Influence of Data Distribution on Federated Learning Performance in Tumor Segmentation

Abbreviations

IID: independent and identically distributed

HCC: hepatocellular carcinoma

FNH: focal nodular hyperplasia

GBM: glioblastoma

EMD: Earth mover's distance

BD: Bhattacharyya distance

CSD: Chi-square distance

KSD: Kolmogorov–Smirnov distance

NET: non-enhancing tumor region

ET: enhancing tumor region

Summary

Federated deep learning model performance in tumor segmentation on CT and MRI was affected by differences in data distributions, which was strongly negatively correlated with the distance between data distributions.

Key Points

(1) The Dice coefficient ratio between federated and centralized models (*theta*) was strongly negatively correlated to Earth mover's distance (EMD) (r=-0.920), Bhattacharyya distance (BD) (r=-0.893) and chi-square distance (CSD) (r=-0.899) values between data distributions, indicating that federating deep learning model performance in tumor segmentation on CT and MRI decreases as distance between datasets increases.

(2) Data distributions of federated models with significantly different performances ($p<0.05$) from centralized models had significantly higher distances (EMD, BD, and CSD) than those of federated models showing no difference in performance.

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

Abstract

Purpose: To investigate the correlation between differences in data distributions and federated deep learning (Fed-DL) algorithm performance in tumor segmentation on CT and MRI.

Materials and Methods: Two Fed-DL datasets were retrospectively collected (from November 2020 to December 2021), one dataset of liver tumor CT images (named "FILTS" for Federated Imaging in Liver Tumor Segmentation; 3 sites, 692 scans) and one publicly available dataset of brain tumor MRIs (named "FeTS" for Federated Tumor Segmentation; 23 sites, 1251 scans). Scans from both datasets were grouped according to site, tumor type, tumor size, dataset size, and tumor intensity. To quantify differences in data distributions, the following four distance metrics were calculated: Earth mover's distance (EMD), Bhattacharyya distance (BD), Chi-square distance (CSD), and Kolmogorov–Smirnov Distance (KSD). Both federated and centralized nnU-Net models were trained by using the same grouped datasets. Fed-DL model performance was evaluated by using the ratio of Dice coefficients, *theta*, between federated and centralized models trained and tested on the same 80:20 split datasets.

Results: The Dice coefficient ratio (*theta*) between federated and centralized models was strongly negatively correlated with the distances between data distributions, with correlation coefficients of -0.920 for EMD, -0.893 for BD, and -0.899 for CSD. However, KSD was weakly correlated with *theta*, with a correlation coefficient of -0.479.

Conclusion: Performance of Fed-DL models in tumor segmentation on CT and MRI datasets were strongly negatively correlated with the distances between data distributions.

KEYWORDS: federated deep learning, tumor segmentation, data distribution

1 Introduction

Over the past decade, deep learning (DL) has been successfully applied in various medical imaging applications such as tumor segmentation [1]. However, state-of-the-art performance of DL models depends largely on the use of diverse training data. The establishment of a centralized, large-scale, multi-institutional labelled medical imaging dataset is not only challenging and costly, but compliance with General Data Protection Regulation and Health Insurance Portability and Accountability Act (HIPAA) guidelines is often associated with various legal, privacy, security, and data-ownership obstacles [2].

One way to overcome these obstacles is through federated deep learning (Fed-DL) [3, 4], in which model training is distributed among multiple sites by exchanging model data instead of raw patient data via the network, decoupling the need of a centralized dataset. Several recent works [5-7] have demonstrated that Fed-DL provides a promising solution to training of DL models while protecting patient privacy. Current Fed-DL research focuses mainly on algorithm performance evaluation between centralized and federated trained models. For instance, Sheller et al [8] demonstrated that Fed-DL could achieve similar performance to centralized models when data were split and distributed among 10 sites. Lee et al [9] showed similar findings using an imbalanced number of scans among sites. However, little is known regarding the impact of data difference on Fed-DL model performance in tumor segmentation.

In general, Fed-DL requires that data distributions among sites are independent and identically distributed (IID) to achieve comparable performance to that of a centralized model. However, real-world datasets are often non-IID due to differences in factors such as disease manifestation, imaging protocols, or patient populations, leading to potential degradation of model performance. Zhao et al [10] reported that the accuracy of federated models reduced when the Earth mover's distance (EMD) of non-IID natural image datasets increased, but the authors did not compare federated and centralized models. To provide a benchmark for evaluation of Fed-DL models to differences in data distribution, the Radiological Society of North America (RSNA) launched the first Federated Tumor Segmentation (FeTS) challenge in 2021 focusing on segmentation of brain tumors using MRI [11].

The purpose of our study was to investigate the correlation between the distance of data distributions and performance of Fed-DL models in tumor segmentation on CT and MRI. To the best of our knowledge, this is the first systematic study focusing on the impact of data difference on Fed-DL performance in tumor segmentation. Our specific aims were as follows: (1) build a large multi-institutional hepatic CT dataset for benchmarking Fed-DL performance of liver tumor segmentation; (2) calculate quantitative metrics for measuring the distance (difference) between data distributions; and (3) investigate the correlation between the distances of data distributions and the performances of Fed-DL in tumor segmentation.

2 Materials and Methods

This retrospective, HIPAA-compliant study was approved by the institutional review board for data analysis of internal and external datasets collected at the involved sites, and the need for patient informed consent was waived. All DICOM (Digital Imaging and Communications in Medicine) images were de-identified at the original institutions before being transferred to our study.

2.1 Liver Tumor CT Dataset

We established a hepatic CT dataset for training and validation of Fed-DL models of liver tumor segmentation, which we named “FILTS” for Federated Imaging of Liver Tumor Segmentation. For the construction of FILTS, we retrospectively collected 692 hepatic contrast-enhanced CT scans from three sites, including 131 scans from the Liver Tumor Segmentation (LiTS) challenge (Site A, Europe) [12], 156 scans from Massachusetts General Hospital (MGH) (Site B, the US), and 405 scans from the Second Affiliated Hospital at Zhejiang University School of Medicine (SAHZU) (Site C, China). All scans at Site B and Site C were collected for liver tumor segmentation. The inclusion criteria were as follows: (1) at least one focal liver lesion diagnosed using CT; (2) confirmation of malignant tumors by corresponding pathologic reports; (3) diagnosis of benign lesions through pathological analyses or a combination of typical image performance and clinical data. As a result, fifteen scans (n=15) that did not contain any focal liver lesions were excluded (Figure 1a). The collected scans were acquired by using different imaging protocols at various CT scanners (GE, Siemens, and Philips), with a largely varying in-

plane resolution from 0.52 mm to 1.0 mm and section thickness from 0.45 mm to 6.0 mm (Figure 1b).

LiTS is a publicly available liver CT dataset, which was collected from seven hospitals and research institutions in Europe. As the institute information of each scan was removed, we treated LiTS as Site A with heterogeneous scans. LiTS only provides portal venous phase liver CT images, which were acquired with different CT scanners and acquisition protocols. The primary and secondary tumor types in LiTS are hepatocellular carcinoma (HCC) and metastases (ME). The segmentations provided by the LiTS were reviewed by a senior radiologist (with >10 years' experience of abdominal CT reading). In total, 734 tumors with an average size of 13.6 cm^3 were annotated, and 75% of these tumors smaller than 5 cm^3.

Site B data were mainly HCC scans collected from September 2005 to August 2015 at MGH, acquired on two types of CT scanners (Siemens and GE). Three hepatic phases of CT images were collected: arterial, portal venous, and delayed phase. Tumors were contoured in portal venous phase with reference to arterial phase on an open software, 3D Quantitative Imaging (3DQI, V1.0) (https://3dqi.mgh.harvard.edu) by one junior radiologist (3 years' experience) and confirmed by the senior radiologist. In total, 762 tumors with an average size of 61.4 cm^3 were contoured.

