



**HAL**  
open science

# Federated brain tumor segmentation: an extensive benchmark

Matthis Manthe, Stefan Duffner, Carole Lartizien

► **To cite this version:**

Matthis Manthe, Stefan Duffner, Carole Lartizien. Federated brain tumor segmentation: an extensive benchmark. *Medical Image Analysis*, 2024, 97, pp.103270. 10.1016/j.media.2024.103270 . hal-04719581

**HAL Id: hal-04719581**

**<https://hal.science/hal-04719581v1>**

Submitted on 3 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## Federated brain tumor segmentation: an extensive benchmark

Matthis Manthe<sup>a,b,\*</sup>, Stefan Duffner<sup>b</sup>, Carole Lartizien<sup>a</sup>

<sup>a</sup>Univ Lyon, CNRS, INSA Lyon, UCBL, Inserm, CREATIS UMR 5220, U1294, F-69621 Villeurbanne, France

<sup>b</sup>Univ Lyon, INSA Lyon, CNRS, UCBL, Centrale Lyon, Univ Lyon 2, LIRIS, UMR5205, F-69621 Villeurbanne, France

### ARTICLE INFO

Article history:

**Keywords:** Deep learning, Federated learning, Medical image segmentation, BraTS, Personalized federated learning, Clustered federated learning

### ABSTRACT

Recently, federated learning has raised increasing interest in the medical image analysis field due to its ability to aggregate multi-center data with privacy-preserving properties. A large amount of federated training schemes have been published, which we categorize into global (one final model), personalized (one model per institution) or hybrid (one model per cluster of institutions) methods. However, their applicability on the recently published Federated Brain Tumor Segmentation 2022 dataset has not been explored yet. We propose an extensive benchmark of federated learning algorithms from all three classes on this task. While standard FedAvg already performs very well, we show that some methods from each category can bring a slight performance improvement and potentially limit the final model(s) bias toward the predominant data distribution of the federation. Moreover, we provide a deeper understanding of the behaviour of federated learning on this task through alternative ways of distributing the pooled dataset among institutions, namely an Independent and Identical Distributed (IID) setup, and a limited data setup. Our code is available at ([https://github.com/MatthisManthe/Benchmark\\_FeTS2022](https://github.com/MatthisManthe/Benchmark_FeTS2022)).

© 2024 Elsevier B. V. All rights reserved.

### 1. Introduction

Deep learning showed state-of-the-art performance on a variety of medical image analysis tasks Baur et al. (2019); Futrega et al. (2022); Maier et al. (2017); Zeineldin et al. (2022). However, training precise deep learning models requires a large amount of sensitive and high-quality data. Data centralization possibilities are limited: restrictive legislations such as the General Data Protection Regulation (GDPR) in Europe, the overall sensitivity of medical images, as well as the high labelling cost of each exam by clinical expert, make the aggregation of large amounts of data complicated if not prohibited.

Federated learning was defined in McMahan et al. (2017) with the seminal algorithm called *Federated Averaging*

(*FedAvg*) as a decentralized privacy-preserving machine learning paradigm enabling numerous institutions to collaboratively train a model on their cumulated data without ever sharing them. Even if the privacy of standard federated learning methods is a field of research in itself Rodríguez-Barroso et al. (2023); Yang et al. (2023); Zhu et al. (2019), classical federated learning methods are a significant step towards effective decentralized training conformed with privacy and ethical requirements.

We have seen a growing interest in federated learning in the medical field as it could drastically increase the data quantity used to train a model at a limited privacy cost. It became a popular research topic on a variety of tasks, from fMRI analysis Li et al. (2020c) to brain anomaly detection Bercea et al. (2022), based on fully Qu et al. (2022a), weakly Lu et al. (2022); Yang et al. (2021), or un- Wu et al. (2022) supervised models, even mixing these levels of supervision Wicaksana et al. (2022).

\*Corresponding author:  
e-mail: [matthis.manthe@insa-lyon.fr](mailto:matthis.manthe@insa-lyon.fr) (Matthis Manthe)

There are overall more and more incentives to perform multi-centric research Mårtensson *et al.* (2020); Wachinger *et al.* (2021), as well as more available public datasets to do so Bilic *et al.* (2023). In particular, the Federated brain Tumor Segmentation (FeTS) initiative published the institution-wise partitioning of the BraTS2021 dataset Baid *et al.* (2021); Karargyris *et al.* (2021); Pati *et al.* (2021); Reina *et al.* (2022), composed of around 1200 multi-modal brain MRI scans of gliomas from 23 clinical institutions along with their multi-label tumor segmentation masks including the whole tumor outline as well as the enhancing and necrotic areas. A significant amount of pioneering works showed that the federated learning paradigm was attractive on this task Sheller *et al.* (2020); Pati *et al.* (2022), with two FeTS challenges organized in 2021 and 2022 Pati *et al.* (2021); Isik-Polat *et al.* (2022); Khan *et al.* (2022); Mächler *et al.* (2021), as well as some independent works exploring other aspects of decentralized learning on this valuable segmentation task Camajori Tedeschini *et al.* (2022); Li *et al.* (2019).

In the past few years, a very large amount of research work has been done by the machine learning community on the general topic of federated learning addressing different limitations. It has been clearly identified now that the statistical distribution of data across institutions, potentially heterogeneous, can drastically alter the performance of standard federated learning methods such as *FedAvg* Li *et al.* (2020a); Karimireddy *et al.* (2020). On this specific problem, a large number of lines of research have been defined, which can be categorized into three classes: global methods McMahan *et al.* (2017), i.e. training one model generalizing on the data of every institution; personalized methods Li *et al.* (2021b), i.e. training one personalized model for each institution while each benefiting from the federation; and hybrid methods Sattler *et al.* (2021), i.e. building a set of models each designed to perform well on a set of institution or samples. Most of these state-of-the-art (SOTA) methods are outside of the scope of previous works on BraTS dataset and have never been explored on the brain tumor segmentation task yet. Moreover, novel federated methods proposed in both the medical image analysis and machine learning communities mainly focus on patient-level classification tasks, with only few works on segmentation tasks.

We propose a benchmark of several global, personalized and hybrid state-of-the-art federated learning methods on the FeTS2022 dataset. Its purpose is to first investigate the performance of standard federated learning methods on the latest version of the dataset, as well as to highlight which class of methods tend to give better performances considering its data distribution. We complement existing works on FeTS dataset by exploring the application of these methods in a significantly different federated setup, from chosen local optimizer to computational constraints. The main contributions of our work are as follows

- We propose, to the best of our knowledge, the first benchmark of personalized and clustered federated learning methods on brain tumor segmentation, showing the potential of both of these paradigms for this task,
- We fixed a common estimated time budget enabling

the fastest algorithms to reach their validation plateau, analysing results close to convergence.

- We analyse the performance biases between institutions brought by each federated optimizer.
- We defined each method in a common formal framework with some adaptations for our federated setup, and publicly available implementations.

Moreover, we explore additional synthetic partitions of the dataset, namely a randomly redistributed version as well as a version with a limited amount of data per institution with a more controlled heterogeneity. We bring with these secondary experiments a better understanding of the behaviour of federated learning methods on this task.

In section 2, we detail the current trends in both the federated learning community and its application to neuro-image analysis. In section 3, we motivate the choice of each method included in this benchmark and describe them. We define the experimental setup in section 4, and explore the results in section 5. We discuss the choices made as well as the limitations of the benchmark and conclude in section 6.

## 2. Related work

### 2.1. Federated learning

The first published federated learning algorithm was *Federated Averaging (FedAvg)* McMahan *et al.* (2017), where each participating institution performs multiple gradient steps locally on a received model between server-side averaging steps. We can underline two main research axes devoted to the improvement of this algorithm in terms of convergence speed and the final accuracy of the obtained global model.

The first one is to perceive the aggregation server-side as the approximation of a gradient, thus exploring federated algorithms as a branch of standard stochastic optimization theory. Thus was explored, for example, the addition of server-side learning rate, momentum and adaptivity Jhunjunwala *et al.* (2023); Reddi *et al.* (2022). A second line of work is to perceive the aggregation step as a model fusion task. Some authors proposed to match neurons or channels between local models before averaging the weights by the Hungarian algorithm (*FedMA* Wang *et al.* (2020a)) or graph matching Liu *et al.* (2022), or by allocating semantic channel positions prior to training in the network for a better averaging (*Fed<sup>2</sup>* Yu *et al.* (2021)).

It was quickly shown that the statistical heterogeneity of local datasets could harm the convergence the more local steps are performed between server synchronizations Wang *et al.* (2020b); Zhao *et al.* (2018). Numerous authors proposed improvements over *FedAvg* in this specific case. Most of them alter the local optimization to limit the "client drift", namely the quick overfitting of local optimizations towards local optima, negating the efficiency of server aggregations. Proximal regularization was first proposed (*FedProx* Li *et al.* (2020a)), followed by several important methods trying to guide local optimizations away from local optima. Examples are *SCAF-FOLD*, leveraging control variates to promote the convergence

of local optimizations toward the global optima Karimireddy et al. (2020), or *MOON* based on a local model contrastive loss Li et al. (2021a). Other authors explored the application of robust optimization methods locally, such as adaptations of Elastic Weight Consolidation (*FedCurv* Shoham et al. (2019)) or generative replay (*FedReplay* Qu et al. (2022a)) from the continual learning field, or employed local sample synthesis Tang et al. (2022); Zhu and Luo (2022). Some authors also used server-side ensemble distillation from auxiliary data to improve the aggregated global model before sending it to participants (*FedDF* Lin et al. (2021) or *FedBE* Chen and Chao (2020)). Finally, some work focuses on neural architectures that are more suited to federated learning, showing that batch normalization is not adapted to local distribution shifts Li et al. (2021c); Wang et al. (2023) or that transformer models might be easier to train in a federated setup than convolutional ones Qu et al. (2022b).

## 2.2. Personalized and hybrid federated learning

In parallel to the previously described methods, some authors suggested that training a plurality of models instead of training a single global model performing well on the data of every institution could help bridge the gap between centralized and federated performance in the case of high statistical heterogeneity.

The first line of research, named **personalization**, aims at training one model per institution each performing better than a global model obtained by standard federated learning and than what could be obtained by isolated institutions. As a straightforward example, personalization can be performed by simply fine-tuning locally a model obtained by *FedAvg*, which was shown quite efficient Mansour et al. (2020). The idea of partial model sharing quickly followed, with the earliest methods retaining local the first (*LG-FedAvg* Liang et al. (2020)) or the last (*FedPer* Arivazhagan et al. (2019)) local layers, or the batch normalization parameters (*FedBN* Li et al. (2020e)) of a neural network. Numerous works then followed this line of work (*CD<sup>2</sup>-FedAvg* Shen et al. (2022), *PartialFixedShare* Pillutla et al. (2022), *PartialFed* Sun et al. (2021)).

The personalization task can be linked to a large number of subfields in machine learning, and multiple formulations were proposed. Authors interpreted personalization in the meta-learning framework, with the objective of training a server-side meta-model easily fine-tunable on local datasets. In that vein, adaptations of the model-agnostic meta-learning algorithm (*MAML* Finn et al. (2017)) were explored (*PerFedAvg* Fallah et al. (2020) or *pFedMe* Dinh et al. (2020)). Alternatively, each local dataset of a federation can be viewed as a task in the multi-task learning framework. Thus, authors tried to adapt multi-task regularization methods to the federated setup, such as *Ditto* Li et al. (2021b) or *FedEM* Marfoq et al. (2022a).

More specific to the federated setup, some researchers explored the combination of local and global methods to perform personalization. The first works explored the definition of personalization objectives as finding the best interpolation between local and global models (*Mapper* Mansour et al. (2020), *Soft-Pull* Xu et al. (2022)), while a recent work also proposed to use a combination of a global federated learning model and local kNN models to perform robust personalization (*kNN-Per* Mar-

foq et al. (2022b)). We can also note an interesting idea of leveraging a hypernetwork server-side to output the parameters (or a subset of them) of personalized models (*pFedHN* Shamsian et al. (2021)).

Finally, tangentially to multi-task adaptations, some authors explored personalization methods based on the computation of similarities between local datasets through different proxies, either using the loss of a local model on the datasets of other institutions (*FedFOMO* Zhang et al. (2021)), or distances in the parameter space (*FedAMP* Huang et al. (2021)) to use them as local aggregation weights. This implicit hypothesis of institutions sharing more similarity with a subset of the federation also motivated a refinement of the personalization framework into a **clustered** (or hybrid) one. In this framework, the objective is to train a set of models, each specialized for a set of institutions. Cluster federated learning also implies the computation of similarities between institutions: some authors experimented with the loss of the clustered models to incrementally compute clusters of institutions during training (*IFCA* Ghosh et al. (2020)), or computed clusters based on the cosine similarity matrix between local updates (*CFL* Sattler et al. (2021)).

## 2.3. Federated learning in brain tumor segmentation

Pioneering works already established the potential of federated learning in brain tumor segmentation. A first proof of concept was proposed in Sheller et al. (2020), where standard *FedAvg* was compared to cyclic weight transfer on BraTS 2018 dataset, highlighting that even in a cross-silo setting, the latter did not scale well due to catastrophic forgetting from one institution to the other, while the former reached a close match in performance with centralized training. This work was rapidly followed by experiments with differential privacy and local momentum restart Li et al. (2019). Furthermore, two editions of the Federated brain Tumor Segmentation Challenge, FeTS 2021 and 2022 were hosted Pati et al. (2021). They were dedicated to the impact analysis of institution sampling, frequency of server synchronisation and aggregation weights strategies in *FedAvg*. Several novel methods were proposed during this challenge. Isik-Polat et al. (2022); Khan et al. (2022); Mächler et al. (2021, 2023) Namely, *FedCostwAvg* Mächler et al. (2021) was the winner of 2021's edition, weighting the local updates by a linear combination of the local sample size and the local loss improvement from one round to another. The authors further improved their method on the second edition, named *FedPIDAvg* Mächler et al. (2023), by adding a third term in the weight of the local updates: a form of momentum on the local losses. Other participants experimented with *FedAvg-M*, *FedAdam*, robust aggregation schemes such as median, trimmed mean (discounting local updates that are too far from the median), Top-K mean (discounting local updates that did not sufficiently improve the local loss compared to others), and other forms of loss based server-side aggregations Rawat et al. (2022); Tuladhar et al.. Finally, other authors explored during the challenge the usage of adaptive number of local epochs per round and learning rate decay with *FedAvg* as well as *FedNova* and *FedAvg-M* Isik-Polat et al. (2022); Tuladhar et al.. Finally, an outstanding real world application of federated learning on 71 institutions

worldwide on brain tumor segmentation was recently published Pati et al. (2022). The authors showed that applying *FedAvg* on such a massive coalition of institutions significantly improves the performance of the final model compared to what could be obtained with only the public BraTS datasets.

In comparison to this first ever challenge in federated medical image segmentation, we made the choice to use a smaller 3D U-Net. Considering our limited computational capacity, we chose the lowest possible training time budget enabling some variants of *FedAvg* to reach a validation plateau, getting closer to convergence than FeTS challenges. We also focused on recently proposed new classes of SOTA methods such as personalization and clustered methods which were not implementable with the coding constraints of the challenges. We used standard SGD locally instead of Adam to stay consistent with the federated learning literature. In essence, we try to complement the results of the FeTS challenges with a different federated setup. Moreover, along with proposing rigorous experiments of SOTA methods, we hope that our work can help identifying valuable "types" of federated learning methods for large scale real-world implementations such as Pati et al. (2022).

### 3. Methods

As we could not experiment with every state-of-the-art federated learning methods, we selected the most representative and established ones from different groups of federated techniques (from adaptive and variance reduction-based global methods, finetuning-based or multi-task-based personalized methods, etc.). We define each of them with the potential adaptations which were required to apply them on the 3D brain tumor segmentation task of interest.

#### 3.1. Notations and centralized problem

Let  $K$  be the number of institutions each with a local dataset  $D_k := \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k}$ , with  $x_{k,i} \in \mathcal{X} = \mathbb{R}^{m \times h \times w \times d}$  the multimodal MRI scan to segment,  $y_{k,i} \in \mathcal{Y} = \{0, 1\}^{l \times h \times w \times d}$  its associated multi-label ground-truth segmentation map,  $n_k$  the local dataset size of institution  $k$  and  $N = \sum_{k=1}^K n_k$  the total number of samples. We note  $w \in \mathcal{W} = \mathbb{R}^p$  the parameters of the neural network to be optimized for the downstream task.

Given a loss function  $l : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ , the classical formulation of the neural network optimization problem in a centralized setup can be defined as

$$w^* = \underset{w \in \mathcal{W}}{\operatorname{argmin}} \sum_{i=1}^N l(w, x_i, y_i) \quad (1)$$

over the pooled dataset  $D = \bigcup_{k=1}^K D_k$ .

#### 3.2. Global federated learning

##### 3.2.1. Problem formulation

Under privacy constraints, pooling each institution's local dataset together on a single computation unit is prohibited. Federated learning was proposed (McMahan et al. (2017)) to train a neural network over decentralized data in a collaborative fashion such that raw data never leaves data warehouses of each

participating institution. The decentralized optimization problem can be defined as

$$w^* = \underset{w \in \mathcal{W}}{\operatorname{argmin}} \sum_{k=1}^K \sum_{i=1}^{n_k} l(w, x_{k,i}, y_{k,i}), \quad (2)$$

recovering the centralized one. Explored global federated algorithms are summarized in Algorithm 1.

##### 3.2.2. Federated Averaging

*Federated Averaging (FedAvg)* is a pioneer work proposed by McMahan et al. (2017). They proposed to organize the collaborative training process in several communication rounds. During each round  $t$ , the aggregation server sends the parameters  $w^t$  of the current global model to each participant. They perform  $E$  epochs of local training using  $w^t$  as an initialization to obtain a local update  $\Delta w_k^t = w_k^t - w^t$  (note that in most theoretical works,  $U$  gradient steps per institution are used instead). They then send these updates back to the server, which aggregates them in a single global update applied to the global model following

$$w^{t+1} = w^t + \sum_{k=1}^K p_k \Delta w_k^t, \quad (3)$$

with  $p_k$  an aggregation weight. Both uniform ( $p_k = \frac{1}{K}$ ) and weighted ( $p_k = \frac{n_k}{N}$ ) averaging were proposed, the latter introduced to ensure convergence towards the global objective of Equation 2.

