City Research Online

City, University of London Institutional Repository

Citation: Linardos, A., Pati, S., Baid, U., Edwards, B., Foley, P., Ta, K., Chung, V., Sheller, M., Khan, M. I., Jafaritadi, M., et al (2025). The MICCAI Federated Tumor Segmentation (FeTS) Challenge 2024: Efficient and Robust Aggregation Methods. Machine Learning for Biomedical Imaging, 3(December 2025), pp. 757-774. doi: 10.59275/j.melba.2025-5242

This is the published version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/36578/

Link to published version: https://doi.org/10.59275/j.melba.2025-5242

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

MELBA Journal Machine Learning for Biomedical Imaging

The MICCAI Federated Tumor Segmentation (FeTS) Challenge 2024: Efficient and Robust Aggregation Methods for Federated Learning

Akis **Linardos** [1,2], Sarthak **Pati** [1,2,3], Ujjwal **Baid** [1,2], Brandon **Edwards** [4], Patrick **Foley** [4], Kevin **Ta** [4], Verena **Chung** [5], Micah **Sheller** [3,4], Muhammad Irfan **Khan** [6], Mojtaba **Jafaritadi** [7], Elina **Kontio** [6], Suleiman **Khan** [6], Leon **Mächler** [8], Ivan **Ezhov** [9], Suprosanna **Shit** [9], Johannes C. **Paetzold** [10], Gustav **Grimberg** [11], Manuel A. **Nickel** [9], David **Naccache** [8], Vasilis **Siomos** [12], Jonathan **Passerat-Palmbach** [13], Giacomo **Tarroni** [12,13], Daewoon **Kim** [14], Leonard L. **Klausmann** [15], Prashant **Shah** [4], Bjoern **Menze** [16], Dimitrios **Makris** [17], Spyridon **Bakas** [1,2,3,17,18,19]

1 Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA
2 Center for Federated Learning in Medicine, Indiana University School of Medicine, Indianapolis, IN, USA
3 Medical AI Group, MLCommons, San Francisco, CA, USA
4 Intel Corporation, Santa Clara, CA, USA
5 Sage Bionetworks, Seattle, WA, USA
6 Turku University of Applied Sciences, Turku, Finland
7 Stanford University, Stanford, CA, USA
8 Ecole Normale Supérieure, Paris, France
9 Technical University of Munich, Munich, Germany
10 Weill Cornell Medicine, New York, USA
11 Ezri AI Labs, Paris, France
12 City St George's, University of London, UK
13 Imperial College London, London, UK
14 Seoul National University, Seoul, South Korea
15 Ostbayerische Technische Hochschule (OTH) Regensburg, Germany
16 Universität Zürich, Zürich, Switzerland
17 Kingston University London, London, UK
18 Departments of Radiology and Imaging Sciences; Neurological Surgery; Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA
19 Department of Computer Science, Luddy School of Informatics, Computing and Engineering, Indiana University, Indianapolis, IN, USA

arXiv:2512.06206v1 [cs.CV] 5 Dec 2025

Abstract

We present the design and results of the MICCAI Federated Tumor Segmentation (FeTS) Challenge 2024, focusing on federated learning (FL) for glioma sub-region segmentation in multi-parametric MRI scans, and evaluating novel weight aggregation methods for increased robustness and efficiency. Participating methods from six teams are evaluated using a standardized FL setup and a multi-institutional dataset derived from the BraTS glioma benchmark—a dataset consisting of 1,251 training cases, 219 validation cases, and 570 hidden test cases, with segmentations of enhancing tumor (ET), tumor core (TC), and whole tumor (WT). Teams are ranked by a cumulative scoring system accounting for segmentation performance—measured by Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95)—and communication efficiency assessed through the convergence score. A PID-controller-based approach emerges as the top-performing method, achieving a mean DSC of 0.733, 0.761, and 0.751 for ET, TC, and WT, respectively, with corresponding HD95 values of 33.922mm, 33.623mm, and 32.309mm, while also being the most efficient with a convergence score of 0.764. These results contribute to ongoing advances in FL, building on top-performers from previous challenge iterations and surpassing them, highlighting PID controllers as powerful mechanisms for stabilizing and optimizing weight aggregation in FL. The challenge code is available at `https://github.com/FeTS-AI/Challenge`.

Keywords

federated learning, biomedical challenge, segmentation, aggregation, brain tumor, glioma, glioblastoma

Article informations

`https://doi.org/10.59275/j.melba.2025-5242`
Received: 2025-4-25, Published 2025-12-4
Corresponding author: spbakas@iu.edu

©2025 Akis Linardos et al.. License: CC-BY 4.0
Check for updates

1. Background

While AI has been making strides in all fields, its applicability in healthcare has been mainly hindered by data scarcity, with most studies focusing on single-center data (Rajpurkar et al., 2022; Kelly et al., 2019), not able to capture the diversity across patient populations. Models produced in such a restricted manner have questionable generalizability in real-world applications, where data significantly varies from one site to the next. To address the scarcity, collaborative studies are essential, and to remain respectful of privacy constraints (such as HIPPA and GDPR (Annas et al., 2003; Voigt and Von dem Bussche, 2017)), a realizable way forward is through Federated Learning (FL): a framework that distributes models across sites, learns locally from institutional data, while alleviating the obvious privacy risks of data-sharing and hence acting as a catalyst for multi-site healthcare partnerships (McMahan et al., 2017; Yang et al., 2018, 2019; Rieke et al., 2020; Sheller et al., 2020; Yang et al., 2024; Pati et al., 2024). This way models may capture information from the high diversity of the real-world data, remaining fair across different populations.

It is important to note that while FL mitigates the obvious privacy risks by keeping data local, it does not guarantee privacy by itself (Pati et al., 2024; Zhao et al., 2025). Model updates can still leak sensitive information through, for example, membership inference attacks (Hu et al., 2022; Zhang et al., 2022). Complementary approaches to strengthen privacy in federated setups include secure aggregation (Fereidooni et al., 2021; So et al., 2022; Rathee et al., 2023), differential privacy (El Ouadrhiri and Abdelhadi, 2022; Adnan et al., 2022), homomorphic encryption (Xie et al., 2024; Aziz et al., 2023), and confidential computing[1].

For the work presented here, the use case is glioma segmentation, encompassing both low-grade and high-grade gliomas—including glioblastoma, the most common and aggressive type of adult brain tumor. Glioblastoma, despite multimodal treatments involving surgical resection, radiation, and chemotherapy, has a median survival of about 15 months, with less than 10% of patients surviving for over 5 years (Ostrom et al., 2015). The poor prognosis is largely a consequence of glioblastoma complexity, whose pathological heterogeneity leads to treatment resistance Bakas et al. (2024a); Villanueva-Meyer et al. (2024); Bakas et al. (2024b). Routine diagnosis and response assessment in glioblastoma patients is carried out through radiologic imaging (i.e., magnetic resonance imaging (MRI)) (Shukla et al., 2017), through which the tumor subregions may be delineated for follow-up computational analyses and personalized diagnostics (Pati et al., 2020). To enable robust tumor subregion delineation, the International Brain Tumor Segmentation (BraTS) benchmark/challenge (Menze et al., 2014; Bakas et al., 2018, 2017c,a,b; Baid et al., 2021b) has been at the forefront of providing high-quality data and an end-to-end open-source framework that fosters a benchmark environment for fair algorithmic evaluation. It used clinically acquired, multi-parametric MRI (mpMRI), and the evaluated algorithms are publicly available for use by the scientific community (Bakas et al., 2015; Zeng et al., 2016; Kamnitsas et al., 2017; Isensee et al., 2018; McKinley et al., 2018).