Site C data were hepatic CT scans collected from January 2016 to December 2018 at SAHZU, including three types of benign liver tumors (focal nodular hyperplasia (FNH), hemangioma (HEM), and cysts), and three types of malignant liver tumors (HCC, ME, and intrahepatic cholangiocarcinoma (ICC)). Pre-contrast, arterial, and portal venous phase images, acquired on three types of CT scanners (Siemens, GE and Philips), were collected. Tumors were contoured by one junior radiologist (5 years' experience) using an open-source software (ITK-SNAP) [13] and confirmed by the senior radiologist. In total, 585 tumors with an average size of 38.2 cm^3 were contoured.

Figure 2 compares three examples of CT scan from each of the three sites and CT attenuation distributions (histograms) of tumors among the three sites.

2.2 Brain Tumor MRI Dataset

The FeTS 2021 dataset [11] is the first Fed-DL medical image dataset, which was collected from multiple sites with different clinical protocols, and contains 1251 total scans (with both images and segmentations). FeTS consists of a subset of glioblastoma (GBM) scans from the Brain Tumor Segmentation (BraTS) dataset [14] containing institutional information and an additional collection of GBM scans from other independent institutions. Each scan in FeTS includes 4 sequences (pre- and post-contrast T1-weighted, T2-weighted, and T2-FLAIR). These scans have been preprocessed using the same steps, including co-registration, resampling (1 mm *1 mm * 1 mm) and skull stripping. Tumors were contoured by one to four readers sharing the same contouring standard and were then confirmed by experienced neuroradiologists. For our study, we performed segmentation of GBM in post-contrast T1 images, which includes the non-enhancing tumor (NET) and enhancing tumor (ET) regions. Because FeTS contains only GBM (grade 4 glioma), we additionally collected scans of three types of diffuse glioma (astrocytoma, glioblastoma, oligodendroglioma), which were histopathologically proven grade 2-4, from the newly published University of California, San Francisco (UCSF) preoperative diffuse glioma MRI dataset (UCSF-PDGM) [15] to study the impact of different types of glioma on Fed-DL performance. A total of 500 post-contrast T1-weighted MRI scans were added to our study.

2.3 Data Grouping

We grouped both the FILTS and FeTS datasets respectively, according to site, tumor type, tumor size, dataset size, and tumor density (CT) / intensity (MRI) for evaluation of Fed-DL model performance on different types of data distribution. Tumor density refers to CT attenuation and tumor intensity is the normalized MRI signal intensity (Z-score) of tumors.

- **Group 1: Different Sites**

In FILTS, three subsets were treated as from three different sites. Portal venous phase CT scans were selected, as this was the only image type provided by Site A. In FeTS, we selected the three sites providing more than 40 scans (Site 1: 512 scans, Site 4: 47 scans, and Site 18: 382 scans). Sites with a small number of scans were not considered as they could introduce high bias.

- **Group 2: Different Tumor Types**

In FILTS, we grouped hepatic CT images at Site C according to six types of liver tumors: HEM, FNH, Cyst, HCC, ICC, and ME. Arterial phase CT scans were selected as certain liver tumor types, such as HCC and FNH, are better visualized in arterial phase than portal venous phase CT. In FeTS, the MRI scans added from UCSF-PDGM were grouped into three subsets based on the three types of diffuse gliomas.

- **Group 3: Different Tumor Sizes**

In FILTS, we grouped all scans into four subsets corresponding to tumor size thresholds of <15 cm^3, 15 – 50 cm^3, 50 – 130 cm^3, and >130 cm^3. In FeTS, we grouped all scans from the 2 largest institutions into 6 subsets by using thresholds of <3 cm^3, 3 – 10 cm^3, and >10 cm^3 for Site 1, and <2 cm^3, 2 – 12 cm^3, and >12 cm^3 for Site 18, respectively. We chose different thresholds in order to keep the number of scans in each subset balanced.

- **Group 4: Different Dataset Sizes**

To assess the effect of imbalanced training datasets on Fed-DL model performance, we first randomly selected different subsets from each site in FILTS, provided that the total number of scans in the two testing subsets remained the same. Because Site A has the least number of scans (86) among the three sites, we randomly selected a similar number of scans (90) from both Site B and Site C. Then, we decreased the number of scans in Site A by 25% (65 scans) and 50% (43 scans), and increased the same number of scans in Site B and Site C. Thus, the ratio of numbers of scans between two subsets decreased from approximately 1.0 (balanced) to 1/3 (imbalanced). For FeTS, the number of scans between Site 4 and Site 1 and Site 4 and Site 18 were highly imbalanced, with ratio of approximately 1/10. We evaluated the FeTS results in Group 1.

- **Group 5: Different Tumor Densities/Intensities**

We grouped the FeTS scans into four subsets (Q1-Q4) using thresholds of MRI signal intensity (SI) on both non-enhancing tumor (NET) and enhancing tumor (ET) regions, SI-Q1 (n=113|): NET < 25% and ET < 25%, SI-Q12 (n=406): NET < 50% and ET < 50%; SI-Q34 (n=387): NET ≥ 50% and ET ≥ 50%; SI-Q4 (n=116): NET ≥ 75% and ET ≥ 75%, respectively. For FILTS,

certain groups of tumors had large differences in tumor density, such as FNH (hyper) vs Cyst (hypo). We evaluated results for the FILTS dataset in Group 2.

2.4 Data Metrics

- **Distance of Data Distribution**

Data distribution specifies the data range and the relative frequency (probability of occurrence) of each data value. A histogram is the most commonly used statistical method to show data distribution. Four metrics were calculated to quantify distance in data distribution: Earth mover's distance (EMD) (or Wasserstein Distance) [16], Bhattacharyya distance (BD) [17], Chi-square distance (CSD) [18], and Kolmogorov–Smirnov distance (KSD) [19].

- **Performance of Tumor Segmentation**

The Dice coefficient is a most well-known metric to evaluate the performance of segmentation. We used the *theta* coefficient to assess performance between a federated model and a centralized model evaluated on the same dataset, defined as follows:

$$Theta = (\text{Dice of federated model}) / (\text{Dice of centralized model}).$$

In general, *theta* is less than 1.0. A *theta* value close to 1.0 means that the federated model achieves similar performance as that of a centralized model in tumor segmentation. Theta was reported as mean ± standard error, of which the standard error was estimated by method using bivariate first-order Taylor expansion (https://www.stat.cmu.edu/~hseltman/files/ratio.pdf).

We developed a federated implementation of nnU-Net [20] based on a server-client architecture and the Fed-Avg algorithm [21]. For a fair comparison between federated and centralized models, we first ran the nnU-Net planning and preprocessing task on all scans by configuration of a 3D U-Net segmentation pipeline, such as resampling, normalization, patch size, and data augmentation parameters. Then, training of either federated or centralized models employed the same pre-processed data and the same set of hyper-parameters. Federated models were trained on the scheme of one server and two clients, each client containing one sub-dataset. All federated and centralized models were trained on an NVIDIA Tesla P40 GPU cluster with 24 GB memory. Data in each group were randomly split into 80% for training and 20% for testing, and the results from the testing data were evaluated.

More technical details of the Fed-DL implementation were described in the Supplemental Materials.

2.5 Statistical Analysis

Paired t-test was performed to assess the difference in performance between a federated model and a centralized model on the same dataset. A *p*-value less than 0.05 rejects the null hypothesis that mean paired Dice difference between a federated model and a centralized model is zero and indicates statistically significant different performances between federated and centralized models.