As the term *FedAvg* encompasses all of these different variations, we specifically define *FedAvg with fixed local epochs* its variant with each institution performing  $E$  local epochs between each communication round and server-side weighted average aggregation, *FedAvg with fixed iterations* its variant with  $U$  local updates and weighted averaging and *FedAvg with uniform averaging* its variant with  $E$  local epochs and uniform averaging. Fixed batch size is set common to all institutions for all variants.

##### 3.2.3. FedNova

In the specific case of an unbalanced amount of samples per institution, the authors of Wang et al. (2020b) underlined an inconsistency in the optimization of *FedAvg* with a fixed number of local epochs per round and weighted aggregation (i.e.  $p_k = \frac{n_k}{N}$ ). With a unique fixed batch size, institutions owning more samples than others will perform more gradient steps, giving implicitly more weight to their local updates in the aggregation. This motivated the development of a framework to limit this problem for a large amount of SOTA methods, leveraging normalized gradient aggregation. In our study, we chose to apply their adaptation of *FedAvg*, following a simplification of the formulation in Isik-Polat et al. (2022). FedNova reduces to *FedAvg* with a uniform weighting scheme and an analytical server learning rate  $\gamma \geq 1$

$$\gamma = K \sum_{k=1}^K p_k^2 \quad (4)$$

where the aggregation becomes

$$w^{t+1} = w^t + \gamma \frac{1}{K} \sum_{k=1}^K \Delta w_k^t \quad (5)$$

### 3.2.4. SCAFFOLD

To reduce the variability of local updates and enforce convergence of local optimizations toward the global optima, Karimreddy et al. (2020) added the computation of local and global control variates used as a form of momentum during local optimizations. After each local training, an institution updates its local control variate  $c_k^t$  following

$$\Delta c_k^t = -c^t + \frac{1}{s_k \eta_l} \Delta w_k^t \quad (6)$$

$$c_k^{t+1} = c_k^t + \Delta c_k^t \quad (7)$$

with  $s_k$  the number of gradient steps performed by institution  $k$  and  $c^t$  a global control variate, and sends them to the server along with the computed local updates  $\Delta w_k^t$ . In addition to the classical aggregation of FedAvg, the server aggregates control variate updates into the global control variate following

$$c^{t+1} = c^t + \sum_{k=1}^K p_k \Delta c_k^t \quad (8)$$

which is sent back to the institutions. A local update step on a batch  $b_k$  of samples of institution  $k$  is altered by these control variates following Equation 9 (omitting local step and global round superscripts for readability).

$$w_k = w_k - \eta_l (\nabla l(w_k, b_k) - c_k + c). \quad (9)$$

Note that we use the adaptation to batch training of SCAFFOLD published in Reddi et al. (2022) to deal with imbalanced local dataset sizes.

### 3.2.5. FedAdam

Reddi et al. (2022) proposed to extend classical adaptive methods such as Adagrad, Adam and Yogi to federated learning by applying them server-side while giving proofs of their effectiveness in this context. For computational cost reasons, we chose to focus on FedAdam only. At aggregation step  $t$ , the server computes first and second-order momentums

$$\Delta^{t+1} = \beta_1 \Delta^t + (1 - \beta_1) \sum_{k=1}^K p_k \Delta w_k^t \quad (10)$$

$$v^{t+1} = \beta_2 v^t + (1 - \beta_2) \left( \sum_{k=1}^K p_k \Delta w_k^t \right)^2 \quad (11)$$

and aggregates local updates following

$$w^{t+1} = w^t + \eta_s \frac{\Delta^t}{\sqrt{v^t + \tau}}. \quad (12)$$

### 3.2.6. q-FedAvg

Soon after the publication of McMahan et al. (2017), Li et al. (2020b) explored the notion of fairness in federated learning. They showed that FedAvg could accentuate the disparities of performance of the global model between institutions. Extending the pioneering work of Mohri et al. (2019), they proposed to give more weight server-side to updates computed by institutions on which the global model at rounds  $t$  performs worse than others. Formally, they define the fair federated objective as follows

$$w^* = \underset{w \in \mathcal{W}}{\operatorname{argmin}} \sum_{k=1}^K \frac{p_k}{q+1} \left( \frac{1}{n_k} \sum_{i=1}^{n_k} l(w, x_i^k, y_i^k) \right)^{q+1} \quad (13)$$

with  $q \in \mathbb{R}^+$  a hyperparameter enabling to define the desired level of fairness ( $q = 0$  is equivalent to FedAvg, while  $q \rightarrow \infty$  gives agnostic federated learning Mohri et al. (2019), optimizing for the worse performing institution). To optimize this objective, the authors proposed a modification of FedAvg called *q-FedAvg*. Each participant weights its local update by the loss of the latest global model on its data to the power  $q$  before transmitting it to the server

$$\Delta_k^t = \frac{1}{\eta_l} F_k^q(w^t) \Delta w_k^t \quad (14)$$

where the local training loss of institution  $k$  is defined as follows

$$F_k(w) := \sum_{\{x_k, y_k\} \in D_k} l(w, x_k, y_k) \quad (15)$$

They also compute a regularization value  $h_k^t$ , used at the server-side in the aggregation step following

$$h_k^t = q F_k^{q-1}(w^t) \|\Delta w_k^t\|^2 + \frac{1}{\eta_l} F_k^q(w^t) \quad (16)$$

$$w^{t+1} = w^t + \frac{\sum_{k=1}^K \Delta_k^t}{\sum_{k=1}^K h_k^t} \quad (17)$$

### 3.2.7. FedPIDAvg

The winner of the 2022's edition of the FeTS challenge Pati et al. (2021) proposed FedPIDAvg Mächler et al. (2023), a variation of FedAvg with a server-side weighting strategy inspired by PID controllers as follows

$$w^{t+1} = w^t + \sum_{k=1}^K \left( \alpha p_k + \beta \frac{\Delta_k^t}{L^t} + \gamma \frac{m_k^t}{M^t} \right) \Delta w_k^t \quad (18)$$

with  $l_{val,k}^t$  the local validation loss of institution  $k$  at round  $t$ ,  $\Delta_k^t$  the local validation loss improvement,  $m_k^t$  a momentum term on these validation loss values and  $L^t$  and  $M^t$  normalization terms defined as

$$l_{val,k}^t = \sum_{i=1}^{n_k^{val}} l(w_k^{t-1}, x_i^k, y_i^k) \quad (19)$$

$$\Delta_k^t = l_{val,k}^{t-1} - l_{val,k}^t, \quad L^t = \sum_{k=1}^K \Delta_k^t \quad (20)$$

$$m_k^t = \sum_{j=0}^5 l_{val,k}^{t-j}, \quad M^t = \sum_{k=1}^K m_k^t \quad (21)$$

Since our federated setup differs from FeTS2022 challenge's one (c.f. Section 2.3), we had to slightly modify this method for it to converge properly, only accounting for positive validation loss differences in the weighted average to avoid some gradient "ascent" in the latest stages of training. This resulted in modification of  $\Delta l_k^t$  as

$$\Delta l_k^t = \max(0, l_{val,k}^{t-1} - l_{val,k}^t) \quad (22)$$

### 3.3. Personalized federated learning

#### 3.3.1. Problem formulation

In the context of statistical heterogeneity between institutions' datasets, personalized federated learning emerged as the optimization of one model per institution. The subsequent objective can be formalized as the union of the local objectives

$$w_k^* = \underset{w_k \in \mathcal{W}}{\operatorname{argmin}} \sum_{i=1}^{n_k} l(w_k, x_i^k, y_i^k), \forall k \in \{1, \dots, K\} \quad (23)$$

while each model  $w_k$  still benefits from the whole federation.

#### 3.3.2. Local training

The simplest approach fitting the latter objective is for each institution to train a model locally without any communication.

#### 3.3.3. Local finetuning

One can also perform personalization by first training a global model using a global solution such as *FedAvg*, which is then finetuned locally by each institution for several epochs.

#### 3.3.4. Ditto

In Li et al. (2021b), the authors inspired themselves by the field of multi-task learning, where the local dataset of each institution can be seen as a task. They define their regularized bi-level objective for institution  $k$  as follows

$$w_k^* = \underset{w_k \in \mathcal{W}}{\operatorname{argmin}} \left( \sum_{i=1}^{n_k} l(w_k, x_i^k, y_i^k) + \frac{\lambda}{2} \|w_k - w^*\|^2 \right) \quad (24)$$

such that  $w^*$  solves the global problem of Equation 2. They argue that this problem can be solved either by local finetuning over the regularized local objective of Equation 24, or by adding a personalization module in parallel to global training. Since they showed that the latter only provides benefits in case of the presence of malicious users while drastically increasing local computation loads, we implement *Ditto* as the former.

#### 3.3.5. Partial model sharing

A large amount of research work has been devoted to the exploration of partial model sharing in federated learning in the literature (Section 2). We chose to focus in this work to the most established ones, namely *FedPer* Arivazhagan et al. (2019) and *LG-FedAvg* Liang et al. (2020). Both share the same principle: the model parameters  $w$  are partitioned into federated ones  $w_g$  and local ones  $w_l$  such that  $w = w_g \cup w_l$ . Based upon *FedAvg*, local updates are computed similarly and only the federated weights  $w_{gk}$  of institution  $k$  are transmitted to the server,

**Algorithm 1:** Global federated learning algorithms (FedAvg, FedAdam, SCAFFOLD, FedNova, q-FedAvg and FedPIDAvg)

#### 1 Global Federated Learning (Server-side)

**Input:**  $w^0, T, E, \eta_l, \eta_s, \beta_1, \beta_2, \tau, q, \alpha, \beta, \gamma$

2  $\gamma \leftarrow K \sum_{k=1}^K p_k^2$

3 **for**  $t \leftarrow 1$  **to**  $T$  **do**

4     Send  $w^t, c^t$  to each institution

5     **for**  $k \leftarrow 1$  **to**  $K$  **do**

6          $\Delta w_k^t, \Delta c_k^t, l_{val,k}^t$

7          $\leftarrow \text{Local\_Update}(k, w^t, E, c^t)$

8         **Intermediate computations**

9          $\Delta^{t+1} \leftarrow \beta_1 \Delta^t + (1 - \beta_1) \sum_{k=1}^K p_k \Delta w_k^t$

10          $v^{t+1} \leftarrow \beta_2 v^t + (1 - \beta_2) (\sum_{k=1}^K p_k \Delta w_k^t)^2$

11          $c^{t+1} \leftarrow c^t + \sum_{k=1}^K p_k \Delta c_k^t$

12          $\Delta_k^t \leftarrow \frac{1}{\eta_l} F_k^q(w^t) \Delta w_k^t$

13          $h_k^t \leftarrow q F_k^{q-1}(w^t) \|\Delta w_k^t\|^2 + \frac{1}{\eta_l} F_k^q(w^t)$

14          $\Delta l_k^t = \max(0, l_{val,k}^{t-1} - l_{val,k}^t)$

15          $m_k^t = \sum_{j=0}^5 l_{val,k}^{t-j}$

16         **Aggregation**

17          $w^{t+1} \leftarrow w^t + \eta_s \frac{\Delta^{t+1}}{\sqrt{v^{t+1} + \tau}}$

18          $w^{t+1} \leftarrow w^t + \frac{\sum_{k=1}^K \Delta_k^t}{\sum_{k=1}^K h_k^t}$

19          $w^{t+1} \leftarrow w^t + \sum_{k=1}^K p_k \Delta w_k^t$

20          $w^{t+1} \leftarrow w^t + \frac{\gamma}{K} \sum_{k=1}^K \Delta w_k^t$

21          $w^{t+1} \leftarrow w^t + \sum_{k=1}^K (\alpha p_k + \beta \frac{\Delta_k^t}{\sum_{k=1}^K \Delta_k^t} + \gamma \frac{m_k^t}{\sum_{k=1}^K m_k^t}) \Delta w_k^t$

22          $w^{t+1} \leftarrow w^t + \sum_{k=1}^K p_k \Delta w_k^t$

23     **return**  $w^T$

#### 20 Local Update (Institution-side)

**Input:**  $k, w^t, E, c^t$

21  $w_k^t \leftarrow w^t$

22 **for**  $e \leftarrow 1$  **to**  $E$  **do**

23     **for**  $b \in \mathcal{B}_k$  **do**

24          $w_k^t \leftarrow w_k^t - \eta_l \nabla l(w_k^t, b) + \eta_l (c^t - c_k^t)$

25      $\Delta c_k^t \leftarrow -c^t + \frac{1}{s_k \eta_l} \Delta w_k^t$

26      $c_k^{t+1} \leftarrow c_k^t + \Delta c_k^t$

27      $\Delta w_k^t \leftarrow w_k^t - w^t$

28      $l_{val,k}^t \leftarrow \text{Local\_Validation}(w_k^t)$

29     **return**  $\Delta w_k^t, \Delta c_k^t, l_{val,k}^t$

aggregated and sent back to the institutions while the weights  $w_{lk}$  are kept completely local and preserved from one local optimization to the other. In *FedPer* the authors proposed to keep the last layers of a model private, while in *LG-FedAvg* the first layers are kept local.

### 3.4. Clustered federated learning

#### 3.4.1. Problem formulation

Depending on the task, assumptions on the existence of clusters of institutions with similar data distributions can be made. Consequently, some authors Ghosh et al. (2020); Sattler et al. (2021) proposed to define a clustered federated learning objective as the union of each cluster objective

$$w_c^* = \underset{w_c \in \mathcal{W}}{\operatorname{argmin}} \sum_{k \in c} \sum_{i=1}^{n_k} l(w_c, x_i^k, y_i^k), \forall c \in \mathcal{C} \quad (25)$$

with  $\mathcal{C} \subset \mathcal{P}(\{1, \dots, K\})$  a partitioning of the  $K$  institutions, and  $w_c$  the model parameters of cluster  $c$ , while each cluster  $c$  benefits from the whole federation.

#### 3.4.2. Prior clustering

If possible, one can leverage prior knowledge on the distribution of the data between institutions, effectively defining clustering before federated training. This leads to  $\mathcal{C}$  being predefined before federated training while implying the leakage of potentially sensitive information. In the case of our study, we do know that gliomas can be classified as of low or high grade (LGG or HGG respectively, c.f. Section 4.1). We name *Prior CFL HGG/LGG* the federated finetuning (using *FedAvg*) of the best obtained *FedAvg* model on two clusters of institutions: the ones only owning volumes displaying a low-grade tumor (12, 13, 14 and 15) from the others.

#### 3.4.3. Sattler's Clustered federated learning (CFL) Sattler et al. (2021)

If no prior knowledge can be shared independently to federated learning, cluster assignments can be performed in a variety of ways during training. In Sattler et al. (2021), the authors proposed to recursively split clusters of institutions into two during federated training. The method starts with standard *FedAvg* on a single cluster  $c_0 = \{1, \dots, K\}$  containing every institution. A list of clusters of institutions  $\mathcal{C}$  is saved during the whole optimization process, initialized with the initial cluster  $c_0$ . At an arbitrary communication round  $t$ , *FedAvg* is performed independently in parallel on each cluster of the list. At each communication round, the server assesses if each cluster  $c$  in the list  $\mathcal{C}$  must be split into two parts  $c_1$  and  $c_2$  before the next round through the following conditions

$$\left\| \frac{1}{|c|} \sum_{i \in c} \Delta w_i^t \right\| < \epsilon_1 \quad (26)$$

$$\max_{i \in c} \|\Delta w_i^t\| > \epsilon_2 \quad (27)$$

If so, the server splits the cluster  $c$  of institutions in two based on the cosine similarity matrix of received local updates from

these institutions

$$a_{i,j}^t = \frac{\langle \Delta w_i^t, \Delta w_j^t \rangle}{\|\Delta w_i^t\| \|\Delta w_j^t\|}, i, j \in c \quad (28)$$

effectively solving

$$c_1, c_2 = \underset{c_1 \cup c_2 = c}{\operatorname{argmin}} \left( \max_{i \in c_1, j \in c_2} a_{i,j} \right) \quad (29)$$

and splitting cluster  $c$  into  $c_1$  and  $c_2$  if

$$\gamma_{max} < \sqrt{\frac{1 - \max_{i \in c_1, j \in c_2} a_{i,j}}{2}} \quad (30)$$

As the hyperparameters  $\epsilon_1, \epsilon_2$  and  $\gamma_{max}$  defined to choose which cluster to split at which global round are very hard to tune, we slightly simplify their method. Instead of using  $\epsilon_1, \epsilon_2$  and  $\gamma_{max}$ , we only define as a hyperparameter the communication rounds at which each cluster must be split in two, which enables us to avoid relying on local updates norms which can be very heterogeneous in our extremely unbalanced setup.

## 4. Experiments

### 4.1. Federated brain tumor segmentation 2022 challenge dataset

Experiments were led on the *MICCAI's Federated Brain Tumor Segmentation 2022 Challenge dataset (FeTS2022)* (Bakas et al. (2017); Pati et al. (2021); Reina et al. (2022)). This dataset is based on the Brain Tumor Segmentation 2021 Challenge dataset (BraTS2021). It consists of 1251 multi-modal brain MRI scans (T1, T1ce, T2 and FLAIR) of size  $240 \times 240 \times 155$  with isotropic  $1\text{mm}^3$  voxel size along with their multi-label tumor segmentation masks including 4 labels, namely background, enhancing tumor (ET), tumor core (TC) and whole tumor (WT).

