256.

[95] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data", in *Artificial intelligence and statistics*, PMLR, 2017, 1273–82.

[96] T. Mou, X. Jiang, J. Li, B. Yan, Q. Chen, T. Zhang, W. Huang, C. Gao, and Y. Chen, "FedTAM: Decentralized Federated Learning with a Feature Attention Based Multi-teacher Knowledge Distillation for Healthcare", in *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)*, IEEE, 2023, 1246–53.

[97] W. Moulahi, I. Jdey, T. Moulahi, M. Alawida, and A. Alabdulatif, "A blockchain-based federated learning mechanism for privacy preservation of healthcare IoT data", *Computers in Biology and Medicine*, 167, 2023, 107630.

[98] J. Mu, M. Kadoch, T. Yuan, W. Lv, Q. Liu, and B. Li, "Explainable federated medical image analysis through causal learning and blockchain", *IEEE Journal of Biomedical and Health Informatics*, 2024.

[99] E. Mushtaq, Y. F. Bakman, J. Ding, and S. Avestimehr, "Federated alternate training (FAT): Leveraging unannotated data silos in federated segmentation for medical imaging", in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2023, 1–5.

[100] S. Nazir and M. Kaleem, "Federated learning for medical image analysis with deep neural networks", *Diagnostics*, 13(9), 2023, 1532.

[101] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, and A. Y. Zomaya, "Federated learning for COVID-19 detection with generative adversarial networks in edge cloud computing", *IEEE Internet of Things Journal*, 9(12), 2021, 10257–71.

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

[103] S. Park and J. C. Ye, "Multi-task distributed learning using vision transformer with random patch permutation", *IEEE Transactions on Medical Imaging*, 2022.

[104] D. Peketi, V. Chalavadi, C. K. Mohan, and Y. W. Chen, "FLWGAN: Federated Learning with Wasserstein Generative Adversarial Network for Brain Tumor Segmentation", in *2023 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2023, 1–8.

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

[106] J.-F. Rajotte, S. Mukherjee, C. Robinson, A. Ortiz, C. West, J. M. L. Ferres, and R. T. Ng, "Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary", in *Proceedings of the conference on information Technology for Social Good*, 2021, 79–84.

[107] S. Rani, A. Kataria, S. Kumar, and P. Tiwari, "Federated learning for secure IoMT-applications in smart healthcare systems: A comprehensive review", *Knowledge-based systems*, 274, 2023, 110658.

[108] G. D. P. Regulation, "GDPR. 2019", 2019.

[109] M. H. u. Rehman, W. Hugo Lopez Pinaya, P. Nachev, J. T. Teo, S. Ourselin, and M. J. Cardoso, "Federated learning for medical imaging radiology", *The British Journal of Radiology*, 96(1150), 2023, 20220890.

[110] P. Riedel, R. von Schwerin, D. Schaudt, A. Hafner, and C. Späte, "ResNetFed: Federated Deep Learning Architecture for Privacy-Preserving Pneumonia Detection from COVID-19 Chest Radiographs", *Journal of Healthcare Informatics Research*, 7(2), 2023, 203–24.

[111] D. Sachin, B. Annappa, S. Ambasange, and A. E. Tony, "A Multimodal Contrastive Federated Learning for Digital Healthcare", *SN Computer Science*, 4(5), 2023, 674.

[112] P. Saha, D. Mishra, and J. A. Noble, "Rethinking Semi-Supervised Federated Learning: How to co-train fully-labeled and fully-unlabeled client imaging data", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 414–24.

[113] S. S. Sandhu, H. T. Gorji, P. Tavakolian, K. Tavakolian, and A. Akhbardeh, "Medical Imaging Applications of Federated Learning", *Diagnostics*, 13(19), 2023, 3140.

[114] L. Shao, B. Chen, Z. Zhang, Z. Zhang, and X. Chen, "Artificial intelligence generated content (AIGC) in medicine: A narrative review", *Mathematical Biosciences and Engineering*, 21(1), 2024, 1672–711.