We also calculated the trendline and Person's correlation coefficients to evaluate the association between *theta* coefficients and distance measures. The trendline is a linear function, $y = kx + b$, where the independent variable, x, is distance, dependent variable, y, is the *theta* value. The correlation coefficient is a measure of the goodness of fit of a linear relationship between *theta* and distance values. Statistical analyses were performed using MedCalc (version 19.5.6), and graphs were created using Microsoft Excel (version 2210).

2.6 Data Availability

The data and the scripts used to perform study evaluations that support the findings will be made publicly available, without due reservation.

3 Results

3.1 Fed-DL Performance on Grouped Data

Performance of Fed-DL models trained on data grouped by site, tumor type, tumor size, dataset size and tumor density/intensity are listed in Table 1-Table 4, respectively.

- **Different Sites**

We found no evidence of a difference between federated and centralized model performance on datasets grouped by site (*p*-values >0.05; Table 1).

- **Different Tumor Types**

Table 2 shows *theta* values ranging from 0.877 to 0.982 in FILTS and 0.975 to 0.999 in FeTS. Figure 3a and 3b show two examples of distance in data distribution between HCC vs. HEM (small distance) and FNH vs. cyst (large distance). Figure 4 shows the distributions of CT attenuation among six types of liver tumors. Of 14 subsets in the FILTS dataset, 5 had significantly different performances (p-values <0.05) between federated and centralized models.

- **Different Tumor Sizes**

Performance of Fed-DL models trained with different groups of tumor sizes are listed in Table 3a (FILTS) and Table 3b (FeTS), respectively. Average *theta* values were high in both the FILTS (0.980±0.154) and FeTS (0.992±0.075) datasets. Figure 3c and 3d show two examples of distance in data distribution between Size 1 vs Size 2 (small tumors) and Size 3 vs Size 4 (large tumors). Although Dice values were higher for large tumors compared with small tumors, the *theta* values remained similar.

- **Different Dataset Sizes**

Table 4a shows performance values of Fed-DL models trained with different numbers of scans in FILTS. Lower ratios of numbers of scans (i.e. more imbalance) led to lower Dice values in both the federated and centralized models. However, average *theta* values remained similar: 0.973±0.150 (*ratio*=1.0), 0.973±0.159 (*ratio*=0.6), and 0.975±0.128 (*ratio*=0.3). In the FeTS dataset (Table 1b), *theta* values remained high even when the *ratio* of numbers of scans was less than 0.3 (e.g., Site 4 : Site 1 = 47 : 512 = 0.092 and Site 4 : Site 18 = 47 : 382 = 0.123).

- **Different Tumor Densities/Intensities**

Performance of federated models trained with different tumor intensities in the FeTS dataset significantly differed from centralized models (Table 4b). Figure 3e and 3f compare histograms of enhancing brain tumors between Site 1 vs. Site 4 (p=0.32) and SI-Q1 vs. SI-Q4 (p=0.003). For FILTS (Table 2a), tumors with different CT density typically had lower *theta* values, such as FNH (hyper) vs Cyst (hypo) and HCC (hyper) vs Cyst (hypo).

3.2 Correlation Analysis

The distances of data distributions were negatively correlated with *theta*, with correlation coefficients of -0.920, -0.893, -0.899 and -0.479 for EMD, BD, CSD, and KSD, respectively. The trendlines in Figure 5 show a negative slope between distance (EMD, BD, CSD, KSD) and *theta*, indicating lower federated model performance with greater distance between data distribution. The waterfall plots of EMD, BD, CSD, and KSD in Figure 6 show the effect of changes in distance of data distribution on performance of federated models compared with centralized models.

There was a significant difference in performance between federated and centralized models for 10 of the 62 total subsets (groups 1 to 5). Corresponding distances in data distribution also differed significantly between federated and centralized models, with values of 13.527±4.506 (median=13.445) vs. 2.722±2.728 (median=1.691) ($p<0.001$) for EMD, 0.691±0.395 (median=0.472) vs. 0.066±0.117 (median=0.025) ($p=0.001$) for BD, 0.618±0.211 (median=0.531) vs. 0.095±0.137 (median=0.046) ($p<0.001$) for CSD, and 0.271±0.097 (median=0.260) vs. 0.186±0.097 (median=0.170) ($p=0.03$) for KSD, respectively.

4 Discussion

In this study, we investigated the correlation between various distance metrics that measure the difference in data distributions and Fed-DL performance in segmentation of liver tumors on CT and brain tumors on MRI. EMD had the strongest, negative correlation ($r=-0.920$) with federated model performance. We found that the between-site difference of tumor density (CT) / intensity (MRI) distributions influenced the Fed-DL performance, which was demonstrated by both liver tumors on CT and brain tumors on MRI. For liver tumors on CT, it was reflected by different tumor types which had different CT attenuations (density), whereas for brain tumors on MRI, it was reflected by tumor regions with different MRI signal intensity. In other words, the Fed-DL performance in tumor segmentation is affected by the difference of CT attenuation or MRI intensity of tumors at different sites. The magnitude of this difference could be measured by EMD, BD or CSD. Other factors including different tumor sizes or imbalanced dataset sizes did not significantly ($p\geq0.05$) impact overall data distribution and thus had little influence on federated model performance.

Our findings are consistent with those of Lee et al [9] and will have substantial impact on the development of Fed-DL using real-world non-IID data. We observed that a key underlying factor affecting the performance of federated models is the distance in data distributions. To achieve comparable performance with a centralized model, a federated model should be trained using datasets with small distances. Many approaches attempted to solve the issue of non-IID data in Fed-DL from the algorithmic perspective, such as episodic learning in continuous frequency space [22], local batch normalization [23], and cross-site modeling [24]. Motivated by our findings, we propose that data augmentation may be a more feasible and practical solution. For example, use of domain adaptation [25] among different clients to reduce data difference measured by EMD may improve Fed-DL performance, even with basic federated algorithms.

The two most common Fed-DL workflows are server-client and peer-to-peer topology [26], and commonly used aggregation methods include Fed-Avg [21], Base + Personalization layer (FedPer) [27], and Federated Matched Averaging (FedMA) [28]. The server-client architecture with the Fed-Avg aggregation algorithm is the most common scheme of Fed-DL. We applied this federated scheme in our study to demonstrate the generalizability of our findings.

There are only a few publicly available Fed-DL medical imaging datasets, including thorax disease classification on chest radiographs [29, 30], skin lesion image classification [31, 32], prostate MRI segmentation [33], and a retinal image database [34]. In particular, the 2021 RSNA Brain Tumor AI challenge based on FeTS (http://www.synapse.org/brats) has facilitated the first formal community benchmark explicitly for Fed-DL aggregation algorithms [11]. As FeTS contains only a single type of glioma, we added three types of glioma collected from the UCSF diffuse glioma MRI dataset (UCSF-PDGM) [15] to investigate the effect of tumor type on Fed-DL performance. Since FeTS and UCSF-PDGM had different imaging protocols and standards, we did not mix UCSF-PDGM scans with FeTS scans in other data groups.

Our study had several limitations. First, Site A used LiTS, which is a multi-site dataset, whereas datasets at Sites B and C were each acquired from a single site, respectively. Although scans from the same site were acquired by using similar imaging protocols on different CT scanners, they also varied in image resolution and image quality. Nevertheless, Site A data may have impacted study findings due to differences in imaging protocols at multiple sites. Second, tumor type was not reported for scans from Sites A and B. Since this was a retrospective study, tumor

type data could not be obtained through tissue biopsy or postoperative pathologic examination. Third, we did not consider the potential effect of inter-reader variability, as segmentation was performed by different readers using different software at different institutions. This might contribute to performance degradation. However, such variability among sites may be unavoidable in real-world federated setting.

In conclusion, differences in data distribution may affect Fed-DL model performance in medical image segmentation. Model performance was strongly negatively correlated with distance (EMD, BD, and CSD) in data distribution. Reducing data distance may provide a feasible solution to ensure development of a high-performing federated model trained on non-IID data.