The Federated Tumor Segmentation (FeTS) challenge 2021 leveraged the BraTS glioma dataset to present the first challenge ever proposed for FL. The FeTS challenge in 2021 focused on constructing and evaluating a consensus model for the segmentation of gliomas, while its continuation in FeTS 2022 built on the foundation laid by its predecessor, further refining the federated learning techniques and expanding the collaborative network for an even larger dataset (Zenk et al., 2025). Since the first study of FL in healthcare (Sheller et al., 2019, 2020) and following the FeTS 2022 challenge, there have been numerous studies across medical imaging fields that focused on federated learning either in a simulation set up where the "different sites" are actually running on a single machine (Linardos et al., 2022; Ro et al., 2021; Li et al., 2022; Adnan et al., 2022) or real-world application, where the federated learning set up is deployed across actual sites, bringing together cohorts that span the globe (Pati et al., 2022b). Along the trajectory of this growth, there have also been multiple notable tools that foster FL research, such as the MedPerf for federated benchmarking of AI models 'in the wild' (Karargyris et al., 2023), fedJAX for simulation-focused research (Ro et al., 2021), and multiple libraries for FL development (Foley et al., 2022; Roth et al., 2022; Beutel et al., 2020; Ziller et al., 2021; Pati et al., 2022c).

Previous FeTS Challenges (2021 and 2022) Zenk et al. (2025) have presented tasks in two scenarios: one task in an environment that replicates federated learning conditions within one machine using BraTS, and a second task where evaluation is carried out across a real-world federation. Building on the insights from these previous challenges, the FeTS Challenge 2024, the third iteration of this challenge, shifts its focus exclusively to "optimal weight aggregation", testing innovations on the federated aggregation algorithm on a single machine, without the hurdle of real-world deployment. The primary goal of this challenge is to further refine the FL approach by optimizing the aggregation of model weights from different institutions, thereby improving the overall performance and robustness of the consensus models. This focus aims to address some of the remaining challenges in FL, particularly how best to combine the learned parameters from diverse data sources without compromising data privacy.

1. https://confidentialcomputing.io/

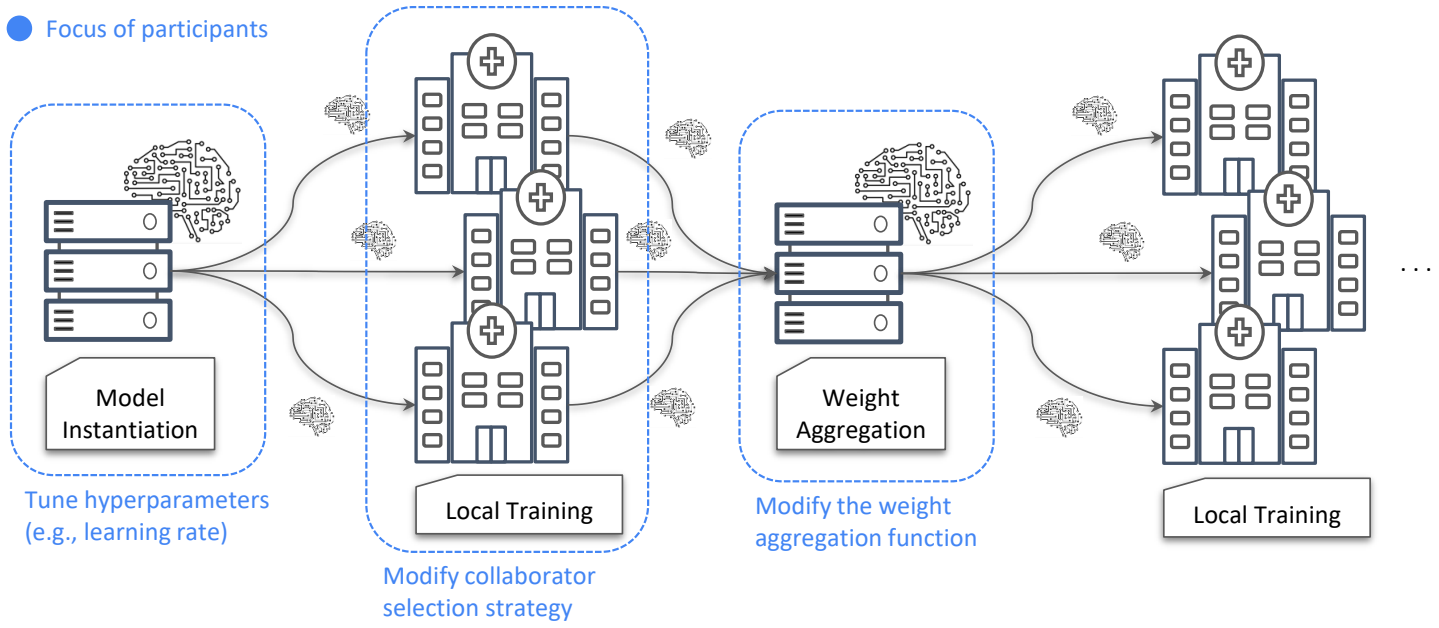

Figure 1: A schematic of the FL-based learning process. The FeTS Challenge tasks participants with contributing their innovations at the three levels: a. hyperparameter tuning, b. collaborator selection, c. weight aggregation function. Note that for this challenge, the individual sites illustrated here are actually data partitions within one computational environment rather than a real-world deployment.

2. Summary

The FeTS Challenge 2024 envisions FL as a transformative paradigm for multi-institutional collaboration, enabling robust, fair, and generalizable models without compromising patient privacy. The challenge aims to refine FL techniques to improve personalized diagnostics and treatment planning, emphasizing on glioma patients.

The task is focused on advancing aggregation methods for federated consensus models (Figure 1), providing tools that enable replicating an FL environment on a single machine. Using a curated multi-parametric MRI dataset from the BraTS glioma benchmark and clinically relevant segmentation metrics (Maier-Hein et al., 2024; Reinke et al., 2024), the challenge ensures real-world relevance while maintaining compliance with privacy regulations, like HIPAA and GDPR. It addresses critical challenges such as data scarcity and heterogeneity in healthcare AI while offering clear innovation opportunities in areas such as weight aggregation algorithms, hyperparameter tuning, and collaborator selection.

It is designed for researchers with programming expertise, through low-code tools GaNDLF (Pati et al., 2021) and openFL (Foley et al., 2022), but with clear "innovation hotspots" in the code, i.e., lines of the script where innovation is encouraged, including the aggregation algorithm itself, the hyper-parameter tuning, and the selection of collaborators per round. This way innovators/participants can seamlessly integrate new ideas in a code that is reproducible and has already been used successfully for three iterations of a challenge (2021, 2022, and now 2024) that attracted participations from research groups around the globe.

3. Resource Availability

3.1 Code Location

To enable reproducibility by the scientific community, the instructions and code used for the rankings have been made publicly available[2]. The GaNDLF framework, which acts as the backbone for AI training and development, is also available for public use[3] (Pati et al., 2021), as well as the backend responsible for the federated learning orchestration (OpenFL by Intel)[4](Reina et al., 2021; Foley et al., 2022).

3.2 Relevant Research

FeTS Challenge 2024 facilitates research on FL through benchmarking algorithms for optimal weight aggregation in FL setups, toward developing models that generalize across diverse clinical datasets without direct data sharing. It also trains research relevant robust AI models to segment glioma subregions in mpMRI data, but bears implications in broader clinical applications (e.g., radiology, cardiology, pathology). Most crucial is the potential of such refined FL techniques to facilitate multi-site collaborations in creating AI models for rare diseases or underserved populations where data availability is limited.

2. https://github.com/FETS-AI/Challenge
3. https://github.com/mlcommons/GaNDLF
4. https://github.com/securefederatedai/openfl

3.3 Licensing

The FeTS Challenge 2024 and its associated resources (FeTS, GaNDLF, and OpenFL) adhere to an Apache License[5]. This ensures that the resources are freely available for research, development, and deployment in various academic and clinical applications.