The real-world partitioning along the 23 acquiring institutions is provided in addition to the samples, enabling the simulation of federated learning. Tumors present on each volume can be of low or high grade (LGG or HGG respectively), defining its level of aggressiveness. We can highlight significant heterogeneity between samples of different institutions in terms of quantity, label distribution and tumor grade. Figure 1 shows the distribution of tumor types for each institution. Institutions 1 and 18 provide a large majority of the samples, accounting for 71% of the total dataset alone while more than 61% of the institutions own less than 15 samples each. This gives an extremely unbalanced setup. Also note that the tumor grade is known for only  $\sim 30\%$  of the patients. We know however that institutions 12, 13, 14 and 15 only own LGG samples, institutions 3 and 4 own samples of both grades, and most other institutions only own HGG samples. Figure 2 shows the distribution of volumes of each label for each institution. The label distribution is not completely homogeneous across institutions, but it is a priori hard to assess if the label shift we see is sufficient to disturb federated training. We can, however, note that samples of institutions 12, 13, 14 and 15 tend to have smaller ET labels, confirmed by the remark on LGG in Bakas et al. (2019) that "they



do not exhibit much contrast enhancement, or edema”. Finally, it must be noted that we do not have access to information about the acquisition protocols of each institution (scanners, parameters, etc.), a feature shift also exists in the dataset.

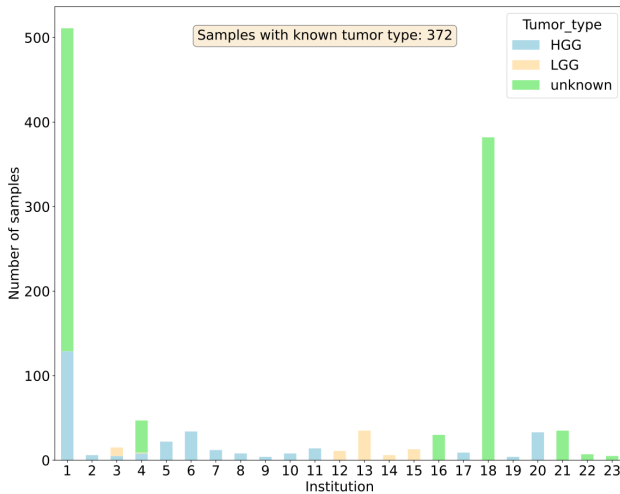


Fig. 1. Number of samples and tumor grade distribution per institution in the challenge setup of FeTS2022 dataset.

## 4.2. Exploring FL methods on FeTS2022 original partitioning

In this first group of experiments, we explore the performance of different training schemes on the full dataset in the challenge setup described in Section 4.1.

### 4.2.1. Performance of centralized and institution 1 training

As non-federated baselines, we trained a model with standard SGD on the pooled dataset, which we call *Centralized* training. Since Institution 1 owns more than 500 volumes, it is by itself capable of training a decent model. We thus train a model on its samples alone, which we name *local institution 1* training, to assess in what manner centralized training and federated methods improve this independent model.

### 4.2.2. Performance of FedAvg variations

As detailed in section 3.2.2, there remains ambiguity about the amount of computation each institution must perform between each communication round in the literature. While theoretical papers such as Karimireddy *et al.* (2020); Li *et al.* (2020d); Wang and Ji (2022) mostly assume a fixed number of gradient steps per institution, in practice most developed methods such as McMahan *et al.* (2017); Reddi *et al.* (2022) implement a fixed number of epochs per institution as described in Algorithm 1. In a context of unbalanced data quantity across institutions and fixed batch size, the notion of an epoch is ill-defined, with institutions owning more samples performing more gradient steps than others, potentially giving larger updates. Moreover, in their pioneering work, McMahan *et al.* (2017) introduces both uniform and weighted server-side aggregations, the latter being introduced to ensure convergence toward the correct optima of Equation 2. We first explore these ideas experimentally to highlight the best way of distributing

computation across institutions with standard *FedAvg* in the challenge setup.

### 4.2.3. Performance of global, personalized and hybrid solutions

We then explore the performance of SOTA federated learning methods described in section 3 compared to the baselines of *local training*, *centralized training* and *FedAvg*.

## 4.3. Performance on alternative data partitionings

Since label and feature shifts and the unbalanced data quantity distribution of the FeTS2022 dataset can alter the behaviour of federated learning algorithms, we complement the proposed benchmark with the application of federated learning methods on two different synthetic partitionings.

### 4.3.1. IID setup

To understand the effect of the heterogeneous data distribution of the original partitioning, we explore the performance of *FedAvg* while redistributing every sample from every institution randomly in an i.i.d and balanced fashion to 23, 10 and 5 synthetic institutions. We call this *the IID setup*.

### 4.3.2. Limited data setup

It took close to 10 years for the organizers and data curators of BraTS datasets to accumulate such an amount of quality data. While this specific case is of great interest, one can argue that the data distribution, with institutions 1 and 18 owning a very large amount of samples compared to the others, might not be representative of most federated learning applications in a medical cross-silo environment. Moreover, we do not have information on the tumor grade for most samples added to the last version of the dataset, as shown in Figure 1, which limits our ability to study the potential impact of label shift. For these reasons, we propose first to remove every sample for which we do not have the latter information. This effectively removes institutions 16, 18, 21, 22 and 23, reducing the number of institutions to 18. While doing so, institution 1 would still own a large number of samples compared to others. We thus randomly remove samples from this institution such that it owns only 35 samples, matching the number of samples of the three other main institutions. The final data distribution in this setup is presented in Figure 3, named the *limited setup*. LGG tumors are now less marginal representing around 30% of the dataset, and data is spread in a more balanced fashion across institutions with overall a limited amount of samples per institution. As experimenting with every implemented method would be too expensive, we chose to limit performance analysis on this setup to classical baselines as well as the best-performing methods in section 4.2.3, namely the global SCAFFOLD method and the pluralist HGG/LGG CFL method.

## 4.4. Experimental setup

### 4.4.1. Model and loss function

As the backbone model, we chose a classical 3D-UNet Ronneberger *et al.* (2015) using Monai’s DynUnet implementation (Cardoso *et al.* (2022); Futrega *et al.* (2021)) with kernel sizes

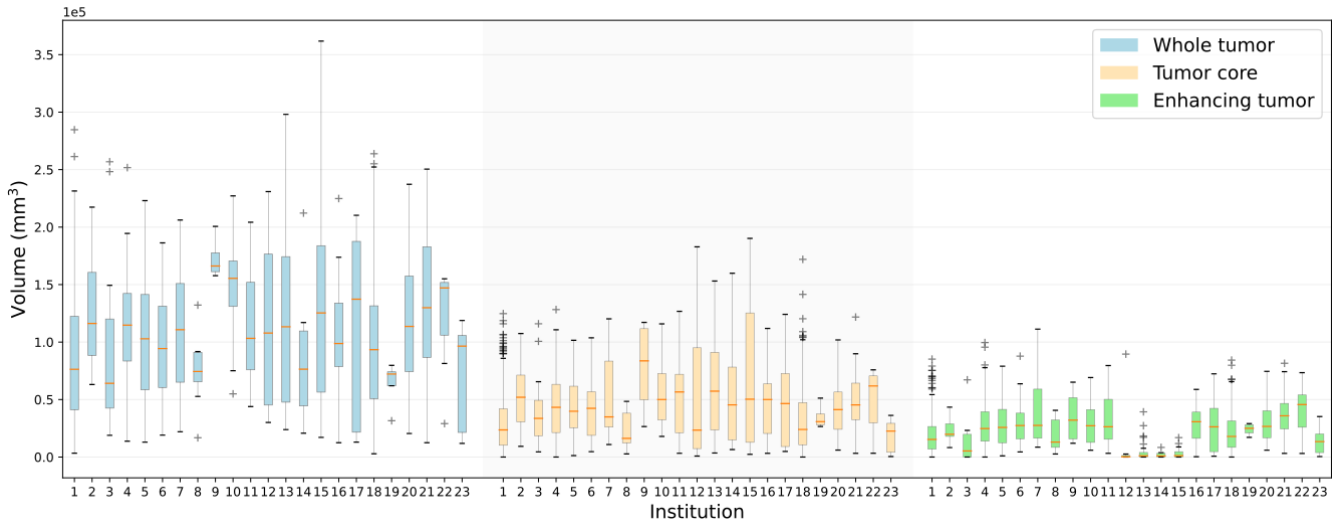


Fig. 2. Label's volume distribution per institution in FeTS2022 dataset.

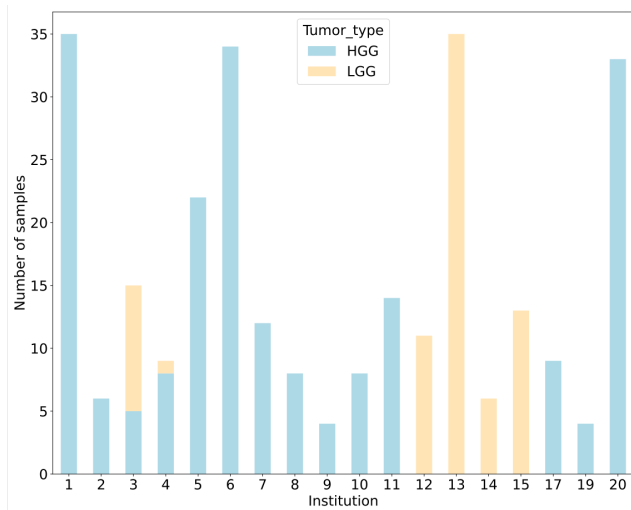


Fig. 3. Number of samples and tumor grade distribution per institution in the limited setup of FeTS2022 dataset. The complete federated subset is composed of 278 patients.

$3 \times 3 \times 3$ , and 4 downsampling paths starting from 32 channels after two input convolutional layers up to 512 channels in the bottleneck of the network, doubling the number of channels after each downsampling layer. The network is composed of around 22.5M parameters. Following the architecture choices in previous versions of BraTS challenge Futrega et al. (2021) and nnUNet Isensee et al. (2021), we use instance normalization layers without scaling parameters. This is even more motivated by the fact that batch normalization can be intrinsically problematic in a federated setup both from performance Li et al. (2021c) and privacy Hatamizadeh et al. (2023) points of view. We use LeakyReLU activations with a negative slope of 0.01 in the whole network. The optimized loss function is a 3D soft Dice loss defined as follows

$$1 - \frac{2 * \sum_i^P p_i g_i + \epsilon}{\sum_i^P p_i + \sum_i^P g_i + \epsilon} \quad (31)$$

where  $p_i \in [0, 1]$  and  $g_i \in \{0, 1\}$  are respectively the prediction and ground truth for voxel  $i$ , and  $\epsilon = 1$  is a smoothing parameter enabling to deal with empty ground truths for TC and ET in some scans. Optimization is performed using patches of size  $128^3$  on the task labels (ET, TC and WT) directly. We say that training has been performed for one epoch when one patch per patient has been processed. Inference on a whole multi-modal volume is performed by sliding window inference method with an overlap of 0.5 between patches in gaussian mode.

#### 4.4.2. Data pre and post-processing

To limit the computational cost of training and inference, we performed, preliminary to any patching step, a per-volume foreground crop based on the non-zero image masks, followed by a zero-padding to a minimum size of 128 on each dimension. Then, to deal with the heterogeneity of image textures induced by acquisitions on different vendor machines or with different acquisition parameters from the different institutions, each modality of each volume was independently normalized with z-score standardization. Whenever a patient is selected in a training batch, a patch of size  $128^3$  is randomly extracted from its associated cropped volume.

We avoided using post-processing steps of the output segmentation maps as in nnUNet Isensee et al. (2021). They would hide imperfections of segmentations from one training algorithm to another, potentially invalidating our conclusions.

#### 4.4.3. Data augmentation

We further apply a variety of data augmentations on the fly during training. Precisely, the following augmentations are applied on an extracted patch, in order:

- *Random flip* on each axis with probability 0.5,
- *Additive Gaussian noise* with a standard deviation of 0.25 with probability 0.15,
- *Gaussian smooth* with a random standard deviation sampled in  $[0.5, 1]$  for each dimension with probability 0.15,

- *Intensity scaling* with factor randomly sampled in [0.7, 1.3] with probability 0.15,
- *Random contrast adjustment* with gamma randomly sampled in [0.5, 2] with probability 0.3.

Note that we do not use spatial augmentations such as random zooms and rotations as in nnUNet Isensee *et al.* (2021). Preliminary experiments tended to show no benefit in performance when using them, while significantly slowing down training as they are relatively expensive augmentations requiring interpolation.

#### 4.4.4. Federated cross-validation

We adapted the paradigm of 5-fold cross-validation to the federated setup to strengthen the statistical significance of the experimental results. This is essential as most institutions do not hold more than 40 volumes. To preserve the representativity of each institution in each decentralized fold, we chose to compute 5 folds on the local dataset of each institution such that a complete decentralized fold is composed of the union of one local fold per institution. Each complete fold serves once as a testing fold. The remaining four other folds are partitioned into a training and validation set following the same principle of locally splitting data into training and validation sets with a proportion of 80-20% for each institution such that the complete decentralized training and validation sets are the union of each local corresponding set. We use the validation sets for hyperparameter optimization and best model selection during training, such that the testing folds effectively give a glimpse of the generalization performances.

#### 4.4.5. Evaluation metrics

Following the choice made by BraTS challenges' organizers, we use a combination of Dice score and Hausdorff distance to evaluate our models. Considering a prediction  $P \in \{0, 1\}^{l \times h \times w \times d}$  and the ground truth  $G \in \{0, 1\}^{l \times h \times w \times d}$ , the Dice score measures the overlap between them

$$DICE(P, G) = \frac{2 * |P \cap G|}{|P| + |G|} \quad (32)$$

while the 95% Hausdorff distance serves as a shape similarity metric, penalizing a significant amount of outlier voxels in the prediction

$$95HD(P, G) = \mathcal{P}_{95}(\{\|p, G\|_2\}_{p \in P} \cup \{\|g, P\|_2\}_{g \in G}) \quad (33)$$

with  $\mathcal{P}_{95}$  symbolizing the 95th-centile. In the case that a label is not present in a volume, occurring for the ET and TC labels of some samples, the Dice score on this label becomes binary while the Hausdorff distance is undefined. In this case, we only include the Dice score in the average metrics, not the Hausdorff distance. For each metric, we record the average and standard deviation over all test cases. Aggregating institution-averaged metrics would give a similar weight in the evaluation to any two institutions which own a different number of samples, giving a fairness-oriented evaluation in the sense of Li *et al.* (2020b). We chose to aggregate metrics on a per-sample basis corresponding

to the classical centralized evaluation and the actual federated optimization objective, while exploring the fairness aspects by also analyzing performances per institution.

#### 4.4.6. Evaluation on BraTS2021 Validation dataset

To ensure the competitiveness of the chosen network architecture and training scheme, we also validate the models trained in a centralized fashion on BraTS2021 validation dataset through the continuous evaluation portal provided by the organizers of the challenges. The ensemble of the five centralized models, each trained on a different set of folds, by a classical logit average is evaluated this way.

#### 4.4.7. Benchmark budget and fairness

Guaranteeing the fairness of the benchmark is non-trivial as we allow for different computational and communication loads between methods without necessarily leading every algorithm to convergence due to our computational limitations. We recorded, for each method, their communication cost as well as the amount of total and parallel SGD steps performed during training respectively defined as

$$S_{total\_SGD\_steps} = R \sum_{k=1}^K s_k \quad (34)$$

$$S_{total\_parallel\_SGD\_steps} = R * \max_{k \in \{1, \dots, K\}} s_k \quad (35)$$

with  $s_k$  the amount of steps performed by institution  $k$  during one round and  $R$  the total number of rounds. We also computed a rough estimate of ideal simulated federated training time, inspired by the time simulation of FeTS challenges. We considered the total time taken by a method to be

$$T_{total} = R * \max_{k \in [K]} (T_k^{computation} + T_k^{communication}) \quad (36)$$

The computation time of an institution  $k$  for a round is composed of its training steps and its validation of the global model

$$T_k^{computation} = s_k * t_k^{batch} + n_k^{eval} * t_k^{eval} \quad (37)$$

with  $n_k^{eval}$  the number of validation patients, and  $t_k^{batch}$  and  $t_k^{eval}$  the training time for one batch and validation time for one patient. Its communication time is composed of its download and upload time

$$T_k^{communication} = S_{model} * t_k^{download} + S_{model} * t_k^{upload} \quad (38)$$

with  $S_{model}$  the size of the model, and  $t_k^{download}$  and  $t_k^{upload}$  the net download and upload speeds.

We estimated  $t_k^{batch}$  and  $t_k^{eval}$  to be 1.86s/batch and 0.80s/patient through multiple centralized epochs using a single V100 GPU giving a realistic computation time of an institution with an extremely high quality system, and  $t_k^{download}$  and  $t_k^{upload}$  to be 20MB/s and 13.3MB/s based on the fastest communicating institution in FeTS challenge<sup>1</sup>. The time simulation

<sup>1</sup>Communication time measures were stored in the source code of FeTS2022 challenge by the organizers.

and these estimated values are only used as a reference for idealistic conditions of federated learning for a relative fairness of the benchmark, and are discussed in the Discussion section 6. With these assumptions, we trained the models for around 20h of simulated time with each algorithm. It corresponds to approximately 300 rounds of FedAvg, which was enough for its fastest variant to be close to convergence. The hyperparameters of all other algorithms were set accordingly as described in the next section.

#### 4.4.8. Hyperparameters

*Centralized* and *Institution 1 local* models were trained for 300 epochs. This number was shown sufficient for the centralized model to be close to convergence i.e. at the beginning of the validation plateau from where it would require an order of magnitude more epochs to potentially improve performance in a slightly significant way. For every federated method, we fixed the number of local epochs to 1. For *FedAvg w/ fixed local iterations*, we fixed the number of local iterations to 10 and trained for 720 communication rounds. Other global solutions as well as *FedPer*, *LG-FedAvg* and *CFL* were trained for 300 communication rounds. It was enough for the fastest converging methods to reach its validation plateau. *Local finetuning* and *Ditto* were trained locally for 30 epochs, starting from the best *FedAvg* model. In *CFL*, the federation was split in two at communication round 200. Both clusters of *HGG/LGG CFL* were trained for 30 communication rounds starting from the best *FedAvg* model after 270 rounds.