[115] L. S. Shapley *et al.*, "A value for n-person games", 1953.

[116] S. Sharma and K. Guleria, "A comprehensive review on federated learning based models for healthcare applications", *Artificial Intelligence in Medicine*, 146, 2023, 102691.

[117] Y. Shen, A. Sowmya, Y. Luo, X. Liang, D. Shen, and J. Ke, "A federated learning system for histopathology image analysis with an orchestral stain-normalization gan", *IEEE Transactions on Medical Imaging*, 2022.

[118] I. Shiri, Y. Salimi, M. Maghsudi, E. Jenabi, S. Harsini, B. Razeghi, S. Mostafaei, G. Hajianfar, A. Sanaat, E. Jafari, *et al.*, "Differential privacy preserved federated transfer learning for multi-institutional 68Ga-PET image artefact detection and disentanglement", *European journal of nuclear medicine and molecular imaging*, 51(1), 2023, 40–53.

[119] S. Singh, S. Rathore, O. Alfarraj, A. Tolba, and B. Yoon, "A framework for privacy-preservation of IoT healthcare data using Federated Learning and blockchain technology", *Future Generation Computer Systems*, 129, 2022, 380–8.

[120] I. Siniosoglou, V. Argyriou, P. Sarigiannidis, T. Lagkas, A. Sarigiannidis, S. K. Goudos, and S. Wan, "Post-processing fairness evaluation of federated models: An unsupervised approach in healthcare", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(4), 2023, 2518–29.

[121] M. F. Sohan and A. Basalamah, "A systematic review on federated learning in medical image analysis", *IEEE Access*, 11, 2023, 28628–44.

[122] R. Tang, H. Liang, Y. Guo, Z. Li, Z. Liu, X. Lin, Z. Yan, J. Liu, X. Xu, W. Shao, *et al.*, "Pan-mediastinal neoplasm diagnosis via nationwide federated learning: a multicentre cohort study", *The Lancet Digital Health*, 5(9), 2023, e560–e570.

[123] S. Tayebi Arasteh, P. Isfort, M. Saehn, G. Mueller-Franzes, F. Khader, J. N. Kather, C. Kuhl, S. Nebelung, and D. Truhn, "Collaborative training of medical artificial intelligence models with non-uniform labels", *Scientific Reports*, 13(1), 2023, 6046.

[124] Z. L. Teo, L. Jin, S. Li, D. Miao, X. Zhang, W. Y. Ng, T. F. Tan, D. M. Lee, K. J. Chua, J. Heng, *et al.*, "Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture", *Cell Reports Medicine*, 2024.

[125] F. Ullah, G. Srivastava, H. Xiao, S. Ullah, J. C.-W. Lin, and Y. Zhao, "A scalable federated learning approach for collaborative smart healthcare systems with intermittent clients using medical imaging", *IEEE Journal of Biomedical and Health Informatics*, 2023.

[126] B. Wang, H. Li, Y. Guo, and J. Wang, "PPFLHE: A privacy-preserving federated learning scheme with homomorphic encryption for healthcare data", *Applied Soft Computing*, 146, 2023, 110677.

[127] D. Wang, C. Han, Z. Zhang, T. Zhai, H. Lin, B. Yang, Y. Cui, Y. Lin, Z. Zhao, L. Zhao, *et al.*, "FedDUS: Lung tumor segmentation on CT images through federated semi-supervised with dynamic update strategy", *Computer Methods and Programs in Biomedicine*, 249, 2024, 108141.

[128] J. Wang, Y. Jin, D. Stoyanov, and L. Wang, "FedDP: Dual personalization in federated medical image segmentation", *IEEE Transactions on Medical Imaging*, 2023.

[129] J. Wang, G. Xie, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, and Y. Jin, "FedMed-GAN: Federated domain translation on unsupervised cross-modality brain image synthesis", *Neurocomputing*, 546, 2023, 126282.