4. Materials & Methods

4.1 Data

This challenge leverages data from BraTS 2021 (Baid et al., 2021a; Bakas et al., 2017c, 2018), a multi-institutional dataset that has been evolving over the span of a decade, and has supported multiple challenges (Menze et al., 2014; Bakas et al., 2017a,b; Adewole et al., 2025; LaBella et al., 2024a; Mehta et al., 2022; LaBella et al., 2024b; Amiruddin et al., 2025; Kofler et al., 2025; Adewole et al., 2023; LaBella et al.; Maleki et al., 2025; LaBella et al., 2024c; de Verdier et al., 2024; Kazerooni et al., 2024; Moawad et al., 2024; Kofler et al., 2023; Li et al., 2024). BraTS 2021 contains mpMRI scans of glioma patients, routinely acquired during standard clinical practice along with their reference standard annotations for the evaluated tumor subregions. These are augmented with metadata that identify the partitioning of the scans in a de-identified manner. Each patient case contains four structural mpMRI scans at the pre-operative baseline timepoint: i) native T1-weighted (T1) and ii) contrast-enhanced T1 (T1-Gd), iii) T2-weighted (T2), and iv) T2 Fluid Attenuated Inversion Recovery (T2-FLAIR).

4.1.1 Data Pre-processing

The exact pre-processing pipeline applied to all the data considered in the present FeTS challenge is identical with the one evaluated and followed by the BraTS challenge and previous FeTS challenge iterations. Input scans (i.e., T1, T1-Gd, T2, T2-FLAIR) are registered to the same anatomical atlas (i.e., SRI-24 (Rohlfing et al., 2010)) using the Greedy diffeomorphic registration algorithm (Yushkevich et al., 2016), ensuring a common spatial resolution of $(1mm^3)$. After completion of the registration process, brain extraction is done to remove any apparent non-brain tissue, using a deep learning approach specifically designed for brain MRI scans with apparent diffuse glioma. This algorithm utilizes a novel training mechanism that introduces the brain's shape prior as knowledge to the segmentation algorithm (Thakur et al., 2020). All pre-processing routines have been made publicly available through the Cancer Imaging Phenomics Toolkit (CaPTk[6]) (Davatzikos et al., 2018; Pati et al., 2019; Rathore et al., 2017) and the FeTS tool (Pati et al., 2022b).

Table 1: Overview of case numbers in training, validation, and test sets. Datasets with source *BraTS'21* are centralized. *based on partitioning 1/2

	Training	Validation	Test
# cases	1251	219	570
# institutions	23/33*	10	12
Source	BraTS'21	BraTS'21	BraTS'21

4.1.2 Annotation Protocol

The skull-stripped scans are annotated to indicate the brain tumor subregions. The annotation process follows a predefined clinically approved annotation protocol that describes the detailed radiologic appearance of each tumor subregion of the MRI scans. In summary, the tumor subregions are:

1. the enhancing tumor (ET), which delineates the hyperintense signal of the T1-Gd, after excluding the vessels.
2. the necrotic tumor core (NCR), which outlines regions appearing dark in both T1 and T1-Gd images (denoting necrosis/cysts), and dark regions in T1-Gd and bright in T1.
3. the tumor core (TC), which includes the ET and NCR, and represents what is typically resected during a surgical operation.
4. the whole tumor (WT), which includes the peritumoral edematous and infiltrated tissue (ED), delineates the regions characterized by the hyperintense abnormal signal envelope on the T2-FLAIR sequence.

During its collection, BraTS followed a strict peer-review process for quality control, where each case was assigned to pairs of annotator-approvers. Annotator experience ranged across various levels of clinical / academic ranks, while approvers were the two experienced board-certified neuroradiologists (with $\geq$ 13 years of glioma experience). The annotators were given flexibility on which tool to use, and whether to follow a complete manual annotation approach, or a hybrid one with automated initial annotations followed by their manual refinements. Afterward, the produced annotations were passed to the corresponding approver, who evaluated them in tandem with the original mpMRI, and either signs them off or, in case of quality issues, returned them to the annotators for refinements. This iterative approach is followed for all cases until their respective annotations reaches satisfactory quality for publication, and noted as final reference standard segmentation labels.

5. http://www.apache.org/licenses/
6. https://www.cbica.upenn.edu/captk

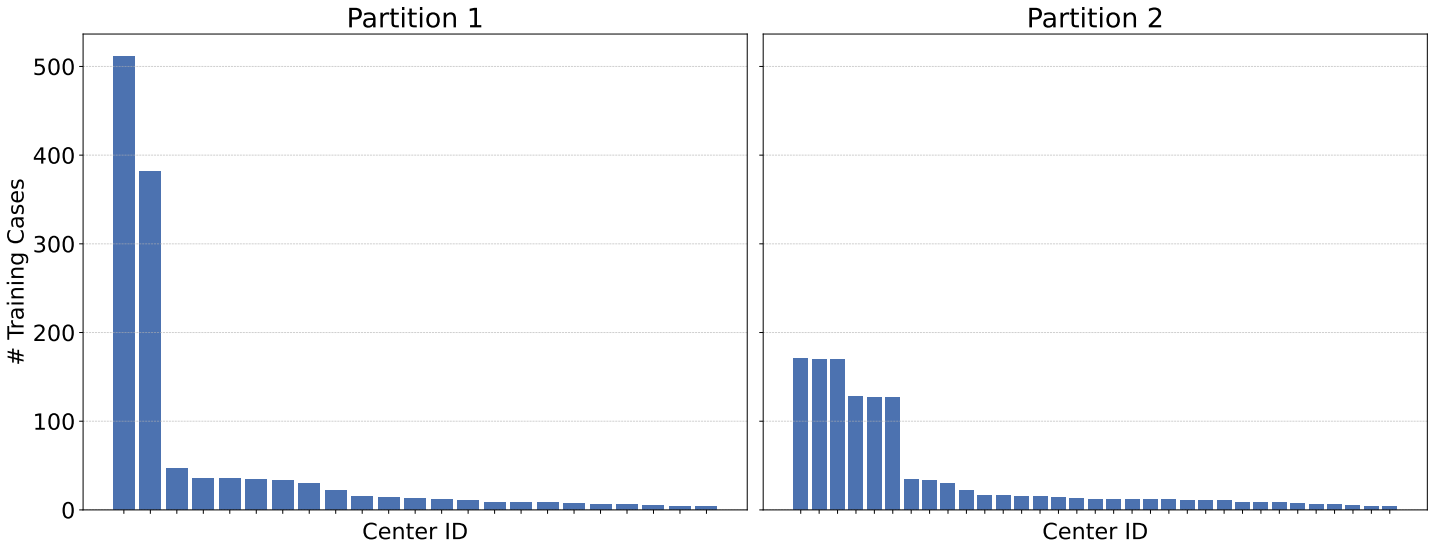

Figure 2: Official partitionings of the training set.

4.1.3 Training, Validation, and Test case characteristics

Of the original BraTS 2021 dataset, only the subset of radiographically visible glioma is included in the FeTS challenge, while cases without apparent enhancement are excluded. The exact numbers can be found in Table 1. Training cases encompass the mpMRI volumes, the corresponding tumor subregion annotations, as well as a pseudoidentifier of the site where the scans are acquired. Validation cases, however, only contain the mpMRI volumes, without any accompanying reference standard annotations or site pseudo-identifiers. We explicitly provide two schemas to partition the provided data (Figure 2):

1. Natural geographical partitioning by institution (partitioning 1, 23 sites)
2. Artificial partitioning using imaging information (partitioning 2, 33 sites), by further sub-dividing each of the 5 largest institutions in partition 1 into three parts after sorting samples by their whole tumor size.