Acknowledgements

LiTS data for liver tumor in CT were obtained from

https://competitions.codalab.org/competitions/17094

FeTS data for brain tumor in MRI were obtained from http://www.synapse.org/brats.

The study was partially supported by the Children's Tumor Foundation.

References

[1] Minaee S, Boykov Y Y, Porikli F, et al. Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligenc 2022, 44(7): 3523 - 3542.

[2] Kaissis G A, Makowski M R, Rückert D, et al. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2020, 2(6): 305-311.

[3] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design. Proceedings of machine learning and systems, 2019, 1: 374-388.

[4] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-19.

[5] Sheller M J, Reina G A, Edwards B, et al. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, 2019: 92-104.

[6] Li W, Milletarì F, Xu D, et al. Privacy-preserving federated brain tumour segmentation. Machine Learning in Medical Imaging: 10th International Workshop, 2019: 133-141.

[7] Roth H R, Chang K, Singh P, et al. Federated learning for breast density classification: A real-world implementation. Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, 2020: 181-191.

[8] Sheller M J, Edwards B, Reina G A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific reports, 2020, 10(1): 1-12.

[9] Lee G H, Shin S Y. Federated learning on clinical benchmark data: performance assessment. Journal of medical Internet research, 2020, 22(10): e20891.

[10] Zhao Y, Li M, Lai L, et al. Federated learning with non-iid data. arXiv preprint arXiv:1806.00582, 2018.

[11] Pati S, Baid U, Zenk M, et al. The federated tumor segmentation (fets) challenge. arXiv preprint arXiv:2105.05874, 2021.

[12] Bilic P, Christ P, Li H B, et al. The liver tumor segmentation benchmark (lits). Medical Image Analysis, 2023, 84: 102680.

[13] Yushkevich P A, Gao Y, Gerig G. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). 2016: 3342-3345.

[14] Menze B H, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging, 2014, 34(10): 1993-2024.

[15] Calabrese E, Villanueva-Meyer J E, Rudie J D, et al. The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. Radiology: Artificial Intelligence, 2022, 4(6): e220058.

[16] Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval. International journal of computer vision, 2000, 40(2): 99-121.

[17] Bhattacharyya A. On a measure of divergence between two multinomial populations. Sankhyā: the indian journal of statistics, 1946, 7(4): 401-406.

[18] Pele O, Werman M. The quadratic-chi histogram distance family. 11th European Conference on Computer Vision. 2010: 749-762.

[19] Massey Jr F J. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association, 1951, 46(253): 68-78.

[20] Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature methods, 2021, 18(2): 203-211.

[21] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. Artificial intelligence and statistics. PMLR, 2017, 54: 1273-1282.

[22] Liu Q, Chen C, Qin J, et al. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1013-1023.

[23] Li X, Jiang M, Zhang X, et al. FedBN: Federated learning on non-iid features via local batch normalization. ICLR 2021.

[24] Guo P, Wang P, Zhou J, et al. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2423-2432.

[25] Li X, Gu Y, Dvornek N, et al. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Medical Image Analysis, 2020, 65: 101765.

[26] Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ digital medicine, 2020, 3(1): 119.

[27] Arivazhagan M G, Aggarwal V, Singh A K, et al. Federated learning with personalization layers. arXiv preprint arXiv:1912.00818, 2019.

[28] Wang H, Yurochkin M, Sun Y, et al. Federated learning with matched averaging. arXiv preprint arXiv:2002.06440, 2020.

[29] Irvin J, Rajpurkar P, Ko M, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 590-597.

[30] Wang X, Peng Y, Lu L, et al. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2097-2106.

[31] Kawahara J, Daneshvar S, Argenziano G, et al. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics, 2018, 23(2): 538-546.

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

[33] Litjens G, Toth R, Van De Ven W, et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Medical image analysis, 2014, 18(2): 359-373.

[34] Orlando J I, Fu H, Breda J B, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis, 2020, 59: 101570.

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

Tables

Table 1. Data Distribution and Model Performance Metrics for Different Sites

(a) FILTS Dataset

Subset	EMD	BD	CSD	KSD	Fed-Dice	Cent-Dice	*theta*	*p*-value
Site A, Site B	4.7618	0.0768	0.1387	0.2800	0.7365	0.7533	0.9777±0.1469	0.35
Site A, Site C	2.9766	0.0356	0.0658	0.2900	0.7389	0.7495	0.9859±0.1885	0.37
Site B, Site C	2.0255	0.0391	0.0722	0.3400	0.7986	0.8116	0.9839±0.1467	0.44

(b) FeTS Dataset

Subset	Region	EMD	BD	CSD	KSD	Fed-Dice	Cent-Dice	*theta*	*p*-value
Site 1, Site 4	NET	1.3490	0.0243	0.0418	0.1300	0.8070	0.8084	0.9983±0.0905	0.80
	ET	0.5425	0.0022	0.0042	0.1400	0.8708	0.8812	0.9882±0.0208	0.32
Site 1, Site 18	NE	0.9080	0.0074	0.0145	0.1100	0.8185	0.8220	0.9958±0.0279	0.53
	ET	2.4166	0.0127	0.0244	0.1699	0.8717	0.8867	0.9831±0.0569	0.27
Site 4, Site 18	NET	1.5632	0.0336	0.0551	0.0700	0.7626	0.7656	0.9961±0.1809	0.52
	ET	1.9616	0.0093	0.0180	0.1500	0.8613	0.8700	0.9901±0.0523	0.14

Note.— *P*-value calculated using paired t test. FILTS: Federated Imaging of Liver Tumor Segmentation, FeTS: Federated Tumor Segmentation, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test, Fed-Dice: Dice coefficient of federated deep learning, Cent-Dice: Dice coefficient of centralized learning, NET: non-enhancing tumor region, ET: enhancing tumor region

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

Table 2. Data Distribution and Model Performance Metrics for Different Tumor Types

(a) Different tumor types at Site C of FILTS Dataset

Subset	EMD	BD	CSD	KSD	Fed-Dice	Cent-Dice	*theta*	*p*-value
HEM, HCC	3.3443	0.0699	0.1278	0.2800	0.8016	0.8246	0.9722±0.1655	0.33
HEM, FNH *	14.0659	0.6334	0.6698	0.2200	0.7515	0.8011	0.9381±0.1755	0.04
HEM, Cyst	10.0006	0.4519	0.5353	0.2500	0.8116	0.8381	0.9683±0.1694	0.05
FNH, Cyst *	23.7903	1.4561	0.9029	0.4800	0.6898	0.7871	0.8764±0.1612	0.01
HEM, ICC	1.6521	0.0418	0.0733	0.2900	0.7791	0.7964	0.9782±0.1490	0.42
HCC, ICC	3.5908	0.0818	0.1282	0.3999	0.7685	0.7829	0.9815±0.1315	0.47
FNH, ICC *	14.4280	0.9820	0.7964	0.2699	0.6886	0.7856	0.8765±0.1868	0.01
HEM, ME	1.8266	0.0484	0.0835	0.2900	0.7969	0.8146	0.9783±0.1539	0.28
HCC, ME	4.8865	0.1363	0.2056	0.1800	0.7605	0.7891	0.9638±0.1718	0.11
ICC, ME	1.6564	0.0177	0.0340	0.1300	0.7805	0.7991	0.9767±0.1640	0.37
HCC, FNH	10.9597	0.4585	0.5413	0.3200	0.7898	0.8096	0.9755±0.1389	0.09
FNH, ME *	15.8082	1.0224	0.8256	0.2400	0.7281	0.8053	0.9041±0.2133	0.04
HCC, Cyst *	12.8731	0.6264	0.6343	0.3800	0.6899	0.7537	0.9154±0.1418	0.01
ME, Cyst	8.1806	0.3732	0.4648	0.4400	0.6803	0.7295	0.9326±0.1204	0.05