Optimizations were performed using classical SGD with a batch size of 4 and a weight decay of  $10^{-5}$  (the latter was shown to be beneficial in Wang et al. (2020a) and Yuan et al. (2022a)). Hyperparameter optimization was performed by grid search using the train-validation split of the first four folds only (not on every fold), apart from *Local finetuning* and *Ditto*, to limit the computational cost of the benchmark. *q-FedAvg* is the only method for which the hyperparameters selection was manual and not based on the best validation performance. As a fairness-oriented method, each value of  $q > 0$  decreased the overall performance, we arbitrarily chose one to illustrate its effect. Apart from *q-FedAvg*, an exponential learning rate decay was used multiplying the local learning rate (not the server one if applicable) by 0.995 after each epoch or communication round, roughly dividing by 5 the initial learning rate after 300 epochs. Search spaces and selected values of each hyperparameter for each method with each federated partitioning can be found in Table A.4 of Appendix A.

#### 4.4.9. Implementation

All methods were implemented in Python using Pytorch Paszke et al. (2019) and Monai Cardoso et al. (2022). Training and testing were simulated using 2 NVIDIA V100 32Go GPUs. Training a centralized model for 300 epochs took around 30 hours in this multi-GPU setup.

## 5. Results

As results for a large number of methods are presented throughout this section, all-vs-all hypothesis testing would re-

quire too much correction for it to be beneficial. We instead chose to perform a non-parametric one-tailed Wilcoxon signed-rank test to support each conclusion taken from the results. We provide in parenthesis the p-value associated with each test.

### 5.1. Centralized performance on the BraTS2021 validation data

We first provide the results of the evaluation of the ensemble of five centralized models through the continuous evaluation portal of BraTS2021. This ensemble of models obtained average Dice scores of 0.920, 0.861 and 0.831 and 95% Hausdorff distances (in mm) of 4.744, 9.835 and 18.048 for Whole Tumor, Tumor Core and Enhancing Tumor labels respectively. While each model was trained on around 15% fewer samples than a classically trained model in the challenge setup due to the additional validation sets isolated from training sets, evaluated performances remain close to what could be obtained by the ensemble of larger nnU-Nets which was part of the 2022 challenge's test winner submission Zeineldin et al. (2022) and achieved 0.9213, 0.8718 and 0.8402 Dice scores and 3.82, 8.95 and 16.03 95% Hausdorff distances. We thus emphasize that the proposed benchmark is not performed using a small model, and most exposed results should be translatable to the practical implementation of federated learning on this task.

### 5.2. Exploring FL methods on FeTS2022 original partitioning

We focus in this section on the performance achieved by the selected SOTA federated methods presented in Section 3 on the original data partitioning (23 institutions) of FeTS2022.

#### 5.2.1. Performance of baselines and FedAvg variations

We show in Table 1 dice scores and 95% Hausdorff distances obtained by different variations of *FedAvg* compared to the performance of *Local institution 1* and *Centralized* training, while Figure 4 shows the average dice scores per institution. First, institution 1 is by itself capable of training a very decent model which however does not generalize well to the dataset of other institutions. Furthermore, centralizing the volumes of every institution does bring a significant improvement ( $p \ll 10^{-10}$ ) of the segmentation accuracy (Dice and Hausdorff) on all of them, with still a slight improvement for samples of institution 1. We see in Figure 4 however that the centralized model does not perform as well on institutions 3, 12, 13, 14 and 15 as on others.

*FedAvg* with weighted aggregation and a fixed amount of local epochs per round enables training of a model with very close performances to the *centralized* one, with still a significant gap ( $p \ll 10^{-10}$ ) of one dice point and a slight edge in 95% Hausdorff distance. Performances on institutions 3, 12, 13, 14, and 15 are however significantly worse than in the centralized case (Figure 4) ( $p = 0.04, 5 \times 10^{-4}, 5 \times 10^{-6}, 0.015$  and  $0.02$  respectively). *FedAvg w/ fixed local epochs* does focus on the performance of large and representative institutions, while institutions that appear to own data further away from the average distribution are clearly penalized compared to centralized training.

Compared to the previous variant of *FedAvg*, using a uniform averaging or a fixed amount of iterations per communication round slows down the convergence, with significantly

lower average performance ( $p \ll 10^{-10}$ ) both in terms of dice score and Hausdorff distance. The performance loss compared to *centralized* on institution 13 is however less important. This recovers theory since using both a weighted aggregation and one local epoch per round “doubly” increases the importance to institutions owning a large number of samples, orienting the optimization toward their local objective Wang et al. (2020b). This shows however that a compromise must be made: giving this much weight to institutions 1 and 18 does enable them to “drive” the federation faster toward an efficient global model, at the cost of limiting its performance on institutions owning atypical samples.

We use in the remainder of this article *FedAvg w/ fixed local epochs* as the baseline for the comparison to other SOTA methods.

### 5.2.2. Global solutions

We show in Table 1 dice scores and 95% Hausdorff distances obtained by SOTA global solutions compared to the baselines of *Institution 1 local*, *Centralized* and *FedAvg* training, while Figure 5 shows the average dice scores per institution.

First, *FedAdam* does not improve the convergence of training over standard *FedAvg*. To limit the computational cost and ensure the fairness of comparison to other methods, we limited the size of the grid search over the hyperparameters of this method, with  $\beta_1, \beta_2, \tau$  fixed following the recommendations of Reddi et al. (2022) and two learning rates remaining to tune (c.f. Table A.4). Better performances and convergence might be obtainable with a better choice of hyperparameters, but it is simply unrealistic in a federated learning context.

*SCAFFOLD* on the other hand does provide a significant gain over *FedAvg* in dice scores ( $p = 0.03$ ), while remaining close in Hausdorff distance. We show in Figure 5 that the performance gain can be perceived for most institutions, and in particular for 3, 12, 13, 14 and 15 ( $p = 0.02, 0.09, 8 \times 10^{-6}, 0.016$  and  $0.07$ ). In other words, the optimization and performance distribution experienced when using *SCAFFOLD* is much closer to *centralized* training than with *FedAvg*, at the cost of twice larger communications. We can only recommend using this optimizer for this task if the network infrastructure enables it.

In the same fashion as *SCAFFOLD*, *FedNova* tends to improve the average dice scores for institutions 12, 13, 14 and 15, but significantly reduces the average dice score across all institutions ( $p \ll 10^{-10}$ ).

*q-FedAvg* does provide a fairer dice score distribution, while however bringing the performance of every other institution significantly down ( $p \ll 10^{-10}$ ).

*FedPIDAvg* gives similar performance to *FedAvg*, while giving a smoother validation loss. The weighting dynamics of this method are very interesting for this task. The large institutions are given higher weights in the first rounds of training as they bring their local validation loss down way faster than others, promoting faster initial convergence. In the later rounds of training, more weight is given to other institutions which continue improving their loss, which justifies a better performance on institutions 12, 13, 14 and 15 than standard *FedAvg*. How-

ever, similarly to *FedAdam*, it might suffer from fixed  $\alpha, \beta$  and  $\gamma$  hyperparameters following Mächler et al. (2023), as well as the randomness of local validations with a very small amount of samples.

### 5.2.3. Personalized and hybrid solutions

We show in Table 1 dice scores and 95% Hausdorff distances obtained by SOTA personalized and hybrid solutions compared to the baselines of *Institution 1 local*, *Centralized* and *FedAvg* training, while Figure 6 shows the average dice scores per institution.

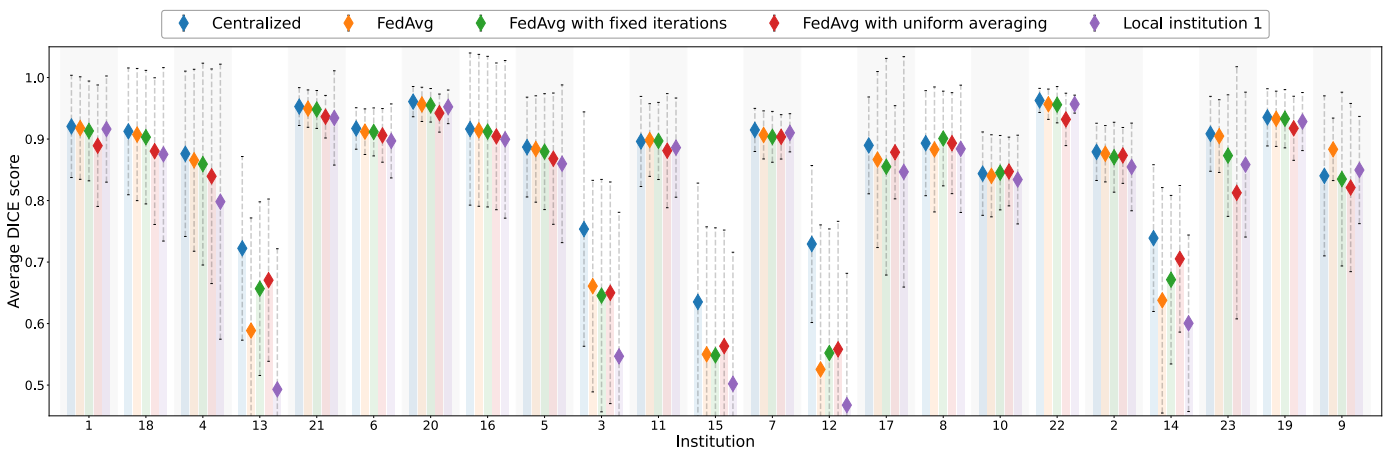
We show first that *local finetuning* does bring a significant performance improvement in dice score ( $p = 10^{-8}$ ) while slightly increasing the average Hausdorff distance. We can note that overall, segmentation performance of the whole tumor (WT) is made a little bit more irregular with a slightly larger Hausdorff distance, but a clear improvement is perceived on the tumor core label ( $p = 3 \times 10^{-4}$ ). Institution-wise, we can see that *local finetuning* does improve the performance for every institution owning a large number of samples, especially for institution 13 for which it enables to limit the performance gap between *centralized* training and *FedAvg* ( $p = 0.0013$ ). Benefits become unclear for smaller institutions: finetuning a model with more than 22M parameters on only a few samples can definitely lead to overfitting. Even when selecting the best model based on validation performance, *local finetuning* is not adapted to very small institutions. The regularization in *Ditto* does not enable solving the problems of *local finetuning*, giving very similar performances.

The explored partial model-sharing methods clearly decrease the performance of the model for most institutions ( $p = 0.78$  and  $10^{-11}$  for *FedPer* and *LG-FedAvg* respectively when including all samples, and  $p = 10^{-5}$  and  $10^{-17}$  when taking out samples from institutions 1 and 18 from the tests). They are also quite unadapted to the small number of samples of most institutions. The fewer samples an institution has the more the performance decreases compared to *FedAvg*, which is coherent with the fact that they have to be able to train from scratch several layers of the network. We can note an interesting difference between *FedPer* and *LG-FedAvg*. The latter gives decent performance only on institutions 1 and 18 owning a large amount of data, equivalent to the local model they can obtain alone, with drastically decreased performances for institutions owning 40 samples or fewer. The former still gives decent performances for institutions 5, 6, 16, 20 and 21 which owns between 22 and 35 samples each. **In that sense, the federation of the first layers of the model seems to be very important for better performance**, or they at least require to be trained on a larger amount of samples than the last layers to be meaningful.

This section’s results motivate the study of clustering-based personalization methods for this task, potentially addressing issues caused by small local sample sizes. We show however that the pure gradient similarity-based clustered method *CFL* does not provide a significant performance gain over other methods, even compared to classical *FedAvg*. While we remarked that institutions 12, 13 and 14 were always clustered together, the computed clusters vary considerably from one fold to another.

**Table 1. Dice scores and 95% Hausdorff distances (mm) obtained by local, centralized and SOTA federated learning algorithms on FeTS2022 (average  $\pm$  std over samples of all folds)**

Algorithm	Mean DICE	DICE WT	DICE TC	DICE ET	Mean 95HD
<b>Centralized</b>	<b>0.903 <math>\pm</math> 0.109</b>	<b>0.929 <math>\pm</math> 0.072</b>	<b>0.907 <math>\pm</math> 0.154</b>	<b>0.873 <math>\pm</math> 0.177</b>	<b>4.565 <math>\pm</math> 6.561</b>
Local training institution 1	0.871 $\pm$ 0.159	0.903 $\pm$ 0.112	0.868 $\pm$ 0.221	0.841 $\pm$ 0.210	7.342 $\pm$ 11.403
FedAvg variants					
FedAvg w/ fixed local epochs	0.891 $\pm$ 0.130	0.924 $\pm$ 0.079	0.892 $\pm$ 0.187	0.858 $\pm$ 0.195	4.742 $\pm$ 6.270
FedAvg w/ fixed local iterations	0.889 $\pm$ 0.124	0.921 $\pm$ 0.081	0.895 $\pm$ 0.167	0.852 $\pm$ 0.196	5.054 $\pm$ 6.949
FedAvg w/ uniform averaging	0.871 $\pm$ 0.128	0.906 $\pm$ 0.087	0.876 $\pm$ 0.173	0.831 $\pm$ 0.203	6.541 $\pm$ 9.306
Global federated methods					
FedNova	0.877 $\pm$ 0.124	0.910 $\pm$ 0.083	0.883 $\pm$ 0.172	0.839 $\pm$ 0.195	5.824 $\pm$ 7.532
FedAdam	0.886 $\pm$ 0.135	0.922 $\pm$ 0.082	0.888 $\pm$ 0.180	0.847 $\pm$ 0.209	6.026 $\pm$ 10.919
<i>SCAFFOLD</i>	<i>0.897 <math>\pm</math> 0.113</i>	<i>0.925 <math>\pm</math> 0.079</i>	<i>0.904 <math>\pm</math> 0.154</i>	<i>0.863 <math>\pm</math> 0.187</i>	<i>4.839 <math>\pm</math> 7.140</i>
q-FedAvg q=1.0	0.803 $\pm$ 0.165	0.861 $\pm$ 0.137	0.781 $\pm$ 0.223	0.768 $\pm$ 0.225	12.994 $\pm$ 18.027
FedPIDAvg	0.891 $\pm$ 0.122	0.921 $\pm$ 0.081	0.896 $\pm$ 0.172	0.857 $\pm$ 0.192	5.163 $\pm$ 7.141
Personalized and clustered federated methods					
Local finetuning	0.895 $\pm$ 0.120	0.924 $\pm$ 0.079	0.900 $\pm$ 0.164	0.861 $\pm$ 0.194	5.114 $\pm$ 7.233
Ditto	0.894 $\pm$ 0.123	0.924 $\pm$ 0.082	0.900 $\pm$ 0.164	0.859 $\pm$ 0.197	5.123 $\pm$ 7.379
FedPer	0.886 $\pm$ 0.135	0.922 $\pm$ 0.085	0.894 $\pm$ 0.173	0.842 $\pm$ 0.224	6.289 $\pm$ 11.335
LG-FedAvg	0.881 $\pm$ 0.138	0.914 $\pm$ 0.094	0.887 $\pm$ 0.176	0.841 $\pm$ 0.214	6.987 $\pm$ 11.776
CFL 2 clusters	0.891 $\pm$ 0.129	0.923 $\pm$ 0.081	0.890 $\pm$ 0.187	0.859 $\pm$ 0.194	5.114 $\pm$ 7.780
<i>Prior CFL HGG/LGG</i>	<i>0.896 <math>\pm</math> 0.115</i>	<i>0.925 <math>\pm</math> 0.080</i>	<i>0.901 <math>\pm</math> 0.162</i>	<i>0.863 <math>\pm</math> 0.185</i>	<i>4.753 <math>\pm</math> 6.568</i>
IID setup					
FedAvg IID 23 clients 300 rounds	0.878 $\pm$ 0.126	0.912 $\pm$ 0.087	0.885 $\pm$ 0.165	0.837 $\pm$ 0.203	6.120 $\pm$ 8.852
FedAvg IID 10 clients 300 rounds	0.888 $\pm$ 0.122	0.920 $\pm$ 0.079	0.895 $\pm$ 0.164	0.849 $\pm$ 0.198	5.372 $\pm$ 7.853
FedAvg IID 5 clients 300 rounds	0.897 $\pm$ 0.111	0.925 $\pm$ 0.075	0.903 $\pm$ 0.158	0.864 $\pm$ 0.183	4.700 $\pm$ 6.713

**Fig. 4. Average Dice scores for local, centralized and FedAvg variations training in the challenge setup of FeTS2022 dataset per institution, in decreasing order of number of samples (c.f. Figure 1). Institutions 17 to 9 (on the right) each own from 10 to 4 samples in total. Errors bars represent  $\pm$  one standard deviation. Centralized training (blue) is the upper baseline. Drops in performance are perceived for every variants of FedAvg compared to Centralized, with significantly larger ones for institutions 12, 13, 14, 15 and 3. FedAvg with fixed epochs (orange) converges faster than other variants to a close match with Centralized performances at the cost of larger gaps for previously cited institutions.**

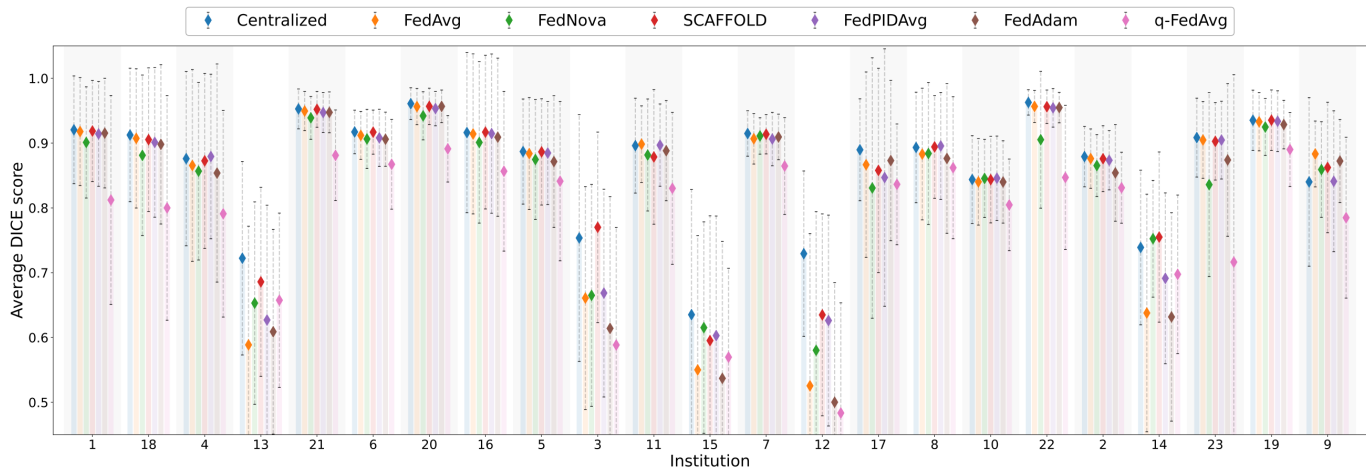
We could not find a concrete link between institutions clustered together for each fold. This method seems to suffer from patch training and the extremely unbalance setup. We show however that applying FedAvg to clusters of institutions that are defined on prior knowledge of the tumor grade does bring a significant performance gain on institutions 12 to 15 which were isolated together ( $p = 10^{-5}$ ). This however requires the communication of potentially sensitive information on the data distribution of each institution which may be acceptable or not. Overall,

this clearly motivates the study of hybrid personalization in the context of a low amount of samples for most institutions.