80% of the dataset is kept for federated training (train and validation split), while the rest is set aside as a test set to evaluate the performance of submitted algorithms. The test set is never shared with the challenge participants.

4.2 Software Stack

4.2.1 Model Architecture

As the challenge focuses on the development of aggregation methods, the architecture of the segmentation model itself remains fixed across all participants. Following current literature and previous iterations of the challenge, the segmentation model chosen for this role is U-Net (Ronneberger et al., 2015) as the architecture has consistently performed well on medical imaging datasets (Isensee et al., 2018; Thakur et al., 2020; Drozdzal et al., 2016; He et al., 2016; Çiçek et al., 2016; Pati et al., 2021). The U-Net architecture is composed of an encoder-decoder architecture, where the encoder consists of layers performing convolutions and downsampling, while the decoder consists of layers performing transpose-convolution and upsampling. It includes skip connections in every convolution block—i.e., concatenated feature maps paired across the encoder and the decoder layer to improve context and feature re-usability, capturing information at multiple scales/resolutions.

4.2.2 Federated Training

An infrastructure for federated tumor segmentation has been provided to all participants indicating the spots on the code where they are expected to make changes ("innovation hotspots"). The primary objective is to develop methods for effective aggregation of local segmentation model weight updates, given the partitioning of the data into their real-world distribution.

The aggregation mechanism follows extensive prior literature (Sheller et al., 2018, 2020; Isensee et al., 2018; Pati et al., 2021), and is illustrated by Figure 1. Models are trained locally in each individual site and sent back to the aggregator at the end of each FL round. At the start of each round, each collaborator locally validates any model it receives from the central aggregation server (*aggregator*), then trains the model received from the *aggregator* on their local data. The local validation results along with the model updates are then returned to the *aggregator*, which combines model updates from all sites to produce a new consensus model. The consensus model is then passed back to each *collaborator*, and starts a new federated round.

5. Quantitative Performance Evaluation

Participants are called to produce segmentation labels of the different glioma subregions (ET, TC, WT). For each region, the predicted segmentation is compared with the reference standard segmentation using the following metrics:

5.1 Dice similarity coefficient (DSC)

The DSC is a metric commonly used to evaluate the performance of segmentation tasks. It measures the extent of spatial overlap of predicted masks (PM) and the provided reference standard (RS), while considering their union, thereby addressing both over-segmentation and under-segmentation. It is defined as

$$DSC = \frac{2|RS \cap PM|}{|RS| + |PM|}. \quad (1)$$

5.2 Hausdorff distance (HD)

This metric quantifies the distance between the boundaries of the reference standard labels against the predicted labels. It originates from set theory and measures the maximum distance of a point set to the nearest point in another set (Rockafellar and Wets, 2005). This makes the HD sensitive to local differences, as opposed to the DSC, which represents a global measure of overlap. For the specific problem of brain tumor segmentation, local differences are important for properly assessing the quality of the segmentation. In this challenge, the 95$^{\text{th}}$ percentile of the HD between the contours of the two segmentation masks is calculated, which is a variant of HD that is more robust to outlier pixels:

$$HD_{95}(PM, GT) = \max\Big\{ \underset{p \in PM}{P_{95\%}} d(p, GT), \underset{g \in GT}{P_{95\%}} d(g, PM) \Big\}, \quad (2)$$

where $d(x, Y) = \min_{y \in Y} ||x - y||$ is the distance of x to set Y.

5.3 Convergence Score

Convergence Score is used to estimate the efficiency of the model and encompasses the time taken to train and evaluate as well as the communication costs (download and upload of model weights between each collaborator and the central server).

We simulate the cumulative time taken per round, breaking it down to four components: training time T_{train}, validation time T_{val}, model weight download T_{down} and upload time T_{up}. In each round, the simulated time per collaborator k is

$$T_k = T_{\text{down},k} + T_{\text{up},k} + T_{\text{val},k} \cdot N_{\text{val},k} + T_{\text{train},k} \cdot N_{\text{train},k} \quad (3)$$

The total time for each round is $\max_k\{T_k\}$. To simulate a realistic FL setup, $T_{\text{x},k}$ was sampled from a normal distribution: $T_{\text{x},k} \sim \mathcal{N}(\mu_{\text{x},k}, \sigma_{\text{x},k})$, where x can be replaced with train/val/down/up. The parameters of the normal distribution are fixed but different for each client k, and based on time measurements derived from the largest to-date real-world FL study, which used the same model—the FeTS initiative (Pati et al., 2022a).

In each federated round we compute the mean DSC on a fixed split (20%) of the training data and the simulated round time T. Over the course of an experiment, this results in a DSC-over-time curve. A projected DSC curve is computed as $\text{DSC}_{\text{proj}}(t) = \max_{t' \leq t} \text{DSC}(\text{t}')$. The convergence score metric is calculated as the area under that projected DSC-over-time curve:

$$S_{\text{conv}} = \int_{t_0}^{T} \text{DSC}_{\text{proj}}(t)\, dt \quad (4)$$

where:

- S_{conv} is the convergence score. Higher values of this metric indicate enhanced convergence and thus a superior FL approach in terms of efficiency.
- $\text{DSC}_{\text{proj}}(t)$ is the projected DSC at time t.
- The integral runs from t_0 (starting time) to T (final time).

To standardize the time-axis for the convergence score among participants, all FL experiments performed during the challenge are limited to one week of simulated total time, which was a realistically feasible duration based on the experience from the FeTS initiative (Pati et al., 2022a). The FL runs were terminated once the simulated time exceeds one week and the model with the highest validation score before the last round is stored, to assure that a long last round exceeding the time limit does not benefit the participant.

5.4 Ranking Strategy

Before evaluating the submissions on the test set, algorithms are re-trained by the organizers, to ensure reproducible results and to prevent data leakage between federated sites. To standardize comparisons, the rankings are computed independently for each test case before aggregating them across the dataset, so performance is assessed fairly across different cases without being dominated by any single case. DSC scores, which measure the overlap between predicted and reference standard segmentations, are ranked in descending order since higher values indicate better segmentation performance. Conversely, Hausdorff distances, which quantify boundary errors, are ranked in ascending order as lower values correspond to more precise boundary delineations.

A tie-breaking strategy using rank averaging is applied to ensure consistent rankings when multiple teams achieve the same performance on a given test case. Specifically, if k teams have the same score for a particular metric, they all receive the average rank they would occupy if ranked distinctly. For example, if three teams tie for the second-best score, instead of being arbitrarily assigned ranks 2, 3, and 4, they are each assigned the average rank $(2+3+4)/3 = 3$. This method prevents unfair advantages or disadvantages due to arbitrary rank assignment and ensures a smooth aggregation of ranks across multiple cases. The overall ranking score for each team is then computed as the mean of its assigned ranks across all Dice and Hausdorff metrics, leading to a comprehensive performance measure.

Beyond segmentation quality, the framework integrates a communication efficiency score to refine the rankings. The communication metric, which reflects the efficiency of model convergence during training, is ranked separately using the same averaging-based ranking strategy. Since a higher communication score indicates better efficiency, rankings for this metric are assigned in descending order. This additional ranking is incorporated into the final ranking score using a predefined weight w, ensuring that teams that achieve competitive segmentation performance with lower communication costs are favored. The cumulative ranking score, incorporating segmentation and communication efficiency, is computed as the mean of all individual rankings (Dice, Hausdorff, and convergence score).