[130] M. Wang, L. Wang, X. Xu, K. Zou, Y. Qian, R. S. M. Goh, Y. Liu, and H. Fu, "Federated Uncertainty-Aware Aggregation for Fundus Diabetic Retinopathy Staging", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 222–32.

[131] R. Wang, X. Yuan, Z. Yang, Y. Wan, M. Luo, and D. Wu, "RFLPV: A robust federated learning scheme with privacy preservation and verifiable aggregation in IoMT", *Information Fusion*, 102, 2024, 102029.

[132] W. Wang, X. Li, X. Qiu, X. Zhang, V. Brusic, and J. Zhao, "A privacy preserving framework for federated learning in smart healthcare systems", *Information Processing & Management*, 60(1), 2023, 103167.

[133] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey", *Machine Intelligence Research*, 20(4), 2023, 447–82.

[134] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 2097–106.

[135] J. Wicaksana, Z. Yan, and K.-T. Cheng, "FCA: taming long-tailed federated medical image classification by classifier anchoring", *arXiv preprint arXiv:2305.00738*, 2023.

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

[137] B. Wu, S. Lyu, and B. Ghanem, "Ml-mg: Multi-label learning with missing labels using a mixed graph", in *Proceedings of the IEEE international conference on computer vision*, 2015, 4157–65.

[138] H. Wu, B. Zhang, C. Chen, and J. Qin, "Federated semi-supervised medical image segmentation via prototype-based pseudo-labeling and contrastive learning", *IEEE Transactions on Medical Imaging*, 2023.

[139] J. C.-H. Wu, H.-W. Yu, T.-H. Tsai, and H. H.-S. Lu, "Dynamically Synthetic Images for Federated Learning of medical images", *Computer Methods and Programs in Biomedicine*, 242, 2023, 107845.

[140] N. Wu, Z. Sun, Z. Yan, and L. Yu, "FedA3I: Annotation Quality-Aware Aggregation for Federated Medical Image Segmentation against Heterogeneous Annotation Noise", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 14, 2024, 15943–51.

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

[142] X. Wu, J. Pei, X.-H. Han, Y.-W. Chen, J. Yao, Y. Liu, Q. Qian, and Y. Guo, "FedEL: Federated ensemble learning for non-iid data", *Expert Systems with Applications*, 237, 2024, 121390.

[143] Y. Wu, D. Zeng, Z. Wang, Y. Sheng, L. Yang, A. J. James, Y. Shi, and J. Hu, "Federated self-supervised contrastive learning and masked autoencoder for dermatological disease diagnosis", *arXiv preprint arXiv:2208.11278*, 2022.

[144] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu, "Federated contrastive learning for volumetric medical image segmentation", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, Springer, 2021, 367–77.

[145] A. Xu, W. Li, P. Guo, D. Yang, H. R. Roth, A. Hatamizadeh, C. Zhao, D. Xu, H. Huang, and 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, 20866–75.

[146] R. Yan, L. Qu, Q. Wei, S.-C. Huang, L. Shen, D. L. Rubin, L. Xing, and Y. Zhou, "Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging", *IEEE Transactions on Medical Imaging*, 42(7), 2023, 1932–43.

[147] Y. Yan, C.-M. Feng, Y. Li, R. S. M. Goh, and L. Zhu, "Federated Pseudo Modality Generation for Incomplete Multi-Modal MRI Reconstruction", *arXiv preprint arXiv:2308.10910*, 2023.

[148] Y. Yan, H. Wang, Y. Huang, N. He, L. Zhu, Y. Xu, Y. Li, and Y. Zheng, "Cross-modal vertical federated learning for mri reconstruction", *IEEE Journal of Biomedical and Health Informatics*, 2024.

[149] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, and K.-T. Cheng, "Variation-aware federated learning with multi-source decentralized medical image data", *IEEE Journal of Biomedical and Health Informatics*, 25(7), 2020, 2615–28.

[150] Q. Yang, J. Zhang, W. Hao, G. P. Spell, and 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, 3845–53.

[151] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications", *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2), 2019, 1–19.