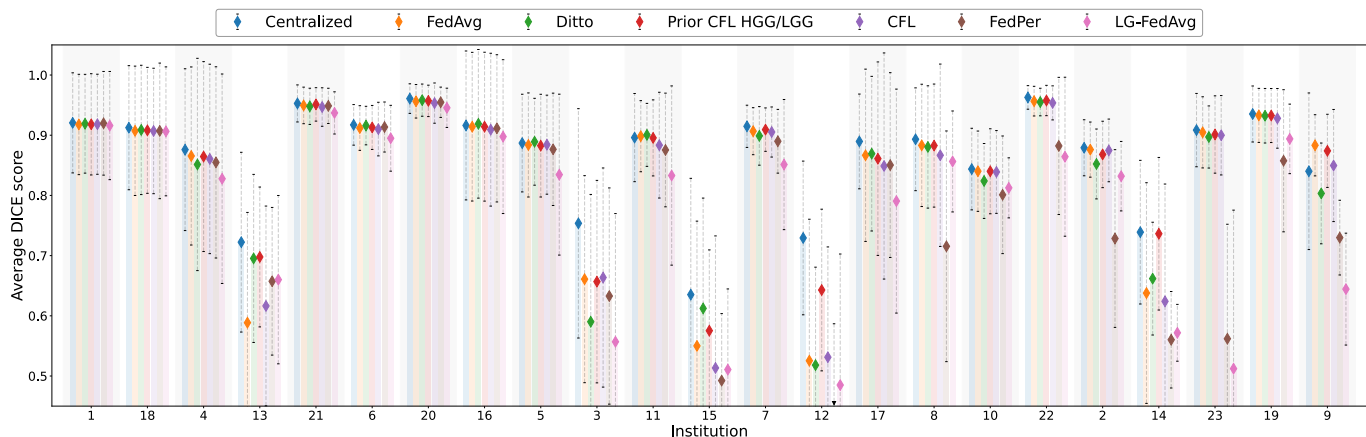
### 5.3. Cost estimations on FeTS2022 original partitioning

We show in Table 2 cost estimations of the different algorithms applied on the complete FeTS2022 federated dataset.

Comparing FedAvg w/ fixed epochs and w/ fixed iterations is highly dependent on the hypotheses on the cost of computations



**Fig. 5.** Average Dice scores of *centralized* and state-of-the-art global federated training in the challenge setup of FeTS2022 dataset per institution, in decreasing order of number of samples (c.f. Figure 1). Institutions 17 to 9 each own from 10 to 4 total samples. Errors bars represent  $\pm$  one standard deviation. Only *SCAFFOLD* (red) improves upon the federated baseline *FedAvg* (orange) by mitigating the drop in performance for institutions 12, 13, 14, 15 and 3 while maintaining it on others. *q-FedAvg* (pink) and *FedNova* (green) homogenizes DICE scores across institutions but decrease the overall performance.



**Fig. 6.** Average Dice scores of *centralized* and state-of-the-art personalized and hybrid federated training in the challenge setup of FeTS2022 dataset per institution, in decreasing order of number of samples (c.f. Figure 1). Institutions 17 to 9 each own from 10 to 4 total samples. Errors bars represent  $\pm$  one standard deviation. *Local finetuning* was removed from this graph as it gives very similar DICE distribution as *Ditto*. These two methods can improve the performance on large enough institutions like 13. The performance of partial model-sharing methods *LG-FedAvg* and *FedPer* collapses for very small institutions. *Prior CFL HGG/LGG* mitigates the drop in performance of *FedAvg* on institutions 12 to 15.

and communications. Our choice of choosing a realistic lower bound on the communication costs while neglecting stragglers effect clearly advantages *FedAvg w/ fixed iterations*, but this algorithm still slightly, but significantly, underperforms compared to *FedAvg w/ fixed epochs*, although being less biased. It also consumes much more energy since it requires around three times more total SGD steps than *FedAvg w/ fixed epochs*. *SCAFFOLD* does necessitate two times more communications than *FedAvg*. This increase appears relatively negligible in the simulated training time. We allowed for slightly more training time for *Local finetuning* and *Ditto*, as finetuning-based methods are generally used after global training ended. These additional local steps do not require synchronization between institutions anymore, with the main improvement being brought by these methods generally in the first one to five local epochs before overfitting.

#### 5.4. IID setup

We show in Table 1 dice scores and 95% Hausdorff distances obtained by applying *FedAvg* in the IID setup described in Section 4.3.1 compared to the baselines of *Institution 1 local*, *Centralized* and *FedAvg* training, while Figure 7 shows the average dice scores per institution. Validation and test splits were preserved for every fold, only the training samples were redistributed such that performances between the original and IID setup remain comparable.

Interestingly, we show that applying *FedAvg* after redistributing the training samples in an i.i.d and balanced fashion to 23 synthetic institutions leads to a significant decrease in performance compared to the original data distribution ( $p \ll 10^{-10}$ ), both in terms of dice score and Hausdorff distance. In that sense, this again motivates the fact that institutions 1 and 18 owning the vast majority of samples in the original FeTS2022

**Table 2. Cost evaluation of each algorithm of the benchmark in the challenge setup.**

Algorithm	Total SGD steps	Total maximum parallel SGD steps	Total communications (number of floats $\times 10^9$ )	Estimated training time (h)
<i>Local institution 1</i>	24600	24600	0	18.2
<i>Centralized</i>	59400	59400	0	44.6
FedAvg variants				
<i>FedAvg w/ fixed local epochs</i>	59400	24600	13.5	19.1
<i>FedAvg w/ fixed local iterations</i>	165600	7200	32.5	19.1
Global federated methods				
<i>FedNova</i> <i>FedAdam</i> <i>FedPIDAvg</i> <i>q-FedAvg</i>	59400	24600	13.5	19.1
<i>SCAFFOLD</i>	59400	24600	27.0	19.9
Personalized and clustered federated methods				
<i>Local finetuning</i> <i>Ditto</i>	65340	27060	13.5	20.9
<i>FedPer</i>	59400	24600	13.5	19.1
<i>LG-FedAvg</i>	59400	24600	13.4	19.1
<i>CFL</i> <i>Prior CFL HGG/LGG</i>	59400	24600	13.5	19.1

setup drives the federated training toward a decent global model, even in the presence of statistical heterogeneity between local data distributions. Performances on institutions 12 to 15 are on the contrary slightly larger than *FedAvg*, since in the IID setup *FedAvg* does not prioritize the performance of samples owned by institutions 1 and 18.

Performance increases when we decrease the number of synthetic institutions, with five synthetic institutions being small enough to enable *FedAvg* in the IID setup to outperform *FedAvg* in the original setup on average, without statistically significant differences between medians ( $p = 0.53$ ). In this configuration, each of the five synthetic institutions owns around 160 samples, much less than institutions 1 and 18 in the original setup. While institutions 1 and 18 perform a larger number of updates, enabling them to drive *FedAvg* training in the challenge setup, a lower amount per institution is required in the IID setup to match its performance. We hypothesize that this convergence speed gap can be caused by the statistical heterogeneity in the original distribution.

### 5.5. Limited data setup

We show in Table 3 dice scores and 95% Hausdorff distances obtained by applying *FedAvg* in the limited data setup described in Section 4.3.2 compared to the baselines of *Institution 1 local*, *Centralized* and *FedAvg* training, with average Dice scores per institutions in Figure B.10 of Appendix Appendix B.

The performance is overall much worse than in the previous cases, due to the drastic reduction of the total sample size. We recover in this setup the same difference between *FedAvg* and *Centralized* training, with a relatively small but significant gap ( $p = 10^{-10}$ ). *SCAFFOLD* in this setup however does not improve the performance over *FedAvg*, it is actually the opposite

in median dice score ( $p = 0.002$ ). *HGG/LGG prior CFL* does not either significantly improve the segmentation performance over *FedAvg* in this setup ( $p = 0.13$ ), although a slight gain might be perceivable in Tumor Core. The overall amount of data seems too low for clustered finetuning to be useful. The performance on institutions 12, 13, 14, 15 and 3 remains worse than others for every optimizer, including *Centralized* training. Their samples tend to be harder to be segmented by a U-Net, with any optimization seeming to prioritize the performance of the predominant HGG samples.

## 6. Discussion

### 6.1. Results synthesis and takeaways

We proposed an extensive benchmark of SOTA federated methods on different data partitionings of the FeTS2022 dataset. In the challenge setup, standard *FedAvg* with a fixed amount of local epochs and weighted averaging performs very closely to *Centralized* training. The difference in performance between variants of *FedAvg* is clear, we think that the mismatch between practice and theory could motivate the analysis of local epochs-based federated optimizers theoretically. From all the global methods experimented with, *SCAFFOLD* is the only one significantly improving the average segmentation performance compared to baselines. It seems like this federated optimizer behaves much more like *Centralized* training than *FedAvg*. This however does not hold in the limited setup. While personalization methods were hampered by the overall low amount of samples per institution, *local finetuning* and *Ditto* could still provide small benefits for large enough institutions. Considering the data distribution of the dataset, with LGG and HGG samples



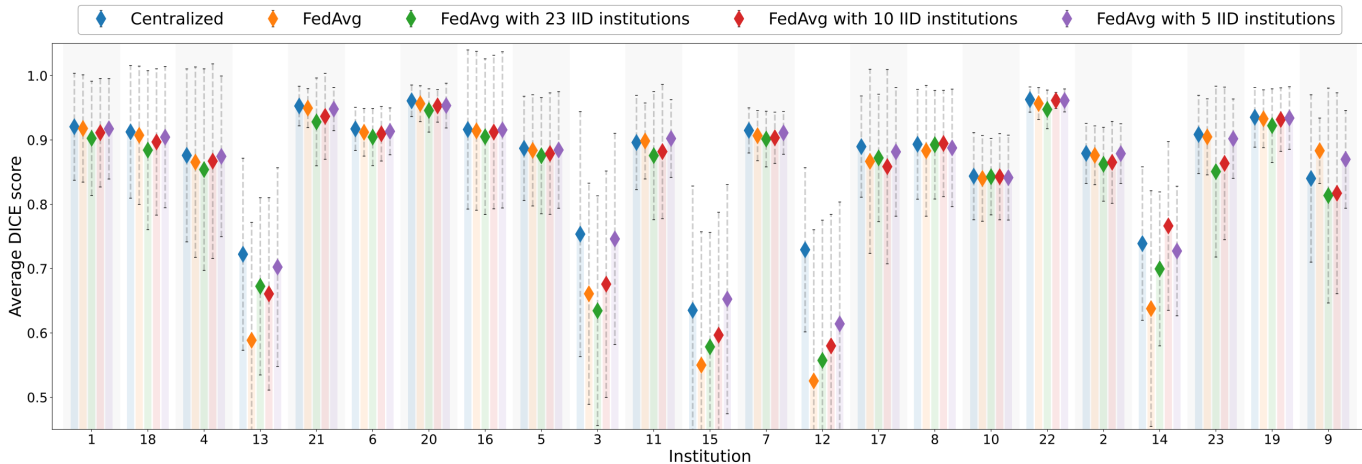


Fig. 7. Average Dice scores of *centralized* and *FedAvg* training in the challenge and *IID* setups of FeTS2022 dataset per institution, in decreasing order of number of test samples (c.f. Figure 1). Errors bars represent  $\pm$  one standard deviation. An *IID* setup limits the drop in performance of *FedAvg* on institutions 12, 13, 14, 15 and 3.

Table 3. Dice scores and 95% Hausdorff distances (mm) obtained by *centralized* and selected state-of-the-art federated training in the *limited setup* of FeTS2022 dataset (average  $\pm$  std over samples of all folds)

Algorithm	Mean DICE	DICE WT	DICE TC	DICE ET	Mean 95HD
Centralized	0.815 $\pm$ 0.161	0.904 $\pm$ 0.098	0.822 $\pm$ 0.188	0.719 $\pm$ 0.308	7.823 $\pm$ 11.962
FedAvg w/ fixed local epochs	0.806 $\pm$ 0.164	0.896 $\pm$ 0.099	0.812 $\pm$ 0.187	0.711 $\pm$ 0.307	8.715 $\pm$ 11.572
SCAFFOLD	0.805 $\pm$ 0.160	0.898 $\pm$ 0.085	0.811 $\pm$ 0.185	0.707 $\pm$ 0.304	7.995 $\pm$ 10.613
Local finetuning	0.806 $\pm$ 0.165	0.891 $\pm$ 0.105	0.808 $\pm$ 0.196	0.718 $\pm$ 0.305	9.332 $\pm$ 12.678
Prior CFL HGG/LGG	0.807 $\pm$ 0.162	0.897 $\pm$ 0.092	0.819 $\pm$ 0.162	0.706 $\pm$ 0.305	8.928 $\pm$ 12.651

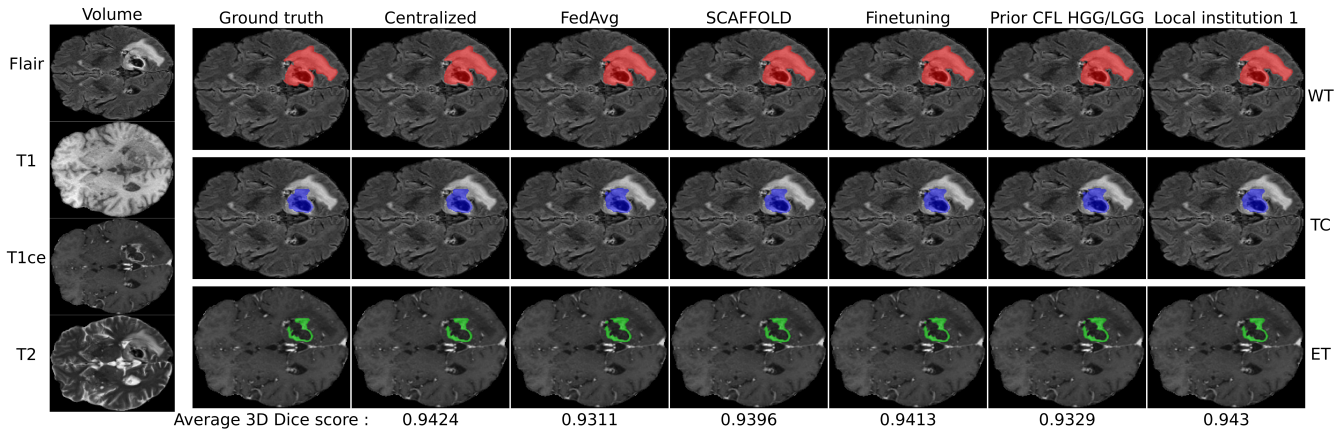
presenting dissimilar label distributions, clustered FL based on this prior knowledge significantly improves the segmentation performance in the challenge. We think that this class of methods is of particular interest in cross-silo medical image analysis tasks with enough data .

For a more qualitative comparison, we provide in Figure 8 and 9 the ground truth and segmentation maps output by models trained with various methods on axial slices of two chosen samples. The first one on Figure 8 is a sample from Institution 1 segmented accurately by most methods, representative of a large number of volumes in the dataset. The second one on Figure 9 is an LGG sample from Institution 13 on which models underperform, representative of most samples from institutions 12 to 15. We see in Figure 8 that the segmentations are very similar. A 0.5 to 1.0 % Dice score difference does not change fundamentally their shape, although some mistakes in the ET label appear in the segmentation made the model trained with *FedAvg* compared to *Centralized*. On the contrary, large differences in segmentation results can be seen in figure 9, where a large part of the TC label was omitted by the model trained with *FedAvg* compared to baselines. From both of these examples, *SCAFFOLD* seems to behave much more like centralized training, even when comparing shapes of the obtained segmentations.