Mathematically, the final ranking score R_t for team t is given by:

$$R_t = \frac{1}{N+w} \sum_{m \in M} r_{t,m} + \frac{w}{N+w} r_{t,comm} \tag{5}$$

where:

- M is the set of segmentation metrics (Dice and Hausdorff) evaluated for each tumor subregion (ET, TC, WT).
- $N = |M|$ is the total number of segmentation rankings per test case, calculated as the number of metrics (2: Dice and Hausdorff) multiplied by the number of tumor subregions (3: ET, TC, WT), giving $N = 2 \times 3 = 6$,
- $r_{t,m}$ is the rank of team t for metric m,
- $r_{t,comm}$ is the rank of team t based on communication efficiency,
- w is the weight assigned to the communication metric.

This formulation ensures a balanced evaluation of segmentation accuracy and computational efficiency, where the influence of communication efficiency is modulated by the chosen weight w. When $w = 0$, rankings depend solely on segmentation performance, whereas increasing w gives more importance to communication efficiency.

6. Limitations

As a simulation-based setup, the FeTS 2024 challenge comes with limitations. While we approximate weight download and upload times in our convergence score metric, it does not fully reflect the communication heterogeneity across sites, which includes variable bandwidth, latency, dropped connections, and asynchronous updates. Second, differences in hardware availability (e.g., GPU memory, CPU load, and system failures) are abstracted away, but these factors strongly influence performance in real-world studies. Finally, security protocols, and the operational complexities of coordinating institutions are not represented here.

Despite these limitations, a simulated setup ensures reproducibility, fairness, and scalability for benchmarking innovations in aggregation methods. Such simulations are necessary prior to real-world deployments as they assess multiple promising aggregation methods without the expensive nature of actual large-scale coordination. They are essentially a preliminary step to future work that would bridge simulation and deployment, incorporating realistic communication and hardware variability into benchmarking protocols.

7. Participating Methods

A total of 6 submissions by 5 teams from four continents were submitted to FeTS Challenge 2024. The six proposed algorithms are outlined in table 2 and their innovations are as follows:

- **Federated Tick-Tack by ReMIC**: Fed Tick-Tack presents a novel approach to federated learning by introducing a two-phase aggregation technique designed to enhance model robustness and accuracy, especially in heterogeneous data environments. Instead of updating models at the end of each communication round, Fed Tick-Tack alternates between two distinct phases: Tick and Tack. In the Tick phase, an aggregated model is distributed to selected collaborators, who then locally train and return their models, weighted according to their individual importance (i.e. a weight assigned to each collaborator that reflects their model's influence on the overall aggregated model). The Tack phase focuses on updating these weights by calculating the differences between consecutive model proposals. This iterative adjustment ensures that the final model not only reflects the most recent learning but also adapts to the individual progress of each collaborator. Furthermore, Fed Tick-Tack supports both scalar and parameter-specific weight adjustments, offering flexibility in how collaborators' contributions are integrated. This method also introduces ranked batches to organize collaborators based on performance, toward balanced and efficient training rounds.
- **Clustered Approach with Bias-Variance Bal-**

Table 2: Teams and the corresponding papers of their submissions. Note that HTTUAS submitted 2 papers, proposing two distinct methods abbreviated in the text based on the technique they're deploying (Rec for Recommender Engine and RL for Reinforcement Learning).

Team	Method (Paper Title)
SNU	FedPOD: the deployable units of training for federated learning
rigg	FedPID, an aggregation method for Federated Learning (Mächler et al., 2024)
HTTUAS (Rec)	Recommender Engine for Client Selection in Federated Brain Tumor Segmentation (Khan et al., 2024b)
HTTUAS (RL)	Election of Collaborators via Reinforcement Learning for Federated Brain Tumor Segmentation (Khan et al., 2024a)
Flair	Adaptive Federated Learning for Brain Tumor Segmentation: A Clustered Approach with Bias-Variance Balancing
ReMIC	Federated Tick-Tack

ancing by Flair: Flair's approach integrates client selection, hyper-parameter tuning, and aggregation methods to enhance federated model training. The method involves clustering clients using k-means based on their training times, which minimizes idle times and ensures that clients with smaller datasets are effectively utilized. An adaptive strategy is used for hyper-parameter selection, including dynamic adjustment of local epochs and a modified cosine annealing schedule for learning rates. The aggregation method improves upon traditional FedAvg by incorporating Validation Loss Ratio (VLR) and an Overfit Penalty (OFP) to balance contributions based on validation performance and to address overfitting.

- **Recommender Engine for Client Selection by HTTUAS**: The study introduces a novel client selection protocol for Federated Learning in brain tumor segmentation, leveraging a recommender engine based on Non-Negative Matrix Factorization (NNMF) combined with a hybrid content-based and collaborative filtering approach. The NNMF decomposes historical performance metrics to identify suitable collaborators, while a fallback mechanism ensures continued operation in the absence of sufficient data. Additionally, this method presents Harmonic Similarity Weighted Aggregation (HSimAgg), an enhancement of the SimAgg (Khan et al., 2021) algorithm that uses harmonic mean aggregation to robustly handle outliers and extreme values, improving the accuracy and reliability of the federated model.

- **Election of Collaborators via Reinforcement Learning by HTTUAS** and similarity-weighted aggregation (SimAgg) (Khan et al., 2021) approach designed to optimize collaborator selection in federated brain tumor segmentation. The method employs multi-armed bandit algorithms, specifically Epsilon-greedy (EG) and Upper Confidence Bound (UCB), to manage the selection of collaborators and enhance model generalization. RL-HSimAgg balances exploration and exploitation to promote effective training across diverse datasets by dynamically choosing collaborators based on their performance. The approach also incorporates a similarity-weighted aggregation method to handle outliers by using harmonic mean, thereby improving robustness in FL environments.

- **FedPID by rigg**: Building on their previous participations FedCostWAvg and FedPIDAvg (Mächler et al., 2021, 2022), FedPID refines the aggregation strategy by incorporating improvements in how the integral term is computed. Unlike FedPIDAvg, which used a simple integration of the loss function, FedPID measures the global drop in cost since the first round. This method integrates a weighted averaging scheme combining dataset sizes, recent cost reductions, and global cost changes to update the model. Additionally, FedPID addresses varying dataset sizes by modeling them with a Poisson distribution, adjusting training iterations accordingly. This approach aims to enhance model performance by balancing local improvements with global progress and handling dataset size variability effectively.

- **FedPOD by SNU**: FedPOD also builds upon the foundations of FedPIDAvg (Mächler et al., 2022), to optimize both learning efficiency and communication costs in federated learning. FedPOD complements FedPIDAvg in two ways: (a) it includes outlier nodes that would otherwise be excluded and (b) eliminates the need for historical participant data. Due to these modules, FedPOD aims to better handle skewed data distributions and participant variability. Also, FedPOD is designed to work with Kubernetes' POD units, allowing for dynamic scaling of computational resources through Kubernetes' auto-scaling functionality.

8. Results

All teams are ranked according to Equation (5). For each of the 570 testing subjects, three tumor regions—ET, TC, and WT—are evaluated using two segmentation measures: the

Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD_{95}). This results in a total of $570 \times 3 \times 2 = 3,420$ individual rankings across all cases and metrics. As each of these metrics is accounted for three times (one for each modality), we also incorporate convergence score into the ranking process with a weighting factor of $w = 3$. This results in an adjusted total of $570 \times 3 \times 3 = 5,130$ rankings summed per team.

To provide an overview of segmentation performance, Table 3 presents the mean DSC and HD95 scores for each team. Higher DSC values indicate better segmentation accuracy, whereas lower HD95 values signify more precise boundary delineation.