Note.— *P*-value calculated using paired t test; (*) denotes significantly different performances (*p*-values <0.05) between federated and centralized models. FILTS: Federated Imaging of Liver Tumor Segmentation, HCC: hepatocellular carcinoma, FNH: focal nodular hyperplasia, HEM: hemangioma, ME: metastases, ICC: intrahepatic cholangiocarcinoma, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test, Fed-Dice: Dice coefficient of federated deep learning, Cent-Dice: Dice coefficient of centralized learning

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

(b) FeTS Dataset

Subset	Region	EMD	BD	CSD	KSD	Fed-Dice	Cent-Dice	*theta*	*p*-value
DA, GBM	NET	0.5206	0.0054	0.0107	0.0900	0.7875	0.7960	0.9893±0.1834	0.56
	ET	3.5145	0.0309	0.0566	0.2000	0.8680	0.8746	0.9924±0.0322	0.29
DA, OG	NET	0.6596	0.0240	0.0406	0.3800	0.6748	0.6924	0.9746±0.2273	0.23
OG, GBM	NET	0.4098	0.0122	0.0220	0.3600	0.7960	0.7969	0.9989±0.1213	0.93

Note.— *P*-value calculated using paired t test. FeTS: Federated Tumor Segmentation, DA: Diffuse Astrocytoma, GBM: Glioblastoma, OG: Oligodendroglioma, NET: non-enhancing tumor region, ET: enhancing tumor region, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test, Fed-Dice: Dice coefficient of federated deep learning, Cent-Dice: Dice coefficient of centralized learning

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

Table 3. Data Distribution and Model Performance Metrics for Different Tumor Sizes

(a) FILTS Dataset

Subset	EMD	BD	CSD	KSD	Fed-Dice	Cent-Dice	*theta*	*p*-value
Size 1, Size 2	0.9238	0.0071	0.0139	0.1100	0.7357	0.7481	0.9835±0.1581	0.33
Size 1, Size 3	2.3366	0.0244	0.0459	0.1099	0.7270	0.7472	0.9730±0.1635	0.39
Size 1, Size 4	1.8415	0.0158	0.0298	0.1700	0.7289	0.7461	0.9769±0.1932	0.51
Size 2, Size 3	1.7254	0.0134	0.0250	0.0800	0.7906	0.8065	0.9803±0.1443	0.52
Size 2, Size 4	1.2116	0.0126	0.0233	0.1300	0.8016	0.8145	0.9842±0.1610	0.45
Size 3, Size 4	0.7824	0.0114	0.0217	0.0900	0.8293	0.8431	0.9836±0.1037	0.39

Note.—Size 1: tumors ≤ 15 cm3; Size 2: 15 cm3 < tumors ≤ 50 cm3; Size 3: 50 cm3 < tumors ≤ 130 cm3; Size 4: tumors > 130 cm3. P-value calculated using paired t test. FILTS: Federated Imaging of Liver Tumor Segmentation, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test, Fed-Dice: Dice coefficient of federated deep learning, Cent-Dice: Dice coefficient of centralized learning

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

(b) FeTS Dataset

Subset	Region	EMD	BD	CSD	KSD	Fed-Dice	Cent-Dice	*theta*	*p*-value
Size 1, Size 2	NET	2.1026	0.0203	0.0388	0.0400	0.7553	0.7749	0.9747±0.1108	0.15
	ET	0.2556	0.0007	0.0014	0.1600	0.8574	0.8617	0.9950±0.0347	0.21
Size 1, Size 3	NET	2.2500	0.0246	0.0453	0.0490	0.7524	0.7618	0.9876±0.0802	0.65
	ET	0.6471	0.0024	0.0047	0.2400	0.8721	0.8744	0.9973±0.0293	0.43
Size 2, Size 3	NET	0.1547	0.0017	0.0034	0.0499	0.8660	0.8666	0.9993±0.0605	0.79
	ET	0.4390	0.0011	0.0022	0.0899	0.9048	0.9082	0.9963±0.0527	0.59
Size 4, Size 5	NET	0.5629	0.0074	0.0144	0.1600	0.7263	0.7369	0.9855±0.1715	0.25
	ET	0.7704	0.0042	0.0070	0.1100	0.8276	0.8343	0.9920±0.0688	0.39
Size 4, Size 6	NET	0.6532	0.0113	0.0204	0.1000	0.7987	0.8074	0.9892±0.0372	0.33
	ET	0.7267	0.0046	0.0087	0.1700	0.8472	0.8497	0.9970±0.0561	0.34
Size 5, Size 6	NET	0.1913	0.0035	0.0050	0.0600	0.8554	0.8558	0.9995±0.0962	0.84
	ET	0.7350	0.0024	0.0046	0.1300	0.8655	0.8688	0.9962±0.1009	0.37

Note.— Site 1 was grouped into three subsets using tumor core volume as follows: Size 1 ≤ 3 cm^3; Size 2: 3 cm^3 < tumors ≤ 10 cm^3; Size 3: >10 cm^3. Site 18 was grouped into three subsets as fellows: Size 4 ≤ 2 cm^3; Size 5: 2 cm^3 < tumors ≤ 12 cm^3; Size 6: >12 cm^3. P-value calculated using paired t test. FeTS: Federated Tumor Segmentation, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test, Fed-Dice: Dice coefficient of federated deep learning, Cent-Dice: Dice coefficient of centralized learning

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

Table 4. Data Distribution and Model Performance Metrics for Imbalanced Dataset Sizes in FILTS and Different Tumor Intensities in FeTS

(a) Imbalanced numbers of scans in FILTS

Subset (Ratio)	EMD	BD	CSD	KSD	Fed-Dice	Cent-Dice	*theta*	*p*-value
Site A1 : Site B2 (0.96)	4.5940	0.0736	0.1335	0.2800	0.7327	0.7507	0.9760±0.1758	0.32
Site A2 : Site B3 (0.59)	5.1150	0.0828	0.1485	0.2300	0.7053	0.7357	0.9587±0.1844	0.12
Site A1 : Site C2 (0.96)	5.0405	0.0799	0.1374	0.1690	0.7198	0.7535	0.9553±0.1667	0.26
Site A2 : Site C3 (0.59)	4.4345	0.0612	0.1086	0.1000	0.6854	0.7073	0.9690±0.1619	0.40
Site A3 : Site C5 (0.32)	3.7734	0.0453	0.0826	0.1100	0.6918	0.7128	0.9705±0.1456	0.29
Site B2 : Site C2 (1.00)	1.2402	0.0377	0.0683	0.2000	0.7716	0.7823	0.9863±0.1072	0.18
Site C4 : Site B3 (0.59)	1.5885	0.0361	0.0658	0.2300	0.7608	0.7745	0.9823±0.1288	0.50
Site B4 : Site C3 (0.59)	1.3235	0.0421	0.0751	0.2300	0.7592	0.7738	0.9811±0.1628	0.34
Site B5 : Site C5 (0.32)	1.1844	0.0424	0.0755	0.1900	0.7536	0.7689	0.9801±0.1100	0.42

Note.— Number of scans: Site A1 (86), Site B2 (90), Site C2 (90); Site A2 (65), Site B3 (111), Site C3 (111); Site A3 (43), Site B4 (65), Site C4 (65); Site B5 (43), Site C5 (133). *P*-value calculated using paired t test. FILTS: Federated Imaging in Liver Tumor Segmentation, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test, Fed-Dice: Dice coefficient of federated deep learning, Cent-Dice: Dice coefficient of centralized learning