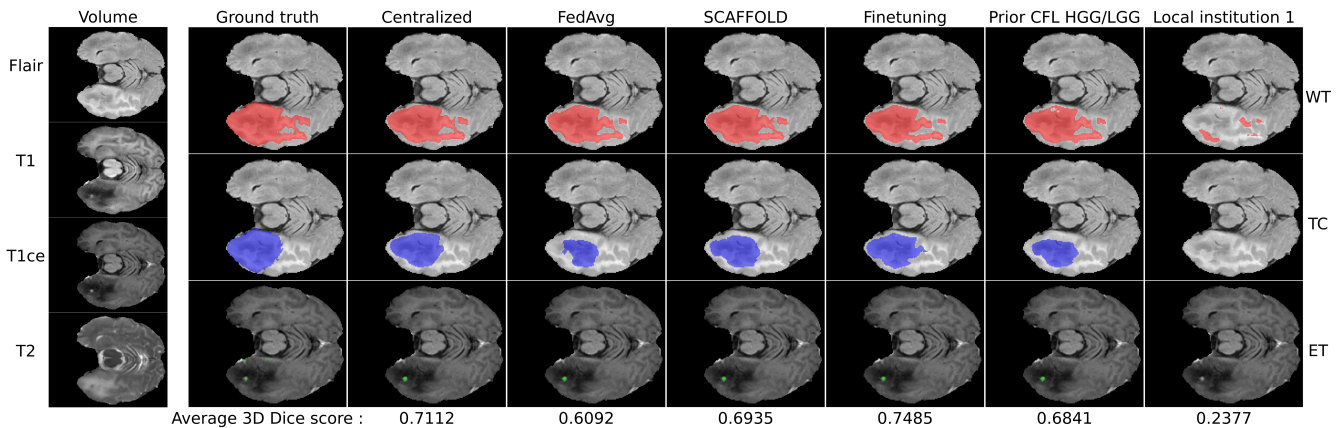
It is intriguing to see how good the performance of standard *FedAvg* is compared to centralized training in the challenge setup. Compared to the *IID* setup with 23 institutions, *FedAvg* in the challenge setup requires fewer communication

rounds to achieve close to centralized performance. We think it is mainly due to institutions 1 and 18 each "centralizing" a large number of samples, driving federated training with large updates. While unbalanced setups could have a priori been seen as a problem, extreme unbalance might actually be exploitable to accelerate federated training. Data-sharing strategies were already proposed in the literature to accelerate federated training, sharing only a fraction of the total data to the server Zhao et al. (2018). We show here that the presence of a large institution, potentially built as a coalition of entities inside the federation, with a sufficiently representative set of samples can be of interest in cross-silo setups. We do not know what type and level of heterogeneity such an extreme unbalance would be robust to, but the goal of newly defined optimizers in this setup becomes to limit the bias induced by this unbalance toward the large institutions. In essence, this might motivate the complementarity between the deployment of federated learning and the construction of large public medical image data centres: these large public datasets could be leveraged at server-side to become the institution driving the others.

Finally, while we tried to roughly balance estimated training times among algorithms, it was based on debatable hypotheses on computation and communication costs. The actual training time of algorithms distributing computation and communication in very different ways such as *FedAvg w/ fixed local epochs* and *fixed local iterations* would be highly dependent on the actual federated infrastructure in practice. The reported estimated training times in Table 2 are comparable between each other and



**Fig. 8.** Qualitative comparison on example axial slice from patient n°00440 from institution 1, representative of most samples of this institution. First column shows, from top to bottom, the original FLAIR, T1, T1ce and T2 MRI images. On the second column are ground truth segmentations for this slice, and on next columns are the segmentations obtained by the 3D UNet model trained with the different FL algorithms. The three rows correspond to the three segmentation labels (from top to bottom, Tumor Core (TC), Whole tumor (WT) and Enhancing Tumor (ET)). *FedAvg* and other federated methods give a well-performing model on the majority of the samples of the dataset.



**Fig. 9.** Qualitative comparison on axial slice from patient n°01514 from institution 13. The figure is structured in the same way as Figure 8. For atypical samples from institutions 12, 13, 14, 15 and 3, the choice of federated learning algorithm has an important impact on the model's segmentation.

only serve as a way to get a fair benchmark, but would not be representative of real federated training times as we proposed a very simplistic estimation model (we do not consider random stragglers for example).

## 6.2. Choice of SOTA methods

We emphasize in this section that the choice of SOTA methods to include in the benchmark was not random. We covered as many lines of research as possible matching the task problems in Section 2 and tried to include in the benchmark one method per "class". Difficulties and benefits from each method can in some cases be translatable to methods of the same class (for example, most partial model-sharing methods would be harmed by the very low amount of samples most institutions own).

Moreover, we voluntarily did not consider some SOTA methods in this benchmark. Ensemble knowledge distillation-based methods such as *FedDF* Lin et al. (2021) require access to public or generated data server-side, which seemed impractical in our medical context. Moreover, most distillation-based methods require the communication of segmentation maps, which

can end up being expensive relative to the size of a convolutional model. *pFedHN* Shamsian et al. (2021) could not scale enough to be used in our case, the segmentation model to be generated by pFedHN's hypernetwork is too large.

We also focused on methods requiring close to the same amount of communication and computation as *FedAvg*, for fair comparability but also simply due to computational constraints, up to a maximum of two times larger computational and communication cost compared to *FedAvg*. We could not experiment with *FedFOMO* Zhang et al. (2021), requiring too many forward passes per communication round.

## 6.3. Limitations of the proposed benchmark

We focused in this study on the performance side of federated learning for brain tumor segmentation. While federated learning was initially introduced as a privacy-preserving framework, it was quickly shown that this setup can still be attacked, potentially leaking information on local data to the server or an individual attacker Rodríguez-Barroso et al. (2023). We do not cover in this study the combination of SOTA federated learning

methods with privacy-preserving techniques such as differential privacy Dwork and Roth (2013) or heavier cryptographic primitives for computational reasons and leave this for future work. For the same reason, we could not incorporate communication optimization techniques such as gradient sparsification or quantization Kairouz *et al.* (2021).

We only use standard SGD as the local optimizer in every experiments. It was shown in a variety of works that local adaptivity, if not handled well, might lead to inconsistencies in training Li *et al.* (2019); Wang *et al.* (2021); Jin *et al.* (2022). As there is no consensus on how local momentum should be applied in a federated setup, for the "safety" of this benchmark and its consistency with the implemented SOTA methods such as *SCAFFOLD*, we chose standard SGD as the local optimizer. The exploration of local optimizers such as SGD with different types of momentum or Adam is a valuable line of exploration left for future work.

Cross-silo federated learning involving health institutions requires them to invest in a technical pole, buy GPU(s) and set up a complex and secure network environment for this task alone. Assuming the full availability of each platform for federated training seems natural, justifying the assumption of full participation of each institution in each federated round.

In this context, centralized federated learning (with a central server orchestrating training) is more secure and demands less communication than fully-decentralized (peer-to-peer) learning on a complete communication graph while giving the exact same results. Indeed, each institution would behave exactly as a central server in this scenario. We thus focused exclusively on the centralized setup, letting the exploration of gossip algorithms on sparse communication graphs for other works.

We only investigated generalization performance of trained models on institutions taking part in training and not on unseen institutions as in Yuan *et al.* (2022b). The choice of institutions to leave out of training for unseen tests would greatly impact the federated learning behaviour in our case, and performing a nested-cross-validation on this aspect was out of our computational limitations. This would be a great line of work, especially considering the recently published BraTS 2023 cluster of datasets Adewole *et al.* (2023); Kazerooni *et al.* (2024).

Finally, our choice to report performance of *Prior CFL HGG/LGG* was not aimed at advocating the use of this specific prior-based CFL method for this task. It showed the potential benefits of using domain knowledge in federated learning, that there exists at least one institution clustering of FeTS2022 institutions for which finetuning on each cluster gives a significant performance improvement, and thus that current SOTA gradient-based CFL methods are not completely able to capture the intricacies of the dataset, highlighting a need for further research. The practicality of clustering data as a function of the tumor grade can be questioned. It seems reasonable to assume the availability or the possibility of acquiring this information for training samples, even in other practical scenarios than FeTS. For just acquired test volumes however, this tumor grade information would most probably be lacking. It then opens a new question: how can we use in practice *Prior CFL* on test samples for which we have very limited information.

## 7. Conclusion

We experimented in this work with a large variety of SOTA federated learning methods on the recently published FeTS2022 dataset and tried to isolate classes of federated methods showing potential on this task. We showed that both SOTA global, personalized and hybrid solutions can improve the segmentation performance compared to standard *FedAvg*, with *SCAFFOLD* and hybrid federated learning seeming the most adapted for cross-silo medical image segmentation tasks. We hope that this study can be helpful both for future methodological works as well as real-world application of federated learning on similar tasks.

## Acknowledgement

This work was partially supported by the Agence Nationale de la Recherche under grant ANR-20-THIA-0007 (IADoc@UdL). It was granted access to the HPC resources of IDRIS under the allocation 2023-AD011013327R1 made by GENCI.

## References

- Adewole, M., Rudie, J.D., Gbadamosi, A., Toyobo, O., Raymond, C., Zhang, D., Omidiji, O., Akinola, R., Suwaid, M.A., Emegoakor, A., Ojo, N., Aguh, K., Kalaiwo, C., Babatunde, G., Ogunleye, A., Gbadamosi, Y., Iorpagher, K., Calabrese, E., Aboian, M., Linguraru, M., Albrecht, J., Wiestler, B., Kofler, F., Janas, A., LaBella, D., Kzerooni, A.F., Li, H.B., Iglesias, J.E., Farahani, K., Eddy, J., Bergquist, T., Chung, V., Shinohara, R.T., Wiggins, W., Reitman, Z., Wang, C., Liu, X., Jiang, Z., Familiar, A., Van Leemput, K., Bukas, C., Piraud, M., Conte, G.M., Johansson, E., Meier, Z., Menze, B.H., Baid, U., Bakas, S., Dako, F., Fatade, A., Anazodo, U.C., 2023. The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa). URL: <http://arxiv.org/abs/2305.19369>, doi:10.48550/arXiv.2305.19369. arXiv:2305.19369 [physics].
- Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S., 2019. Federated Learning with Personalization Layers. arXiv:1912.00818 [cs, stat] URL: <http://arxiv.org/abs/1912.00818>. arXiv: 1912.00818.
- Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., Prevedello, L.M., Rudie, J.D., Sako, C., Shinohara, R.T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., Schaffter, T., Yu, T., Zheng, J., Moawad, A.W., Coelho, L.O., McDonnell, O., Miller, E., Moron, F.E., Oswood, M.C., Shih, R.Y., Siakallis, L., Bronstein, Y., Mason, J.R., Miller, A.F., Choudhary, G., Aggarwal, A., Besada, C.H., Derakhshan, J.J., Diogo, M.C., Do-Dai, D.D., Farage, L., Go, J.L., Hadi, M., Hill, V.B., Iv, M., Joyner, D., Lincoln, C., Lotan, E., Miyakoshi, A., Sanchez-Montano, M., Nath, J., Nguyen, X.V., Nicolas-Jilwan, M., Jimenez, J.O., Ozturk, K., Petrovic, B.D., Shah, C., Shah, L.M., Sharma, M., Simsek, O., Singh, A.K., Soman, S., Statevych, V., Weinberg, B.D., Young, R.J., Ikuta, I., Agarwal, A.K., Cambron, S.C., Silbergleit, R., Dusoi, A., Postma, A.A., Letourneau-Guillon, L., Perez-Carrillo, G.J.G., Saha, A., Soni, N., Zaharchuk, G., Zohrabian, V.M., Chen, Y., Cekic, M.M., Rahman, A., Small, J.E., Sethi, V., Davatzikos, C., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J., Freymann, J.B., Kirby, J.S., Wiestler, B., Crivellaro, P., Colen, R.R., Kotrotsou, A., Marcus, D., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H., Wiest, R., Jakab, A., Weber, M.A., Mahajan, A., Menze, B., Flanders, A.E., Bakas, S., 2021. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. URL: <http://arxiv.org/abs/2107.02314>, doi:10.48550/arXiv.2107.02314. arXiv:2107.02314 [cs].
- Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data 4.

- Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., Wiestler, B., Colen, R., Kotrotsou, A., Lamontagne, P., Marcus, D., Milchenko, M., Nazeri, A., Weber, M.A., Mahajan, A., Baid, U., Gerstner, E., Kwon, D., Acharya, G., Agarwal, M., Alam, M., Albiol, A., Albiol, A., Albiol, F.J., Alex, V., Allinson, N., Amorim, P.H.A., Amrutkar, A., Anand, G., Andermatt, S., Arbel, T., Arbelaez, P., Avery, A., Azmat, M., B., P., Bai, W., Banerjee, S., Barth, B., Batchelder, T., Batmanghelich, K., Battistella, E., Beers, A., Belyaev, M., Bendszus, M., Benson, E., Bernal, J., Bharath, H.N., Biros, G., Bisdas, S., Brown, J., Cabezas, M., Cao, S., Cardoso, J.M., Carver, E.N., Casamitjana, A., Castillo, L.S., Catà, M., Cattin, P., Cerigues, A., Chagas, V.S., Chandra, S., Chang, Y.J., Chang, S., Chang, K., Chazalon, J., Chen, S., Chen, W., Chen, J.W., Chen, Z., Cheng, K., Choudhury, A.R., Chylla, R., Clérigues, A., Colleman, S., Colmeiro, R.G.R., Combalia, M., Costa, A., Cui, X., Dai, Z., Dai, L., Daza, L.A., Deutsch, E., Ding, C., Dong, C., Dong, S., Dudzik, W., Eaton-Rosen, Z., Egan, G., Escudero, G., Estienne, T., Everson, R., Fabrizio, J., Fan, Y., Fang, L., Feng, X., Ferrante, E., Fidon, L., Fischer, M., French, A.P., Fridman, N., Fu, H., Fuentes, D., Gao, Y., Gates, E., Gering, D., Gholami, A., Gierke, W., Glocker, B., Gong, M., González-Villá, S., Grosgees, T., Guan, Y., Guo, S., Gupta, S., Han, W.S., Han, I.S., Harmuth, K., He, H., Hernández-Sabaté, A., Herrmann, E., Himthani, N., Hsu, W., Hsu, C., Hu, X., Hu, X., Hu, Y., Hu, Y., Hua, R., Huang, T.Y., Huang, W., Van Huffel, S., Huo, Q., HV, V., Iftekharruddin, K.M., Isensee, F., Islam, M., Jackson, A.S., Jambawalikar, S.R., Jesson, A., Jian, W., Jin, P., Jose, V.J.M., Jungo, A., Kainz, B., Kamnitsas, K., Kao, P.Y., Karnawat, A., Kellermeier, T., Kermi, A., Keutzer, K., Khadir, M.T., Khened, M., Kickingereder, P., Kim, G., King, N., Knapp, H., Knecht, U., Kohli, L., Kong, D., Kong, X., Koppers, S., Kori, A., Krishnamurthi, G., Krivov, E., Kumar, P., Kushibar, K., Lachinov, D., Lambrou, T., Lee, J., Lee, C., Lee, Y., Lee, M., Lefkovits, S., Lefkovits, L., Levitt, J., Li, T., Li, H., Li, W., Li, H., Li, X., Li, Y., Li, H., Li, Z., Li, X., Li, Z., Li, X., Li, W., Lin, Z.S., Lin, F., Lio, P., Liu, C., Liu, B., Liu, X., Liu, M., Liu, J., Liu, L., Llado, X., Lopez, M.M., Lorenzo, P.R., Lu, Z., Luo, L., Luo, Z., Ma, J., Ma, K., Mackie, T., Madabushi, A., Mahmoudi, I., Maier-Hein, K.H., Majji, P., Mammen, C.P., Mang, A., Manjunath, B.S., Marcinkiewicz, M., McDonagh, S., McKenna, S., McKinley, R., Mehl, M., Mehta, S., Mehta, R., Meier, R., Meinel, C., Merhof, D., Meyer, C., Miller, R., Mitra, S., Moiyadi, A., Molina-Garcia, D., Monteiro, M.A.B., Mrukwa, G., Myronenko, A., Nalepa, J., Ngo, T., Nie, D., Ning, H., Niu, C., Nuechterlein, N.K., Oermann, E., Oliveira, A., Oliveira, D.D.C., Oliver, A., Osman, A.F.I., Ou, Y.N., Ourselin, S., Paragios, N., Park, M.S., Paschke, B., Pauloski, J.G., Pawar, K., Pawlowski, N., Pei, L., Peng, S., Pereira, S.M., Perez-Beteta, J., Perez-Garcia, V.M., Pezold, S., Pham, B., Phophalia, A., Piella, G., Pillai, G.N., Piraud, M., Pisov, M., Popli, A., Pound, M.P., Pourreza, R., Prasanna, P., Prkowska, V., Pridmore, T.P., Puch, S., Puybareau, E., Qian, B., Qiao, X., Rajchl, M., Rane, S., Rebsamen, M., Ren, H., Ren, X., Revanuru, K., Rezaei, M., Rippel, O., Rivera, L.C., Robert, C., Rosen, B., Rueckert, D., Safwan, M., Salem, M., Salvi, J., Sanchez, I., Sánchez, I., Santos, H.M., Sartor, E., Schellingerhout, D., Scheufeke, K., Scott, M.R., Scussel, A.A., Sedlar, S., Serrano-Rubio, J.P., Shah, N.J., Shah, N., Shaiikh, M., Shankar, B.U., Shboul, Z., Shen, H., Shen, D., Shen, L., Shen, H., Shenoy, V., Shi, F., Shin, H.E., Shu, H., Sima, D., Sinclair, M., Smedby, O., Snyder, J.M., Soltaninejad, M., Song, G., Soni, M., Stawiaski, J., Subramanian, S., Sun, L., Sun, R., Sun, J., Sun, K., Sun, Y., Sun, G., Sun, S., Suter, Y.R., Szilagyi, L., Talbar, S., Tao, D., Tao, D., Teng, Z., Thakur, S., Thakur, M.H., Tharakan, S., Tiwari, P., Tochon, G., Tran, T., Tsai, Y.M., Tseng, K.L., Tuan, T.A., Turlapov, V., Tustison, N., Vakilopoulou, M., Valverde, S., Vanguri, R., Vasiliev, E., Ventura, J., Vera, L., Vercauteren, T., Verrastro, C.A., Vidyaratne, L., Vilaplana, V., Vivekanandan, A., Wang, G., Wang, Q., Wang, C.J., Wang, W., Wang, D., Wang, R., Wang, Y., Wang, C., Wang, G., Wen, N., Wen, X., Wenginger, L., Wick, W., Wu, S., Wu, Q., Wu, Y., Xia, Y., Xu, Y., Xu, X., Xu, P., Yang, T.L., Yang, X., Yang, H.Y., Yang, J., Yang, H., Yang, G., Yao, H., Ye, X., Yin, C., Young-Moxon, B., Yu, J., Yue, X., Zhang, S., Zhang, A., Zhang, K., Zhang, X., Zhang, L., Zhang, X., Zhang, Y., Zhang, L., Zhang, J., Zhang, X., Zhang, T., Zhao, S., Zhao, Y., Zhao, X., Zhao, L., Zheng, Y., Zhong, L., Zhou, C., Zhou, X., Zhou, F., Zhu, H., Zhu, J., Zhuge, Y., Zong, W., Kalpathy-Cramer, J., Farahani, K., Davatzikos, C., van Leemput, K., Menze, B., 2019. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. URL: <http://arxiv.org/abs/1811.02629>. arXiv:1811.02629 [cs, stat].
- Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2019. Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images, in: Crimi, A., Bakas, S., Kuijff, H., Keyvan, F., Reyes, M., van Walsum, T. (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, Cham. pp. 161–169. doi:10.1007/978-3-030-11723-8\_16.
- Bercea, C.I., Wiestler, B., Rueckert, D., Albarqouni, S., 2022. Federated disentangled representation learning for unsupervised brain anomaly detection. *Nature Machine Intelligence* 4, 685–695. URL: <https://www.nature.com/articles/s42256-022-00515-2>. doi:10.1038/s42256-022-00515-2. number: 8 Publisher: Nature Publishing Group.
- Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., Lohöfer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., Lipková, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., Hülsemeyer, C., Beetz, M., Ettliger, F., Antonelli, M., Bae, W., Bellver, M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.W., Georgescu, B., Giró-i Nieto, X., Gruen, F., Han, X., Heng, P.A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., Jäger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., van Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B., 2023. The Liver Tumor Segmentation Benchmark (LiTS). *Medical Image Analysis* 84, 102680. URL: <https://www.sciencedirect.com/science/article/pii/S1361841522003085>. doi:10.1016/j.media.2022.102680.
- Camajori Tedeschini, B., Savazzi, S., Stoklasa, R., Barbieri, L., Stathopoulos, I., Nicoli, M., Serio, L., 2022. Decentralized Federated Learning for Healthcare Networks: A Case Study on Tumor Segmentation. *IEEE Access* 10, 8693–8708. doi:10.1109/ACCESS.2022.3141913. conference Name: IEEE Access.
- Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A., 2022. MONAI: An open-source framework for deep learning in healthcare. URL: <http://arxiv.org/abs/2211.02701>. doi:10.48550/arXiv.2211.02701. arXiv:2211.02701 [cs].
- Chen, H.Y., Chao, W.L., 2020. FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning. URL: <https://openreview.net/forum?id=dgtpE6gKjHn>.
- Dinh, C.T., Tran, N.H., Nguyen, T.D., 2020. Personalized federated learning with moreau envelopes, in: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA. pp. 21394–21405.
- Dwork, C., Roth, A., 2013. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 211–407. URL: <http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042>. doi:10.1561/04000000042.
- Fallah, A., Mokhtari, A., Ozdaglar, A., 2020. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc.. pp. 3557–3568. URL: <https://proceedings.neurips.cc/paper/2020/hash/24389bfe4fe2eba8bf9aa9203a44cdd-Abstract.html>.
- Finn, C., Abbeel, P., Levine, S., 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, in: *International Conference on Machine Learning*, PMLR. pp. 1126–1135. URL: <https://proceedings.mlr.press/v70/finn17a.html>. ISSN: 2640-3498.