FedPID and FedPOD methods presented by rigg and SNU respectively are top performers on DSC across all tumor types, indicating superior segmentation accuracy and consistency. The two methods also achieve competitive median Hausdorff Distance at 95th percentile (HD95) scores, demonstrating robustness in handling extreme boundary cases. The best DSC performance is achieved by SNU, with values of 0.733 (ET), 0.751 (WT), and 0.761 (TC). The best HD95 scores (i.e., lowest values) are more distributed: rigg achieves the lowest HD95 for ET (32.246 mm) and TC (31.705 mm), while HTTUAS (Rec) attains the lowest for WT (28.228 mm). These results demonstrate that no single method consistently outperforms across all metrics, highlighting trade-offs between segmentation accuracy and robustness.

Table 3: Mean DSC and HD95 for each team in the FeTS Challenge 2024. The highest DSC values (higher is better) and the lowest HD95 values (lower is better) are highlighted in bold.

Team	DSC (↑)			HD95 (↓)		
	ET	WT	TC	ET	WT	TC
SNU	**0.733**	0.751	**0.761**	33.922	32.309	33.623
rigg	0.722	**0.754**	0.748	**32.246**	31.122	**31.705**
HTTUAS (Rec)	0.682	0.738	0.716	34.023	**28.228**	32.911
HTTUAS (RL)	0.668	0.702	0.699	32.930	28.991	31.372
Flair	0.658	0.651	0.681	42.637	27.893	44.622
ReMIC	0.620	0.645	0.644	45.724	29.030	46.426

Visualizing the distribution of the performance, we also observe substantial variability in performance across different samples, with rigg and SNU exhibiting the lowest variance (Figure 3).

In terms of communication efficiency, FedPOD (SNU) also has the best convergence score **by a significant margin** (Figure 4). In terms of the other methods, HTTUAS (REC) provides a solid alternative with slightly higher DSC values than HTTUAS (RL) but with more variability in HD95, suggesting a trade-off between segmentation accuracy and robustness. Flair shows good DSC results but has higher variability and below average communication efficiency, reflecting that its dynamic hyper-parameter adjustments and clustering strategy might not be worth the computational overload.

The final ranking of each team is determined by summing all individual rankings and computing the cumulative ranking score as described in Equation (5). This approach ensures that both segmentation accuracy and computational efficiency are considered in a balanced manner. The winner of FeTS Challenge 2024 is the FedPOD method of SNU, while the FedPID method is also on the top 3 4. Even if we remove the communication efficiency from the ranking assessment, these two methods remain on the top 3 of the leaderboard. As both are based on the PID-controller, this showcases the effectiveness of this foundation for weight aggregation techniques.

9. Discussion

The MICCAI FeTS Challenge 2024 explored FL for glioma sub-region segmentation in brain mpMRI scans, focusing on innovations in weight aggregation. By analyzing the results of six distinct approaches, we identify trends and insights, highlighting both strengths and areas for improvement.

Two particular methods were of highest interest: FedPID and FedPOD, which achieved first and second ranking respectively if we only account for performance on the DSC and HD95, then third and first if we also account for communication efficiency (Table 4). Their relative variability on performance in the box plots was also the most similar in terms of robustness (Figure 3). The similarity in performance may be due to the fact that both these methods build on a predecessor from FeTS Challenge 2022 (Zenk et al., 2025), FedPIDAvg (Mächler et al., 2022), whose first iteration was FedCostWAvg (Mächler et al., 2021) presented in FeTS Challenge 2021. In both these challenges, the respective method achieved competitive ranking, and now two top performers are building on it further, suggesting this method presents a highly reliable baseline for innovation on the algorithm of FL.

Methods such as HTTUAS (Rec) and Flair illustrate the challenge of balancing segmentation accuracy with communication efficiency. HTTUAS (Rec) performed competitively in DSC while maintaining lower variability in HD95. Flair's adaptive strategies yielded moderate results but were hindered by higher computational demands. Communication efficiency was a critical metric in the final ranking, in which FedPOD outperformed all others by a significant margin, underscoring its relevance for real-world FL applications. In particular, it showcased what could be a critical innovation for FL efficiency: reframing the process as deployable Kubernetes units. This could be a highly valuable modification to

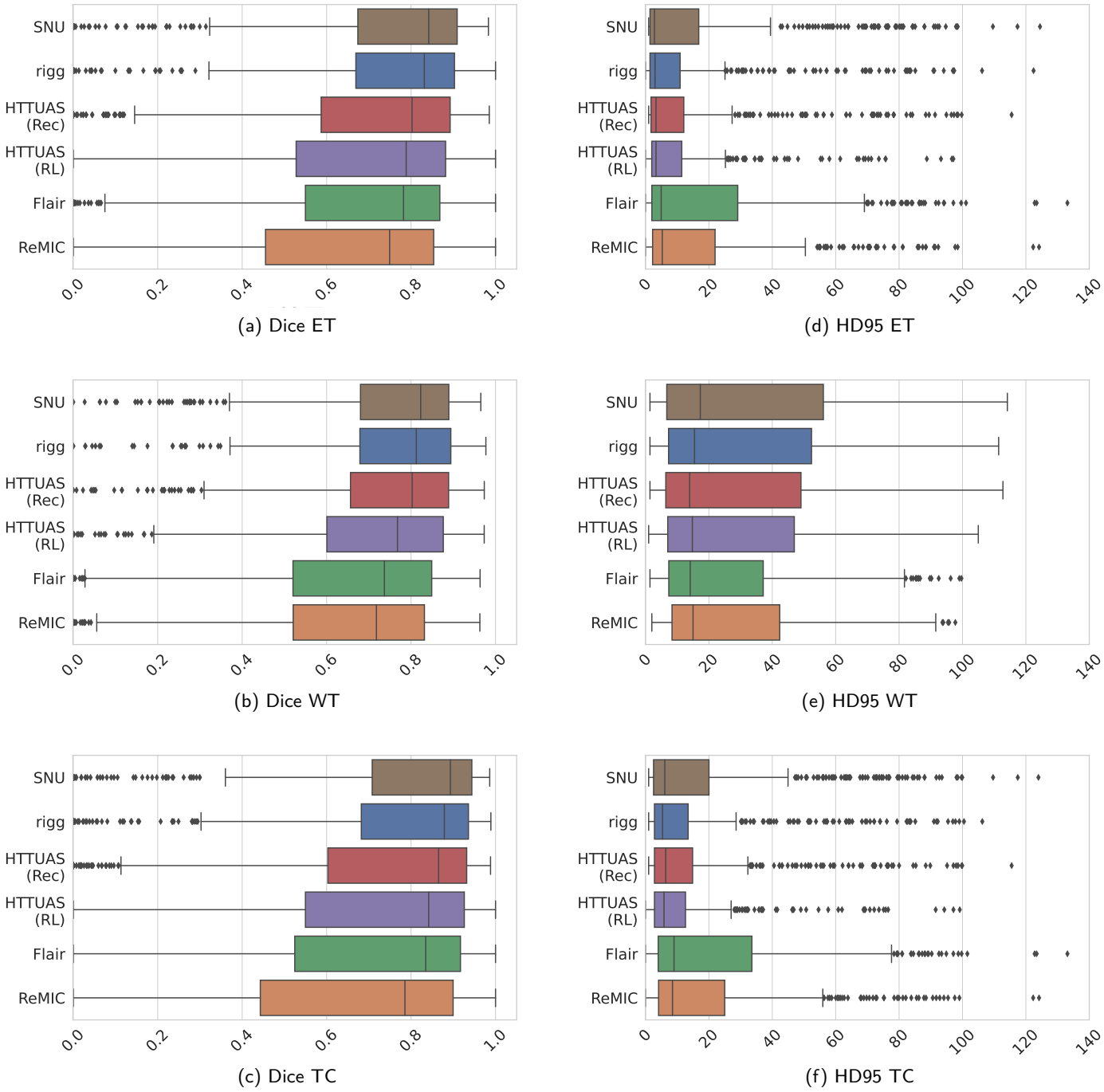

Figure 3: Box plots for Dice and HD95 metrics, illustrating the span of segmentation performance across different participating methods.

the algorithm for multi-site collaborations, especially under resource constraints.