[152] Q. Yang, Q. Zhu, M. Wang, W. Shao, Z. Zhang, and D. Zhang, "Self-Supervised Federated Adaptation for Multi-Site Brain Disease Diagnosis", *IEEE Transactions on Big Data*, 2023.

[153] Y. Yang, X. Liu, T. Gao, X. Xu, P. Zhang, and G. Wang, "Dense Contrastive-based Federated Learning for Dense Prediction Tasks on Medical Images", *IEEE Journal of Biomedical and Health Informatics*, 2024.

[154] Z. Yang, Y. Chen, H. Huangfu, M. Ran, H. Wang, X. Li, and Y. Zhang, "Dynamic corrected split federated learning with homomorphic encryption for U-shaped medical image networks", *IEEE Journal of Biomedical and Health Informatics*, 2023.

[155] G. Yue, P. Wei, T. Zhou, Y. Song, C. Zhao, T. Wang, and B. Lei, "Specificity-aware Federated Learning with Dynamic Feature Fusion Network for Imbalanced Medical Image Classification", *IEEE Journal of Biomedical and Health Informatics*, 2023.

[156] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization", *arXiv preprint arXiv:1710.09412*, 2017.

[157] R. Zhang, Z. Fan, Q. Xu, J. Yao, Y. Zhang, and Y. Wang, "Grace: A generalized and personalized federated learning method for medical imaging", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 14–24.

[158] L. Zhao and J. Huang, "A distribution information sharing federated learning approach for medical image data", *Complex & Intelligent Systems*, 9(5), 2023, 5625–36.

[159] Y. Zhao, Q. Liu, X. Liu, and K. He, "Medical Federated Model with Mixture of Personalized and Sharing Components", *arXiv preprint arXiv:2306.14483*, 2023.

[160] Z. Zheng, Y. Hayashi, M. Oda, T. Kitasaka, K. Misawa, and K. Mori, "Federated 3D multi-organ segmentation with partially labeled and unlabeled data", *International Journal of Computer Assisted Radiology and Surgery*, 2024, 1–14.

[161] B. Zhou, T. Miao, N. Mirian, X. Chen, H. Xie, Z. Feng, X. Guo, X. Li, S. K. Zhou, J. S. Duncan, *et al.*, "Federated transfer learning for low-dose PET denoising: a pilot study with simulated heterogeneous data", *IEEE transactions on radiation and plasma medical sciences*, 7(3), 2022, 284–95.

[162] B. Zhou, H. Xie, Q. Liu, X. Chen, X. Guo, Z. Feng, J. Hou, S. K. Zhou, B. Li, A. Rominger, *et al.*, "FedFTN: Personalized federated learning with deep feature transformation network for multi-institutional low-count PET denoising", *Medical image analysis*, 90, 2023, 102993.

[163] J. Zhou, L. Zhou, D. Wang, X. Xu, H. Li, Y. Chu, W. Han, and X. Gao, "Personalized and privacy-preserving federated heterogeneous medical image analysis with pppml-hmi", *Computers in Biology and Medicine*, 169, 2024, 107861.

[164] J. Zhou, L. Zhou, D. Wang, X. Xu, H. Li, Y. Chu, W. Han, and X. Gao, "Personalized and privacy-preserving federated heterogeneous medical image analysis with PPPML-HMI (preprint)", 2023.

[165] S. K. Zhou, H. Greenspan, and D. Shen, *Deep learning for medical image analysis*, Academic Press, 2023.

[166] X. Zhou, W. Huang, W. Liang, Z. Yan, J. Ma, Y. Pan, I. Kevin, and K. Wang, "Federated distillation and blockchain empowered secure knowledge sharing for Internet of medical Things", *Information Sciences*, 662, 2024, 120217.

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

[168] M. Zhu, J. Liao, J. Liu, and Y. Yuan, "FedOSS: Federated Open Set Recognition via Inter-client Discrepancy and Collaboration", *IEEE Transactions on Medical Imaging*, 2023.