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

(b) Different tumor intensities in FeTS

Subset	Region	EMD	BD	CSD	KSD	Fed-Dice	Cent-Dice	*theta*	*p*-value
SI-Q12, SI-Q34	NET *	6.2058	0.2452	0.3437	0.1700	0.7272	0.7626	0.9536±0.1365	0.04
	ET	6.8982	0.0965	0.1605	0.2100	0.8031	0.8432	0.9525±0.0662	0.06
SI-Q1, SI-Q34	NET	8.7306	0.4961	0.5563	0.1900	0.6554	0.6931	0.9456±0.1927	0.06
	ET *	10.1325	0.2137	0.2974	0.2800	0.7599	0.8117	0.9362±0.0351	0.01
SI-Q12, SI-Q4	NET *	11.3050	0.5575	0.5857	0.1900	0.7426	0.7838	0.9475±0.2004	0.04
	ET	9.5900	0.1687	0.2562	0.2000	0.7686	0.8059	0.9537±0.0498	0.09
SI-Q1, SI-Q4	NET *	13.8350	0.8664	0.7306	0.1800	0.5411	0.5947	0.9098±0.2279	0.02
	ET *	12.8243	0.3104	0.3929	0.3000	0.6965	0.7551	0.9224±0.1662	0.003

Note.— SI-Q(signal intensity quarter)1: non-enhancing < 25% and enhancing < 25%; SI-Q12: non-enhancing < 50% and enhancing < 50%; SI-Q34: non-enhancing ≥ 50% and enhancing ≥ 50%; SI-Q4: non-enhancing ≥ 75% and enhancing ≥ 75%. P-value calculated using paired t test; (*) denotes significantly different performances (p-values <0.05) between federated and centralized models. FeTS: Federated Tumor Segmentation, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test, Fed-Dice: Dice coefficient of federated deep learning, Cent-Dice: Dice coefficient of centralized learning

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, rad-ai@rsna.org

Figure Legends

Figure 1. (A) Selection criteria and (B) characteristics for the Federated Imaging in Liver Tumor Segmentation (FILTS) dataset. FNH = focal nodular hyperplasia, HCC = hepatocellular carcinoma, HEM = hemangioma, ICC = intrahepatic cholangiocarcinoma, LiTS = Liver Tumor Segmentation, ME = metastases.

Figure 2. (A-C) Example CT axial images of liver tumors at different sites and (D-F) histograms showing differences in CT attenuation distribution across the three sites.

Figure 3. (A-F) Examples of histogram between two different subsets of tumor in CT Liver Tumor Segmentation (FILTS) and MRI brain tumor segmentation (FeTS) datasets. HCC = hepatocellular carcinoma, HEM = hemangioma, FNH = focal nodular hyperplasia, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test.

Figure 4. CT attenuation distributions of different types of tumors at Site C in CT Liver Tumor Segmentation (FILTS) dataset. HCC = hepatocellular carcinoma, HEM = hemangioma, ME = metastases, FNH = focal nodular hyperplasia, ICC = intrahepatic cholangiocarcinoma.

Figure 5. Correlation coefficients and trendlines between distance metrics (A) EMD, (B) BD, (C) CSD, and (D) KSD and *theta* value. EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test.

Figure 6. Waterfall plots of (A) EMD, (B) BD, (C) CSD, and (D) KSD related to the performances of the federated models in 62 grouped subsets evaluations. EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test. Bar color indicates the p-values calculated using paired t test: Red: <0.05; Yellow: $0.05 \leq$ p-value < 0.10; Blue: ≥ 0.10.

Supplemental Material

4.1 Metrics of Data Distribution

Earth mover's distance (EMD) (or Wasserstein Distance) evaluates the dissimilarity between two data distributions that are represented by histograms [1]. Suppose $P = \{(p_i, u_i)\}_{i=1}^{m}$ and $Q = \{(q_j, v_j)\}_{j=1}^{n}$ are two histograms with sizes *m* and *n*, respectively, in which u_i (v_j) is the *i*th (*j*th) bin and p_i (q_j) is its weight. EMD is defined as the minimum work required to resolve the supply-demand transports:

$$\mathrm{EMD}(P,Q) = \min_{F=\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}}$$

subject to the following constraints:

$$\sum_j f_{ij} \le p_i, \sum_i f_{ij} \le q_j, \sum_{i,j} f_{ij} = \min\left\{\sum_i p_i, \sum_j q_j\right\}, f_{ij} \ge 0$$

where $F = \{f_{ij}\}$ denotes a set of flows, and each flow f_{ij} represents the amount transported from the *i*th bin to the *j*th bin. d_{ij} is the ground distance between the positions u_i and v_j .

For two data distributions *P* and *Q* over the same domain *X*, the Bhattacharyya distance (BD) [2] is defined as

$$\mathrm{BD}(P,Q) = -\ln\left(\mathrm{BC}(P,Q)\right)$$

where $\mathrm{BC}(P,Q) = \sum_{x \in X} \sqrt{p(x)q(x)}$ is the Bhattacharyya coefficient for discrete probability distributions.

The chi-square distance (CSD) between two distributions *P* and *Q* is defined by [3]

$$\mathrm{CSD}(P,Q) = \frac{1}{2} \sum_{x \in X} \frac{[p(x) - q(x)]^2}{p(x) + q(x)}$$

The two-sample Kolmogorov-Smirnov distance (KSD) is defined as the maximum absolute distance between their cumulative distribution functions (CDFs), which is obtained as [4]

$$\mathrm{KSD}(P,Q) = \max_x \left| F_P(x) - F_Q(x) \right|$$

where $F_P(x)$ and $F_Q(x)$ are the CDFs of the distributions P and Q, respectively.

4.2 Federated Deep Learning (Fed-DL) Models of Tumor Segmentation

We employed nnU-Net [5] for segmentation of liver tumors on CT and brain tumors on MRI in the study. Both federated and centralized models applied the same hyper parameters, and the same data augmentation techniques including rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring. Instead of using the original gradients averaging and updating during each mini-batch training, we adopted the Fed-Avg algorithm [6], a more efficient and common Fed-DL training strategy, in which weight averaging was performed during each epoch.

In the training stage, we applied mini-batch optimizer to train the model, and the network was trained for 300 epochs, with one epoch being defined by an iteration over 200 mini-batches. Samples in the mini-batches were chosen randomly from the training scans. Stochastic gradient descent with Nesterov momentum ($\mu = 0.99$) and an initial learning rate of 0.01 were used for learning network weights. The learning rate was decayed throughout the training by following the 'poly' learning rate policy of $(1 - \text{epoch}/\text{epochmax})^{0.9}$. We used the loss function that combines Dice loss with the standard binary cross-entropy (BCE) loss, which is generally the default for segmentation models. Also, the instance normalization was applied for each layer of the model.

In the inference stage, only testing data were used for inference. Segmentations were predicted with a sliding window approach, in which the window size equals the patch size used during training. Adjacent predictions overlap by half of the size of a patch. A Gaussian importance weighting was applied to reduce stitching artifacts and the influence of positions close to borders.

There are several open-source Fed-DL frameworks, such as TensorFlow Federated [7], PySyft [8], and Federated AI Technology Enabler [9]. These frameworks provided some general prototypes of server-client topology and communication mechanism. But they did not provide

some specific networks or optimization strategies that were required in medical image segmentation. Therefore, most of Fed-DL research was built on specific networks or learning strategies on their own. As the nnU-Net framework was too complicated to directly plunge into the existing federated framework, we implemented the federated averaging on the nnU-Net framework. The implemented server-client communication exactly followed the process described in [6].

The scripts of training commands of Fed-DL and centralized learning based on nnU-Net were like:

python FL_training.py 3d_fullres nnUNetTrainerV2 Task002_BrainTumor 1 site1 site2 --npz --use_compressed_data

python CL_training.py 3d_fullres nnUNetTrainerV2 Task002_BrainTumor 1 site1 site2 --npz --use_compressed_data

References for Supplemental Material

1 Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval. International journal of computer vision 2000; 40(2): 99-121.