- Futrega, M., Milesi, A., Marcinkiewicz, M., Ribalta, P., 2021. Optimized U-Net for Brain Tumor Segmentation. URL: <http://arxiv.org/abs/2110.03352>, doi:10.48550/arXiv.2110.03352. arXiv:2110.03352 [cs, eess].
- Futrega, M., Milesi, A., Marcinkiewicz, M., Ribalta, P., 2022. Optimized U-Net for Brain Tumor Segmentation, in: Crimi, A., Bakas, S. (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, Cham. pp. 15–29. doi:10.1007/978-3-031-09002-8\_2.
- Ghosh, A., Chung, J., Yin, D., Ramchandran, K., 2020. An efficient framework for clustered federated learning, in: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA. pp. 19586–19597.
- Hatamizadeh, A., Yin, H., Molchanov, P., Myronenko, A., Li, W., Dogra, P., Feng, A., Flores, M.G., Kautz, J., Xu, D., Roth, H.R., 2023. Do Gradient Inversion Attacks Make Federated Learning Unsafe? *IEEE Transactions on Medical Imaging*, 1–1doi:10.1109/TMI.2023.3239391. conference Name: *IEEE Transactions on Medical Imaging*.
- Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., Zhang, Y., 2021. Personalized Cross-Silo Federated Learning on Non-IID Data. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 7865–7873. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/16960>. number: 9.
- Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation 18, 203–211. URL: <https://www.nature.com/articles/s41592-020-01008-z>, doi:10.1038/s41592-020-01008-z. number: 2 Publisher: Nature Publishing Group.
- Isik-Polat, E., Polat, G., Kocuyigit, A., Temizel, A., 2022. Evaluation and Analysis of Different Aggregation and Hyperparameter Selection Methods for Federated Brain Tumor Segmentation, in: Crimi, A., Bakas, S. (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, Cham. pp. 405–419. doi:10.1007/978-3-031-09002-8\_36.
- Jhunjunhwal, D., Wang, S., Joshi, G., 2023. FedExp: Speeding Up Federated Averaging via Extrapolation. URL: <http://arxiv.org/abs/2301.09604>, doi:10.48550/arXiv.2301.09604. arXiv:2301.09604 [cs].
- Jin, J., Ren, J., Zhou, Y., Lyu, L., Liu, J., Dou, D., 2022. Accelerated federated learning with decoupled adaptive optimization URL: <https://arxiv.org/abs/2207.07223>, doi:10.48550/ARXIV.2207.07223. publisher: arXiv Version Number: 1.
- Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A., Bonawitz, K., Charles, Z.B., Cormode, G., Cummings, R., D’Oliveira, R.G.L., Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, O., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S.U., Sun, Z., Suresh, A.T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F., Yu, H., Zhao, S., 2021. *Advances and Open Problems in Federated Learning*. *Found. Trends Mach. Learn.* doi:10.1561/22000000083.
- Karargyris, A., Umeton, R., Sheller, M.J., Aristizabal, A., George, J., Bala, S., Beutel, D.J., Bittorf, V., Chaudhari, A., Chowdhury, A., Coleman, C., Desinghu, B., Diamos, G., Dutta, D., Feddema, D., Fursin, G., Guo, J., Huang, X., Kanter, D., Kashyap, S., Lane, N., Mallick, I., Mascagni, P., Mehta, V., Natarajan, V., Nikolov, N., Padoy, N., Pekhimenko, G., Reddi, V.J., Reina, G.A., Ribalta, P., Rosenthal, J., Singh, A., Thiagarajan, J.J., Wuest, A., Xenochristou, M., Xu, D., Yadav, P., Rosenthal, M., Loda, M., Johnson, J.M., Mattson, P., 2021. MedPerf: Open Benchmarking Platform for Medical Artificial Intelligence using Federated Evaluation. URL: <http://arxiv.org/abs/2110.01406>, doi:10.48550/arXiv.2110.01406. arXiv:2110.01406 [cs].
- Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T., 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning, in: *International Conference on Machine Learning*.
- Kazerouni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., Bagheri, S., Baid, U., Bergquist, T., Borja, A.J., Calabrese, E., Chung, V., Conte, G.M., Dako, F., Eddy, J., Ezhov, I., Familiar, A., Farahani, K., Haldar, S., Iglesias, J.E., Janas, A., Johansen, E., Jones, B.V., Koffler, F., LaBella, D., Lai, H.A., Van Leemput, K., Li, H.B., Maleki, N., McAllister, A.S., Meier, Z., Menze, B., Moawad, A.W., Nandolia, K.K., Pavaine, J., Piraud, M., Poussaint, T., Prabhu, S.P., Reitman, Z., Rodriguez, A., Rudie, J.D., Sanchez-Montano, M., Shaikh, I.S., Shah, L.M., Sheth, N., Shinohara, R.T., Tu, W., Viswanathan, K., Wang, C., Ware, J.B., Wiestler, B., Wiggins, W., Zapaishchykova, A., Aboian, M., Bornhorst, M., de Blank, P., Deutsch, M., Fouladi, M., Hoffman, L., Kann, B., Lazow, M., Mikael, L., Nabavizadeh, A., Packer, R., Resnick, A., Rood, B., Vossough, A., Bakas, S., Linguraru, M.G., 2024. The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). URL: <http://arxiv.org/abs/2305.17033>, doi:10.48550/arXiv.2305.17033. arXiv:2305.17033 [cs, eess, q-bio].
- Khan, M.I., Jafaritadi, M., Alhoniemi, E., Kontio, E., Khan, S.A., 2022. Adaptive Weight Aggregation in Federated Learning for Brain Tumor Segmentation, in: Crimi, A., Bakas, S. (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, Cham. pp. 455–469. doi:10.1007/978-3-031-09002-8\_40.
- Li, Q., He, B., Song, D., 2021a. Model-Contrastive Federated Learning. URL: <http://arxiv.org/abs/2103.16257>, doi:10.48550/arXiv.2103.16257. arXiv:2103.16257 [cs].
- Li, T., Hu, S., Beirami, A., Smith, V., 2021b. Ditto: Fair and Robust Federated Learning Through Personalization, in: *Proceedings of the 38th International Conference on Machine Learning*, PMLR. pp. 6357–6368. URL: <https://proceedings.mlr.press/v139/li21h.html>. iSSN: 2640-3498.
- Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V., 2020a. Federated Optimization in Heterogeneous Networks. URL: <http://arxiv.org/abs/1812.06127>, doi:10.48550/arXiv.1812.06127. arXiv:1812.06127 [cs, stat].
- Li, T., Sanjabi, M., Beirami, A., Smith, V., 2020b. Fair Resource Allocation in Federated Learning. URL: <https://openreview.net/forum?id=ByexE1SYDr>.
- Li, W., Milletari, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., Feng, A., 2019. Privacy-Preserving Federated Brain Tumour Segmentation, in: Suk, H.I., Liu, M., Yan, P., Lian, C. (Eds.), *Machine Learning in Medical Imaging*, Springer International Publishing, Cham. pp. 133–141. doi:10.1007/978-3-030-32692-0\_16.
- Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S., 2020c. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis* 65, 101765. URL: <https://www.sciencedirect.com/science/article/pii/S1361841520301298>, doi:10.1016/j.media.2020.101765.
- Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z., 2020d. On the Convergence of FedAvg on Non-IID Data. URL: <http://arxiv.org/abs/1907.02189>, doi:10.48550/arXiv.1907.02189. arXiv:1907.02189 [cs, math, stat].
- Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q., 2020e. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. URL: <https://openreview.net/forum?id=6YEQUn0QICG>.
- Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q., 2021c. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. URL: <http://arxiv.org/abs/2102.07623>, doi:10.48550/arXiv.2102.07623. arXiv:2102.07623 [cs].
- Liang, P.P., Liu, T., Ziyin, L., Allen, N.B., Auerbach, R.P., Brent, D., Salakhutdinov, R., Morency, L.P., 2020. Think Locally, Act Globally: Federated Learning with Local and Global Representations. URL: <http://arxiv.org/abs/2001.01523>, doi:10.48550/arXiv.2001.01523. arXiv:2001.01523 [cs, stat].
- Lin, T., Kong, L., Stich, S.U., Jaggi, M., 2021. Ensemble Distillation for Robust Model Fusion in Federated Learning. arXiv:2006.07242 [cs, stat] URL: <http://arxiv.org/abs/2006.07242>. arXiv: 2006.07242.
- Liu, C., Lou, C., Wang, R., Xi, A.Y., Shen, L., Yan, J., 2022. Deep Neural Network Fusion via Graph Matching with Applications to Model Ensemble and Federated Learning, in: *Proceedings of the 39th International Conference on Machine Learning*, PMLR. pp. 13857–13869. URL: <https://proceedings.mlr.press/v162/liu22k.html>. iSSN: 2640-3498.
- Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F.K., Chen, T.Y., Mahmood, F., 2022. Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis* 76, 102298. URL: <https://www.sciencedirect.com/science/article/pii/S1361841521003431>, doi:10.1016/j.media.2021.102298.
- Maier, O., Menze, B.H., von der Gablentz, J., Hani, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens,

- D., Dutil, F., Egger, K., Feng, C., Glocker, B., Götz, M., Haeck, T., Halme, H.L., Havaei, M., Iftekharuddin, K.M., Jodoin, P.M., Kamnitsas, K., Kellner, E., Korvenoja, A., Laroche, H., Ledig, C., Lee, J.H., Maes, F., Mahmood, Q., Maier-Hein, K.H., McKinley, R., Muschelli, J., Pal, C., Pei, L., Rangarajan, J.R., Reza, S.M.S., Robben, D., Rueckert, D., Salli, E., Suetens, P., Wang, C.W., Wilms, M., Kirschke, J.S., Kr Amer, U.M., Munte, T.F., Schramm, P., Wiest, R., Handels, H., Reyes, M., 2017. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical Image Analysis* 35, 250–269. doi:10.1016/j.media.2016.07.009.
- Mansour, Y., Mohri, M., Ro, J., Suresh, A.T., 2020. Three Approaches for Personalization with Applications to Federated Learning. arXiv:2002.10619 [cs, stat] URL: <http://arxiv.org/abs/2002.10619>. arXiv: 2002.10619.
- Marfoq, O., Neglia, G., Bellet, A., Kameni, L., Vidal, R., 2022a. Federated Multi-Task Learning under a Mixture of Distributions. URL: <https://openreview.net/forum?id=YcQx6zhEzRp>.
- Marfoq, O., Neglia, G., Kameni, L., Vidal, R., 2022b. Personalized Federated Learning through Local Memorization. URL: <http://arxiv.org/abs/2111.09360>. doi:10.48550/arXiv.2111.09360. arXiv:2111.09360 [cs, stat].
- McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y., 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data, in: *Artificial Intelligence and Statistics*.
- Mohri, M., Sivek, G., Suresh, A.T., 2019. Agnostic Federated Learning, in: *Proceedings of the 36th International Conference on Machine Learning*, PMLR. pp. 4615–4625. URL: <https://proceedings.mlr.press/v97/mohri19a.html>. ISSN: 2640-3498.
- Mächler, L., Ezhov, I., Kofler, F., Shit, S., Paetzold, J.C., Loehr, T., Wiestler, B., Menze, B., 2021. FedCostWAvg: A new averaging for better Federated Learning. arXiv:2111.08649 [cs] URL: <http://arxiv.org/abs/2111.08649>. arXiv: 2111.08649.
- Mächler, L., Ezhov, I., Shit, S., Paetzold, J.C., 2023. FedPIDAvg: A PID controller inspired aggregation method for federated learning. URL: <http://arxiv.org/abs/2304.12117>, arXiv:2304.12117 [cs].
- Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., Rektorova, I., Bonanni, L., Pardini, M., Kramberger, M.G., Taylor, J.P., Hort, J., Snædal, J., Kulisevsky, J., Blanc, F., Antonini, A., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H., Lovestone, S., Simmons, A., Aarsland, D., Westman, E., 2020. The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study. *Medical Image Analysis* 66, 101714. URL: <https://www.sciencedirect.com/science/article/pii/S1361841520300785>, doi:10.1016/j.media.2020.101714.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. URL: <http://arxiv.org/abs/1912.01703>, doi:10.48550/arXiv.1912.01703. arXiv:1912.01703 [cs, stat].
- Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.H., Reina, G.A., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C., Sako, C., Ghodasara, S., Bilello, M., Mohan, S., Vollmuth, P., Brugnara, G., Preetha, C.J., Sahm, F., Maier-Hein, K., Zenk, M., Bendszus, M., Wick, W., Calabrese, E., Rudie, J., Villanueva-Meyer, J., Cha, S., Ingalhalikar, M., Jadhav, M., Pandey, U., Saini, J., Garrett, J., Larson, M., Jeraj, R., Currie, S., Frood, R., Fatania, K., Huang, R.Y., Chang, K., Balaña, C., Capellades, J., Puig, J., Trenkler, J., Pichler, J., Necker, G., Haunschild, A., Meckel, S., Shukla, G., Liem, S., Alexander, G.S., Lombardo, J., Palmer, J.D., Flanders, A.E., Dicker, A.P., Sair, H.I., Jones, C.K., Venkataraman, A., Jiang, M., So, T.Y., Chen, C., Heng, P.A., Dou, Q., Kozubek, M., Lux, F., Michálek, J., Matula, P., Keřkovský, M., Kopřivová, T., Dostál, M., Vybítal, V., Vogelbaum, M.A., Mitchell, J.R., Farinhas, J., Maldjian, J.A., Yogananda, C.G.B., Pinho, M.C., Reddy, D., Holcomb, J., Wagner, B.C., Ellingson, B.M., Cloughesy, T.F., Raymond, C., Oughourlian, T., Hagiwara, A., Wang, C., To, M.S., Bhardwaj, S., Chong, C., Agzarian, M., Falcão, A.X., Martins, S.B., Teixeira, B.C.A., Sprenger, F., Menotti, D., Lucio, D.R., LaMontagne, P., Marcus, D., Wiestler, B., Kofler, F., Ezhov, I., Metz, M., Jain, R., Lee, M., Lui, Y.W., McKinley, R., Slotboom, J., Radojewski, P., Meier, R., Wiest, R., Murcia, D., Fu, E., Haas, R., Thompson, J., Ormond, D.R., Badve, C., Sloan, A.E., Vadmal, V., Waite, K., Colen, R.R., Pei, L., Ak, M., Srinivasan, A., Bapuraj, J.R., Rao, A., Wang, N., Yoshiaki, O., Moritani, T., Turk, S., Lee, J., Prabhudesai, S., Morón, F., Mandel, J., Kamnitsas, K., Glocker, B., Dixon, L.V.M., Williams, M., Zampakis, P., Panagiotopoulos, V., Tsiganos, P., Alexiou, S., Haliassos, I., Zacharaki, E.I., Moustakas, K., Kalogeropoulou, C., Kardamakis, D.M., Choi, Y.S., Lee, S.K., Chang, J.H., Ahn, S.S., Luo, B., Poisson, L., Wen, N., Tiwari, P., Verma, R., Bareja, R., Yadav, I., Chen, J., Kumar, N., Smits, M., van der Voort, S.R., Alafandi, A., Incecar, F., Wijnenga, M.M.J., Kapsas, G., Gahrman, R., Schouten, J.W., Dubbink, H.J., Vincent, A.J.P.E., van den Bent, M.J., French, P.J., Klein, S., Yuan, Y., Sharma, S., Tseng, T.C., Adabi, S., Niclou, S.P., Keunen, O., Hau, A.C., Vallières, M., Fortin, D., Lepage, M., Landman, B., Ramadas, K., Xu, K., Chotai, S., Chambless, L.B., Mistry, A., Thompson, R.C., Gusev, Y., Bhuvaneshwar, K., Sayah, A., Bencheqroun, C., Belouali, A., Madhavan, S., Booth, T.C., Chelliah, A., Modat, M., Shuaib, H., Dragos, C., Abayazeed, A., Kolodziej, K., Hill, M., Abbassy, A., Gamal, S., Mekhaimar, M., Qayati, M., Reyes, M., Park, J.E., Yun, J., Kim, H.S., Mahajan, A., Muzi, M., Benson, S., Beets-Tan, R.G.H., Teuwen, J., Herrera-Trujillo, A., Trujillo, M., Escobar, W., Abello, A., Bernal, J., Gómez, J., Choi, J., Baek, S., Kim, Y., Ismael, H., Allen, B., Buatti, J.M., Kotrotsou, A., Li, H., Weiss, T., Weller, M., Bink, A., Pouymayou, B., Shaykh, H.F., Saltz, J., Prasanna, P., Shrestha, S., Mani, K.M., Payne, D., Kurc, T., Pelaez, E., Franco-Maldonado, H., Loayza, F., Quevedo, S., Guevara, P., Torche, E., Mendoza, C., Vera, F., Ríos, E., López, E., Velastin, S.A., Ogbole, G., Soneye, M., Oyekunle, D., Odafe-Oyibotho, O., Osobu, B., Shu'aibu, M., Dorcas, A., Dako, F., Simpson, A.L., Hamghalam, M., Peoples, J.J., Hu, R., Tran, A., Cutler, D., Moraes, F.Y., Boss, M.A., Gimpel, J., Veetil, D.K., Schmidt, K., Bialecki, B., Marella, S., Price, C., Cimino, L., Apgar, C., Shah, P., Menze, B., Barnholtz-Sloan, J.S., Martin, J., Bakas, S., 2022. Federated learning enables big data for rare cancer boundary detection. *Nature Communications* 13, 7346. URL: <https://www.nature.com/articles/s41467-022-33407-5>, doi:10.1038/s41467-022-33407-5. number: 1 Publisher: Nature Publishing Group.
- Pati, S., Baid, U., Zenk, M., Edwards, B., Sheller, M., Reina, G.A., Foley, P., Gruzdev, A., Martin, J., Albarqouni, S., Chen, Y., Shinohara, R.T., Reinke, A., Zimmerer, D., Freymann, J.B., Kirby, J.S., Davatzikos, C., Colen, R.R., Kotrotsou, A., Marcus, D., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H., Wiest, R., Jakab, A., Weber, M.A., Mahajan, A., Maier-Hein, L., Kleesiek, J., Menze, B., Maier-Hein, K., Bakas, S., 2021. The Federated Tumor Segmentation (FeTS) Challenge. arXiv:2105.05874 [cs, eess].
- Pillutla, K., Malik, K., Mohamed, A., Rabbat, M., Sanjabi, M., Xiao, L., 2022. Federated Learning with Partial Model Personalization. URL: <http://arxiv.org/abs/2204.03809>, doi:10.48550/arXiv.2204.03809. arXiv:2204.03809 [cs, math].
- Qu, L., Balachandrar, N., Zhang, M., Rubin, D., 2022a. Handling data heterogeneity with generative replay in collaborative learning for medical imaging. *Medical Image Analysis* 78, 102424. URL: <https://www.sciencedirect.com/science/article/pii/S1361841522000755>, doi:10.1016/j.media.2022.102424.
- Qu, L., Zhou, Y., Liang, P.P., Xia, Y., Wang, F., Adeli, E., Fei-Fei, L., Rubin, D., 2022b. Rethinking Architecture Design for Tackling Data Heterogeneity in Federated Learning, in: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, New Orleans, LA, USA. pp. 10051–10061. URL: <https://ieeexplore.ieee.org/document/9880336/>, doi:10.1109/CVPR52688.2022.00982.
- Rawat, A., Zizzo, G., Kadhe, S., Epperlein, J.P., Braghin, S., 2022. Robust learning protocol for federated tumor segmentation challenge .
- Reddi, S.J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B., 2022. Adaptive Federated Optimization. URL: <https://openreview.net/forum?id=LkFG31B13U5>.
- Reina, G.A., Gruzdev, A., Foley, P., Perepelkina, O., Sharma, M., Davidyuk, I., Trushkin, I., Radionov, M., Mokrov, A., Agapov, D., Martin, J., Edwards, B., Sheller, M.J., Pati, S., Moorthy, P.N., Wang, S.h., Shah, P., Bakas, S., 2022. OpenFL: An open-source framework for Federated Learning. *Physics in Medicine & Biology* 67.
- Rodríguez-Barroso, N., Jiménez-López, D., Luzón, M.V., Herrera, F., Martínez-Cámara, E., 2023. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion* 90, 148–173. URL: <https://www.sciencedirect.com/science/article/pii/S1566253522001439>, doi:10.1016/j.inffus.2022.09.011.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J.,