The observed variability across aggregation strategies can be partly explained in medical imaging context: clinical datasets are inherently heterogeneous across sites not only in size but also in characteristics, reflecting differences in scanners, acquisition protocols, and patient populations, and outliers are not infrequent. On the one hand, HTTUAS methods are based on SimAgg assumes similarity of local models to global average, which given non-IID settings of medical imaging is often an incorrect assumption that does not acknowledge the presence of outliers. SimAgg-based methods thus favor high-performing collaborators and may boost overall segmentation accuracy but risk overlooking smaller institutions, raising questions of fairness. On the other hand, algorithms such as FedPID and FedPOD regulate updates across rounds using controller-inspired dynamics, which stabilizes learning under the non-uniform

Table 4: Final cumulative rankings of teams in the FeTS Challenge 2024. The table displays two rankings: A. one that includes convergence score—i.e. **Overall Ranking** ($w = 3$), accounting for computational efficiency. B. one that excludes convergence score—i.e. **Segmentation-Only** ($w = 0$), evaluating exclusively on segmentation quality.

Team	***Overall Ranking*** $(w = 3)$		***Segmentation-Only*** $(w = 0)$	
	Rank	**Score** (R_t)	**Rank**	**Score** (R_t)
SNU	**1**	**2.317**	2	2.976
HTTUAS (Rec)	2	2.744	3	3.116
rigg	3	3.206	**1**	**2.809**
HTTUAS (RL)	4	3.337	4	3.505
Flair	5	4.402	5	4.102
ReMIC	6	4.994	6	4.491

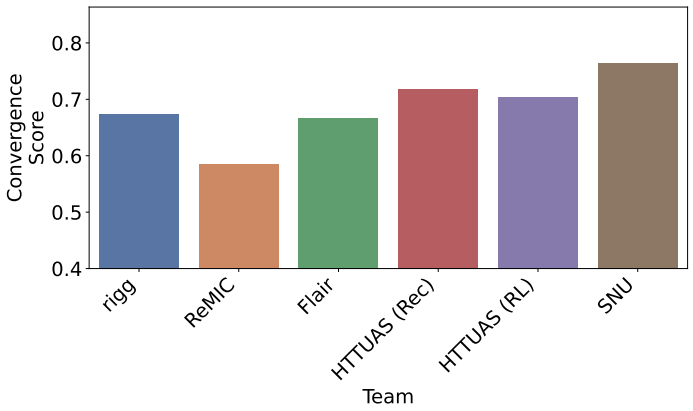

Figure 4: Convergence Score for each team. SNU's approach outperforms all others in this regard.

conditions of medical imaging, and effectively model outliers with a Poisson distribution. While FedPID and its predecessors exclude these outliers, FedPOD has devised a way to include them effectively in the training.

Future work may apply FedPOD and other PID-based methods beyond simulation-based evaluations to real-world deployments, and assess these highlighted benefits in settings of realistic device heterogeneity and communication. FedPOD's inclusion of outliers makes it an interesting candidate for fairness studies, while its compatibility of Kubernetes further enables its application on large-scale real world scenarios. This challenge, as its predecessors, highlights that FL algorithms are ripe for innovation, and sets a foundation for continued improvement.

Data availability

The imaging data used for the FeTS Challenge 2024 are derived from the publicly available BraTS 2021 dataset. Access to the dataset in the format used for this challenge can be obtained through The Cancer Imaging Archive: `https://www.cancerimagingarchive.net/analysis-result/rsna-asnr-miccai-brats-2021/`.

Ethical Standards

This work involves no new human or animal subject data collection. All imaging data used in the FeTS Challenge 2024 are from the publicly available BraTS 2021 dataset, which was de-identified and ethically approved under the respective Institutional Review Boards (IRBs) of contributing institutions. Data use complies with HIPAA and GDPR regulations. No additional IRB approval was necessary for this study. All data derivatives used in the challenge preserve and respect the original dataset's licensing and ethical constraints.

Ethical Standards

The FeTS Challenge 2024 adheres to rigorous ethical standards, ensuring compliance with data privacy and consent regulations. The challenge exclusively uses the BraTS 2021 dataset, which is publicly available and open source. Since no new patient data are collected or shared, the challenge does not involve privacy risks or concerns related to data protection. Furthermore, the challenge investigates FL, a paradigm that, when deployed in the real world, allows data to remain within originating institutions, thereby preserving patient privacy and complying with regulations such as HIPAA and GDPR.

All human data included in the challenge originate from the BraTS dataset, where Institutional Review Board (IRB) approval had been previously obtained. Data collection and processing adhered to protocols ensuring subjects' informed consent or opt-in mechanisms, as per institutional policies.

For derivative data used within the challenge, such as preprocessed MRI scans, compliance with the original dataset's licensing terms is maintained, ensuring ethical use and redistribution of data.

Acknowledgments

Research reported in this publication was partially supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under the award numbers U24CA189523 and U01CA242871. Computational resources used in this research were partially supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH or any other funding body.

Conflicts of Interest

The Intel-affiliated authors (B. Edwards, P. Foley, M. Sheller, and K. Ta) disclose potential competing interests as employees of Intel Corporation. Intel may develop proprietary software that could be perceived as related to the OpenFL open-source project that is the FL backbone of this work. Additionally, this study highlights the feasibility of federated learning for brain tumor segmentation, a domain where Intel could benefit from increased demand for relevant computational products. The remaining authors declare no competing interests.

References

Maruf Adewole, Jeffrey D Rudie, Anu Gbdamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, et al. The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa). *ArXiv*, pages arXiv–2305, 2023.

Maruf Adewole, Jeffrey D Rudie, Anu Gbadamosi, Dong Zhang, Confidence Raymond, James Ajigbotoshso, Oluyemisi Toyobo, Kenneth Aguh, Olubukola Omidiji, Rachel Akinola, et al. The brats-africa dataset: Expanding the brain tumor segmentation data to capture african populations. *Radiology: Artificial Intelligence*, 7 (4):e240528, 2025.

Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific reports*, 12(1):1953, 2022.

Raisa Amiruddin, Nikolay Y Yordanov, Nazanin Maleki, Pascal Fehringer, Athanasios Gkampenis, Anastasia Janas, Kiril Krantchev, Ahmed Moawad, Fabian Umeh, Salma Abosabie, et al. Training the next generation of physicians for artificial intelligence-assisted clinical neuroradiology: Asnr miccai brain tumor segmentation (brats) 2025 lighthouse challenge education platform. *arXiv preprint arXiv:2509.17281*, 2025.

George J Annas et al. Hipaa regulations-a new era of medical-record privacy? *New England Journal of Medicine*, 348(15):1486–1490, 2003.