2 Bhattacharyya, A. On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions. Bulletin of the Calcutta Mathematical Society 1943; 35: 99-109.

3 Pele O, Werman M. The quadratic-chi histogram distance family. European conference on computer vision 2010; 749-762.

4 Massey Jr F J. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association, 1951, 46(253): 68-78.

5 Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 2021; 18(2): 203-211.

6 McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. Artificial intelligence and statistics. PMLR 2017; 1273-1282.

7 TensorFlow Federated: Machine Learning on Decentralized Data. TensorFlow. URL: https://www.tensorflow.org/federated.

8 Ryffel T, Trask A, Dahl M, Wagner B, Mancuso J, Rueckert D. A generic framework for privacy preserving deep learning. arXiv preprint 2018:1811.04017.

9 An Industrial Grade Federated Learning Framework. FATE. URL: https://fate.fedai.org/.

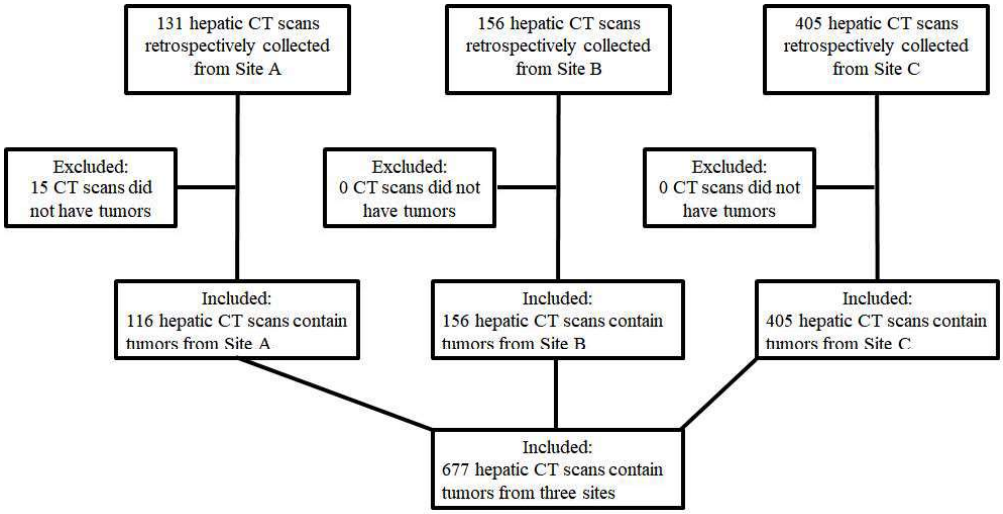

Figure 1. (a) Selection criteria.

Data Name	Data Source	Number of Cases	In-plane Resolution (mm)	Slice Thickness (mm)	Number of Tumors	Tumor Type
Site A	LiTS (publicly available)	131	0.55 – 1.0	0.45 – 6.0	734	HCC and metastases
Site B	A hospital in US	156	0.59 – 0.98	2.0 – 5.0	762	HCC
Site C	A hospital in China	405	0.52 – 0.96	0.8 – 5.0	585	HEM, FNH, Cyst, HCC, ICC, and ME

Age and sex were not listed as they were de-identified.

Figure 1.(b) characteristics for the Federated Imaging in Liver Tumor Segmentation (FILTS) dataset. FNH = focal nodular hyperplasia, HCC = hepatocellular carcinoma, HEM = hemangioma, ICC = intrahepatic cholangiocarcinoma, LiTS = Liver Tumor Segmentation, ME = metastases

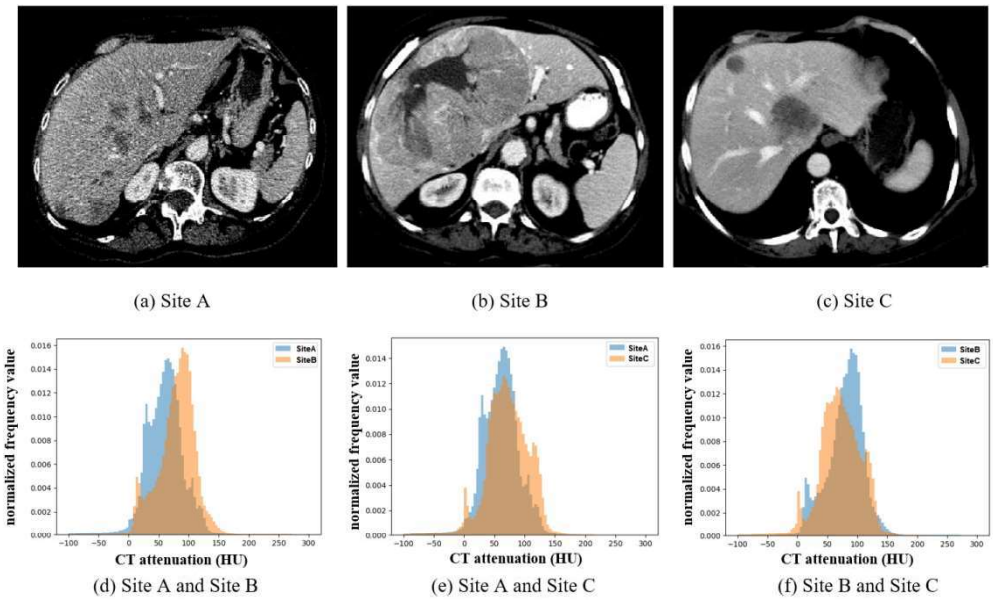

Figure 2. (A-C) Example CT axial images of liver tumors at different sites and (D-F) histograms showing differences in CT attenuation distribution across the three sites.

406x242mm (96 x 96 DPI)

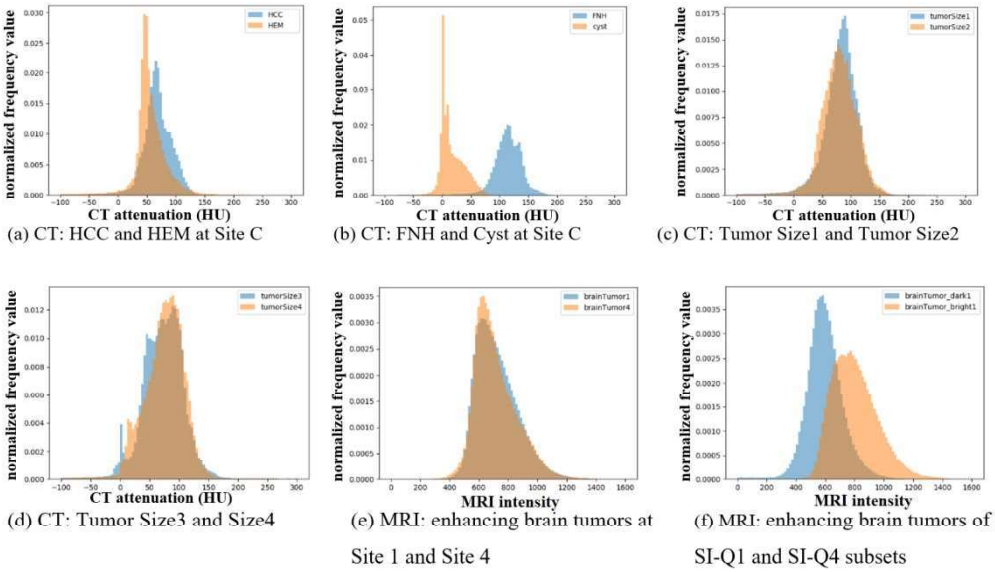

Figure 3. (A-F) Examples of histogram between two different subsets of tumor in CT Liver Tumor Segmentation (FILTS) and MRI brain tumor segmentation (FeTS) datasets. HCC = hepatocellular carcinoma, HEM = hemangioma, FNH = focal nodular hyperplasia, EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test.