- Wells, W.M., Frangi, A.F. (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham. pp. 234–241. doi:10.1007/978-3-319-24574-4\_28.
- Sattler, F., Müller, K.R., Samek, W., 2021. Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems* 32, 3710–3722. doi:10.1109/TNNLS.2020.3015958. conference Name: IEEE Transactions on Neural Networks and Learning Systems.
- Shamsian, A., Navon, A., Fetaya, E., Chechik, G., 2021. Personalized Federated Learning using Hypernetworks, in: *Proceedings of the 38th International Conference on Machine Learning*, PMLR. pp. 9489–9502. URL: <https://proceedings.mlr.press/v139/shamsian21a.html>. iSSN: 2640-3498.
- Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R., Bakas, S., 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports* 10, 12598. URL: <https://www.nature.com/articles/s41598-020-69250-1>, doi:10.1038/s41598-020-69250-1. number: 1 Publisher: Nature Publishing Group.
- Shen, Y., Zhou, Y., Yu, L., 2022. CD $\$$ 2 $\$$ pFed: Cyclic Distillation-guided Channel Decoupling for Model Personalization in Federated Learning. URL: <http://arxiv.org/abs/2204.03880>. arXiv:2204.03880 [cs].
- Shoham, N., Avidor, T., Keren, A., Israel, N., Benditkis, D., Mor-Yosef, L., Zeitak, I., 2019. Overcoming Forgetting in Federated Learning on Non-IID Data. URL: <http://arxiv.org/abs/1910.07796>. arXiv:1910.07796 [cs, stat].
- Sun, B., Huo, H., YANG, Y., Bai, B., 2021. PartialFed: Cross-Domain Personalized Federated Learning via Partial Initialization, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc.. pp. 23309–23320. URL: <https://proceedings.neurips.cc/paper/2021/hash/c429429bf1f2af051f2021dc92a8e8bea-Abstract.html>.
- Tang, Z., Zhang, Y., Shi, S., He, X., Han, B., Chu, X., 2022. Virtual Homogeneity Learning: Defending against Data Heterogeneity in Federated Learning, in: *Proceedings of the 39th International Conference on Machine Learning*, PMLR. pp. 21111–21132. URL: <https://proceedings.mlr.press/v162/tang22d.html>. iSSN: 2640-3498.
- Tuladhar, A., Tyagi, L., Souza, R., Forkert, N.D., . Federated learning using variable local training for brain tumor segmentation, in: Crimi, A., Bakas, S. (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing. pp. 392–404. doi:10.1007/978-3-031-09002-8\_35.
- Wachinger, C., Rieckmann, A., Pölsterl, S., 2021. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis* 67, 101879. URL: <https://www.sciencedirect.com/science/article/pii/S1361841520302437>, doi:10.1016/j.media.2020.101879.
- Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., Khazaeni, Y., 2020a. Federated learning with matched averaging, in: *International Conference on Learning Representations*.
- Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V., 2020b. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc.. pp. 7611–7623. URL: <https://proceedings.neurips.cc/paper/2020/hash/564127c03caab942e503ee6f810f54fd-Abstract.html>.
- Wang, J., Xu, Z., Garrett, Z., Charles, Z., Liu, L., Joshi, G., 2021. Local adaptivity in federated learning: Convergence and consistency. URL: <http://arxiv.org/abs/2106.02305>, doi:10.48550/arXiv.2106.02305, arXiv:2106.02305 [cs, stat].
- Wang, S., Ji, M., 2022. A Unified Analysis of Federated Learning with Arbitrary Client Participation. URL: <http://arxiv.org/abs/2205.13648>, doi:10.48550/arXiv.2205.13648. arXiv:2205.13648 [cs, math, stat].
- Wang, Y., Shi, Q., Chang, T.H., 2023. Why Batch Normalization Damage Federated Learning on Non-IID Data? URL: <http://arxiv.org/abs/2301.02982>. arXiv:2301.02982 [cs].
- Wicaksana, J., Yan, Z., Zhang, D., Huang, X., Wu, H., Yang, X., Cheng, K.T., 2022. FedMix: Mixed Supervised Federated Learning for Medical Image Segmentation. *IEEE Transactions on Medical Imaging* , 1–1doi:10.1109/TMI.2022.3233405. conference Name: IEEE Transactions on Medical Imaging.
- Wu, Y., Zeng, D., Wang, Z., Shi, Y., Hu, J., 2022. Distributed contrastive learning for medical image segmentation. *Medical Image Analysis* 81, 102564. URL: <https://www.sciencedirect.com/science/article/pii/S1361841522002079>, doi:10.1016/j.media.2022.102564.
- Xu, A., Li, W., Guo, P., Yang, D., Roth, H., Hatamizadeh, A., Zhao, C., Xu, D., Huang, H., Xu, Z., 2022. Closing the Generalization Gap of Cross-silo Federated Medical Image Segmentation, *IEEE Computer Society*. pp. 20834–20843. URL: <https://www.computer.org/csdl/proceedings-article/cvpr/2022/694600u0834/1H1iUX8eCbe>, doi:10.1109/CVPR52688.2022.02020.
- Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H.R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., Zhu, W., Carrafiello, G., Patella, F., Cariati, M., Obinata, H., Mori, H., Tamura, K., An, P., Wood, B.J., Xu, D., 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. *Medical Image Analysis* 70, 101992. doi:10.1016/j.media.2021.101992.
- Yang, H., Ge, M., Xue, D., Xiang, K., Li, H., Lu, R., 2023. Gradient Leakage Attacks in Federated Learning: Research Frontiers, Taxonomy and Future Directions. *IEEE Network* , 1–8doi:10.1109/MNET.001.2300140. conference Name: IEEE Network.
- Yu, F., Zhang, W., Qin, Z., Xu, Z., Wang, D., Liu, C., Tian, Z., Chen, X., 2021. Fed2: Feature-Aligned Federated Learning, in: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, New York, NY, USA. pp. 2066–2074. URL: <https://doi.org/10.1145/3447548.3467309>, doi:10.1145/3447548.3467309.
- Yuan, H., Morningstar, W.R., Ning, L., Singhal, K., 2022a. What do we mean by generalization in federated learning?, in: *International Conference on Learning Representations*.
- Yuan, H., Morningstar, W.R., Ning, L., Singhal, K., 2022b. What Do We Mean by Generalization in Federated Learning? URL: [https://openreview.net/forum?id=VimqQq-i\\_Q](https://openreview.net/forum?id=VimqQq-i_Q).
- Zeineldin, R.A., Karar, M.E., Burgert, O., Mathis-Ullrich, F., 2022. Multimodal CNN Networks for Brain Tumor Segmentation in MRI: A BraTS 2022 Challenge Solution. URL: <http://arxiv.org/abs/2212.09310>, doi:10.48550/arXiv.2212.09310. arXiv:2212.09310 [cs, eess].
- Zhang, M., Sapra, K., Fidler, S., Yeung, S., Alvarez, J.M., 2021. Personalized Federated Learning with First Order Model Optimization. Technical Report arXiv:2012.08565. arXiv. URL: <http://arxiv.org/abs/2012.08565>, doi:10.48550/arXiv.2012.08565. arXiv:2012.08565 [cs, stat] type: article.
- Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V., 2018. Federated Learning with Non-IID Data. arXiv:1806.00582 [cs, stat] URL: <http://arxiv.org/abs/1806.00582>. arXiv: 1806.00582.
- Zhu, L., Liu, Z., Han, S., 2019. Deep Leakage from Gradients, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc. URL: <https://papers.nips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html>.
- Zhu, W., Luo, J., 2022. Federated Medical Image Analysis with Virtual Sample Synthesis, in: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer Nature Switzerland, Cham. volume 13433, pp. 728–738. URL: [https://link.springer.com/10.1007/978-3-031-16437-8\\_70](https://link.springer.com/10.1007/978-3-031-16437-8_70), doi:10.1007/978-3-031-16437-8\_70. series Title: Lecture Notes in Computer Science.

**Appendix A. Hyperparameters search and selected values**

**Appendix B. Additional figures of the limited data setup**



Table A.4. Hyperparameters search in the different FeTS2022 partitionings

Method	Parameter	Search space	Selected value
Challenge setup			
<i>Centralized, institution 1 local</i>	learning rate	{0.008, 0.02, 0.05, 0.1, 0.2}	0.1
<i>FedAvg and variations, SCAFFOLD, CFL</i>	learning rate	{0.02, 0.05, 0.1, 0.2, 0.4, 0.8}	0.4
<i>FedNova</i>	learning rate	{0.05, 0.1, 0.2, 0.4}	0.2
<i>FedAdam</i>	local learning rate	{0.01, 0.1, 1.0}	0.1
	server learning rate	{0.0001, 0.001, 0.01, 0.1, 1.0}	0.001
	$\beta_1, \beta_2, \tau$	$\emptyset$	0.9, 0.999, $10^{-8}$
<i>q-FedAvg</i>	learning rate	{0.05, 0.1, 0.2, 0.4, 0.8}	0.4
	$q$	{0.01, 0.1, 1.0, 2.0}	1.0
<i>FedPIDAvg</i>	learning rate	{0.05, 0.1, 0.2, 0.4}	0.4
	$\alpha, \beta, \gamma$	$\emptyset$	0.45, 0.45, 0.1
<i>Local finetuning</i>	learning rate	{0.01, 0.02, 0.05, 0.1, 0.2, 0.4}	per institution
<i>Ditto</i>	learning rate	{0.01, 0.04, 0.1, 0.4}	per institution
	$\lambda$	{0.001, 0.01, 0.1, 1.0}	
<i>FedPer</i>	learning rate	{0.02, 0.05, 0.1, 0.2, 0.4}	0.4
	number of private layers	$\emptyset$	4
<i>LG-FedAvg</i>	learning rate	{0.02, 0.05, 0.1, 0.2, 0.4}	0.2
	number of private layers	$\emptyset$	4
<i>HGG/LGG CFL</i>	learning rate	{0.02, 0.05, 0.1, 0.2}	0.1
IID setup			
<i>FedAvg IID</i>	learning rate	$\emptyset$	0.4
Limited setup			
<i>Centralized, institution 1 local</i>	learning rate	{0.05, 0.1, 0.3, 0.8}	0.1
<i>FedAvg and SCAFFOLD</i>	learning rate	{0.05, 0.1, 0.2, 0.4, 0.8}	0.8
<i>Local finetuning</i>	learning rate	{0.01, 0.02, 0.05, 0.1, 0.2, 0.4}	per institution
<i>HGG/LGG CFL</i>	learning rate	{0.02, 0.05, 0.1, 0.2, 0.4}	0.2

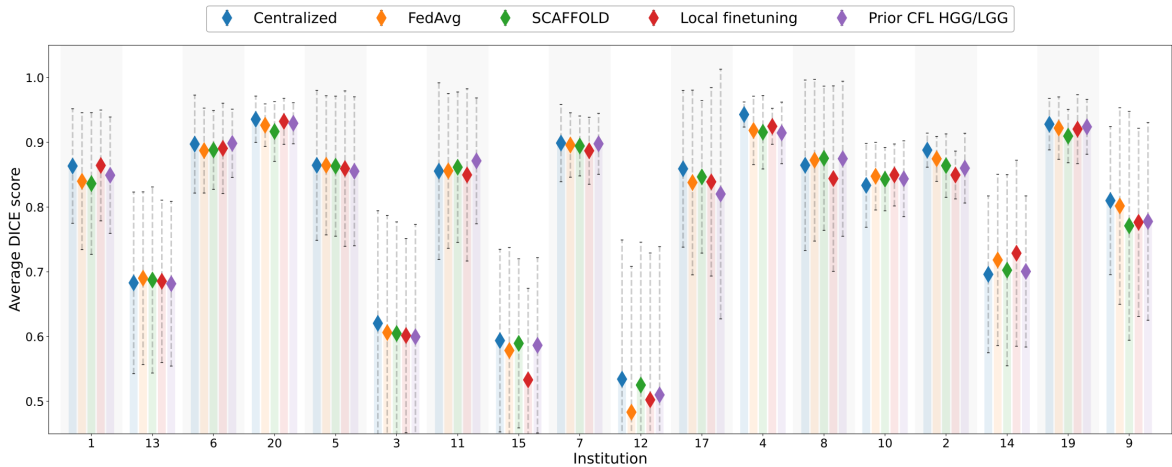


Fig. B.10. Average Dice scores of *centralized* and *FedAvg* training in *Limited data* setup of FeTS2022 dataset per institution, in decreasing order of number of test samples (c.f. Figure 1). Errors bars represent  $\pm$  one standard deviation.