437x247mm (96 x 96 DPI)

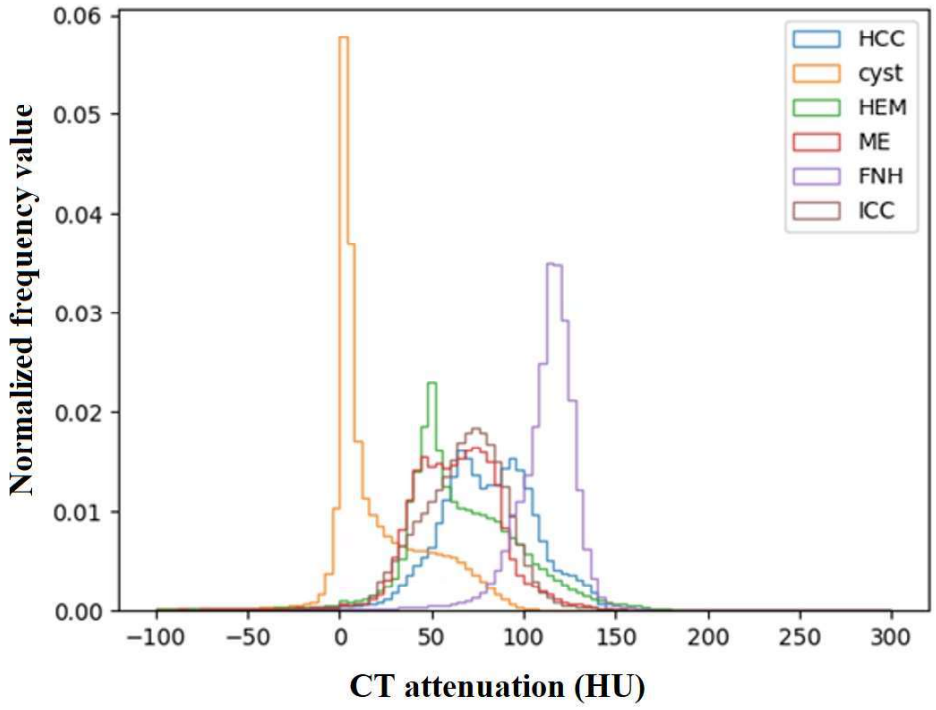

Figure 4. CT attenuation distributions of different types of tumors at Site C in CT Liver Tumor Segmentation (FILTS) dataset. HCC = hepatocellular carcinoma, HEM = hemangioma, ME = metastases, FNH = focal nodular hyperplasia, ICC = intrahepatic cholangiocarcinoma.

236x179mm (120 x 120 DPI)

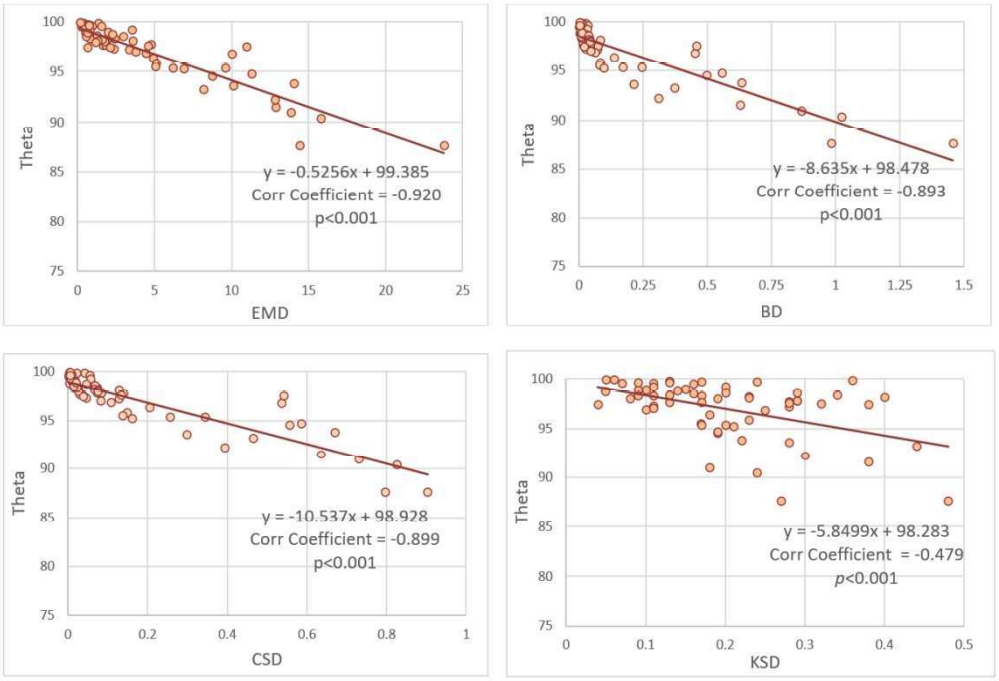

Figure 5. Correlation coefficients and trendlines between distance metrics (A) EMD, (B) BD, (C) CSD, and (D) KSD and theta value. EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test.

290x198mm (120 x 120 DPI)

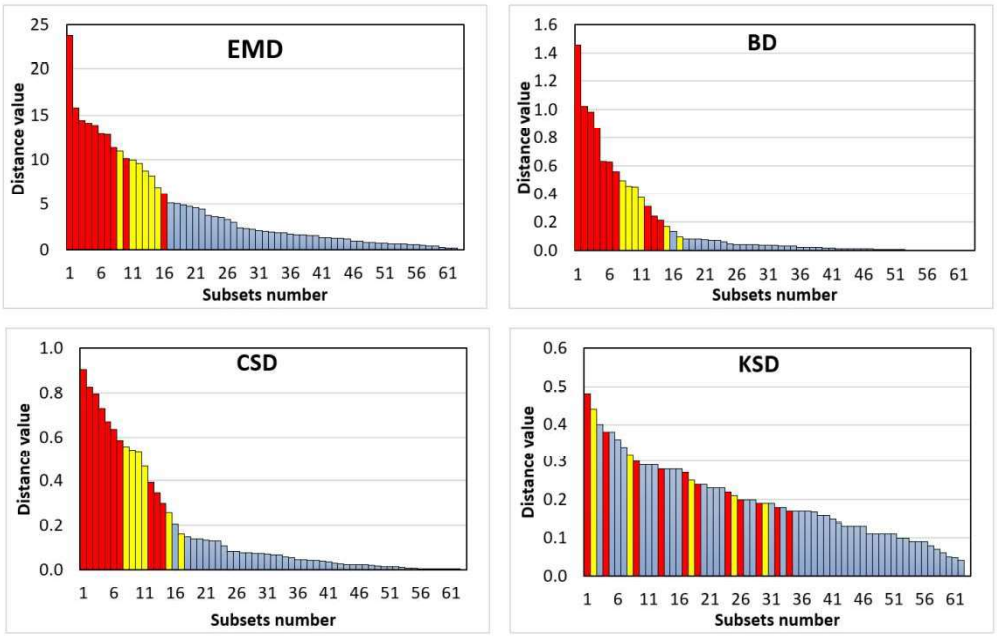

Figure 6. Waterfall plots of (A) EMD, (B) BD, (C) CSD, and (D) KSD related to the performances of the federated models in 62 grouped subsets evaluations. EMD: Earth mover's distance, BD: Bhattacharyya distance, CSD: Chi-square distance, KSD: D-statistic of Kolmogorov–Smirnov test. Color indicates the p-values calculated using paired t test: Red: <0.05; Yellow: 0.05 ≤ p-value < 0.10; Blue: ≥0.10.

384x245mm (96 x 96 DPI)