Rezak Aziz, Soumya Banerjee, Samia Bouzefrane, and Thinh Le Vinh. Exploring homomorphic encryption and differential privacy techniques towards secure federated learning paradigm. *Future internet*, 15(9):310, 2023.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, Ahmed W. Moawad, Luiz Otavio Coelho, Olivia McDonnell, Elka Miller, Fanny E. Moron, Mark C. Oswood, Robert Y. Shih, Loizos Siakallis, Yulia Bronstein, James R. Mason, Anthony F. Miller, Gagandeep Choudhary, Aanchal Agarwal, Cristina H. Besada, Jamal J. Derakhshan, Mariana C. Diogo, Daniel D. Do-Dai, Luciano Farage, John L. Go, Mohiuddin Hadi, Virginia B. Hill, Michael Iv, David Joyner, Christie Lincoln, Eyal Lotan, Asako Miyakoshi, Mariana Sanchez-Montano, Jaya Nath, Xuan V. Nguyen, Manal Nicolas-Jilwan, Johanna Ortiz Jimenez, Kerem Ozturk, Bojan D. Petrovic, Chintan Shah, Lubdha M. Shah, Manas Sharma, Onur Simsek, Achint K. Singh, Salil Soman, Volodymyr Statsevych, Brent D. Weinberg, Robert J. Young, Ichiro Ikuta, Amit K. Agarwal, Sword C. Cambron, Richard Silbergleit, Alexandru Dusoi, Alida A. Postma, Laurent Letourneau-Guillon, Gloria J. Guzman Perez-Carrillo, Atin Saha, Neetu Soni, Greg Zaharchuk, Vahe M. Zohrabian, Yingming Chen, Milos M. Cekic, Akm Rahman, Juan E. Small, Varun Sethi, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R. Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-Andre Weber, Abhishek Mahajan, Bjoern Menze, Adam E. Flanders, and Spyridon Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021a.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021b.

S Bakas, H Akbari, A Sotiras, et al. Segmentation labels for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive, 2017a.

Spyridon Bakas, Ke Zeng, Aristeidis Sotiras, Saima Rathore, Hamed Akbari, Bilwaj Gaonkar, Martin Rozycki, Sarthak Pati, and Christos Davatzikos. Glistrboost: combining multimodal mri segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In *BrainLes 2015*, pages 144–155. Springer, 2015.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017b.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017c.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Spyridon Bakas, Siddhesh P Thakur, Shahriar Faghani, Mana Moassefi, Ujjwal Baid, Verena Chung, Sarthak Pati, Shubham Innani, Bhakti Baheti, Jake Albrecht, et al. Brats-path challenge: Assessing heterogeneous histopathologic brain tumor sub-regions. *arXiv preprint arXiv:2405.10871*, 2024a.

Spyridon Bakas, Philipp Vollmuth, Norbert Galldiks, Thomas C Booth, Hugo JWL Aerts, Wenya Linda Bi, Benedikt Wiestler, Pallavi Tiwari, Sarthak Pati, Ujjwal Baid, et al. Artificial intelligence for response assessment in neuro oncology (ai-rano), part 2: recommendations for standardisation, validation, and good clinical practice. *The Lancet Oncology*, 25(11):e589–e601, 2024b.

Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

Christos Davatzikos, Saima Rathore, Spyridon Bakas, Sarthak Pati, Mark Bergman, Ratheesh Kalarot, Patmaa Sridharan, Aimilia Gastounioti, Nariman Jahani, Eric Cohen, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of medical imaging*, 5(1):011018, 2018.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint arXiv:2405.18368*, 2024.

Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.

Ahmed El Ouadrhiri and Ahmed Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE access*, 10:22359–22380, 2022.

Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Helen Möllering, Thien Duc Nguyen, Phillip Rieger, Ahmad-Reza Sadeghi, Thomas Schneider, Hossein Yalame, et al. Safelearn: Secure aggregation for private federated learning. In *2021 IEEE security and privacy workshops (SPW)*, pages 56–62. IEEE, 2021.

Patrick Foley, Micah J Sheller, Brandon Edwards, Sarthak Pati, Walter Riviera, Mansi Sharma, Prakash Narayana Moorthy, Shih-han Wang, Jason Martin, Parsa Mirhaji, et al. Openfl: the open federated learning library. *Physics in Medicine & Biology*, 67(21):214001, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36: 61–78, 2017.

Alexandros Karargyris, Renato Umeton, Micah J Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, et al. Federated benchmarking of medical artificial intelligence with medperf. *Nature machine intelligence*, 5(7):799–810, 2023.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Deep Gandhi, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation in pediatrics (brats-peds) challenge: focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *arXiv preprint arXiv:2404.15009*, 2024.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.

Muhammad Irfan Khan, Mojtaba Jafaritadi, Esa Alhoniemi, Elina Kontio, and Suleiman A Khan. Adaptive weight aggregation in federated learning for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 455–469. Springer, 2021.

Muhammad Irfan Khan, Elina Kontio, Suleiman A Khan, and Mojtaba Jafaritadi. Election of collaborators via reinforcement learning for federated brain tumor segmentation. *arXiv preprint arXiv:2412.20253*, 2024a.

Muhammad Irfan Khan, Elina Kontio, Suleiman A Khan, and Mojtaba Jafaritadi. Recommender engine driven client selection in federated brain tumor segmentation. *arXiv preprint arXiv:2412.20250*, 2024b.

Florian Kofler, Felix Meissen, Felix Steinbauer, Robert Graf, Eva Oswald, Ezequiel de da Rosa, Hongwei Bran Li, Ujjwal Baid, Florian Hoelzl, Oezguen Turgut, et al. The brain tumor segmentation (brats) challenge 2023: Local synthesis of healthy brain tissue via inpainting. *arXiv preprint arXiv:2305.08992*, 2023.

Florian Kofler, Marcel Rosier, Mehdi Astaraki, Ujjwal Baid, Hendrik Möller, Josef A Buchner, Felix Steinbauer, Eva Oswald, Ezequiel de la Rosa, Ivan Ezhov, et al. Brats orchestrator: Democratizing and disseminating state-of-the-art brain tumor image analysis. *arXiv preprint arXiv:2506.13807*, 2025.

D LaBella, K Schumacher, M Mix, K Leu, S McBurney-Lin, P Nedelec, J Villanueva-Meyer, J Shapey, T Vercauteren, K Chia, et al. Brain tumor segmentation (brats) challenge 2024: Meningioma radiotherapy planning automated segmentation. arxiv 2024. *arXiv preprint arXiv:2405.18383*.

Dominic LaBella, Ujjwal Baid, Omaditya Khanna, Shan McBurney-Lin, Ryan McLean, Pierre Nedelec, Arif Rashid, Nourel Hoda Tahon, Talissa Altes, Radhika Bhalerao, et al. Analysis of the brats 2023 intracranial meningioma segmentation challenge. *arXiv preprint arXiv:2405.09787*, 2024a.

Dominic LaBella, Omaditya Khanna, Shan McBurney-Lin, Ryan Mclean, Pierre Nedelec, Arif S Rashid, Nourel Hoda Tahon, Talissa Altes, Ujjwal Baid, Radhika Bhalerao, et al. A multi-institutional meningioma mri dataset for automated multi-sequence image segmentation. *Scientific data*, 11(1):496, 2024b.

Dominic LaBella, Katherine Schumacher, Michael Mix, Kevin Leu, Shan McBurney-Lin, Pierre Nedelec, Javier Villanueva-Meyer, Jonathan Shapey, Tom Vercauteren, Kazumi Chia, et al. Brain tumor segmentation (brats) challenge 2024: Meningioma radiotherapy planning automated segmentation. *arXiv e-prints*, pages arXiv–2405, 2024c.

Hongwei Bran Li, Gian Marco Conte, Qingqiao Hu, Syed Muhammad Anwar, Florian Kofler, Ivan Ezhov, Koen van Leemput, Marie Piraud, Maria Diaz, Byrone Cole, et al. The brain tumor segmentation (brats) challenge 2023: Brain mr image synthesis for tumor segmentation (brasyn). *ArXiv*, pages arXiv–2305, 2024.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE, 2022.

Akis Linardos, Kaisar Kushibar, Sean Walsh, Polyxeni Gkontra, and Karim Lekadir. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Scientific Reports*, 12(1):3551, 2022.

Leon Mächler, Ivan Ezhov, Florian Kofler, Suprosanna Shit, Johannes C Paetzold, Timo Loehr, Claus Zimmer, Benedikt Wiestler, and Bjoern H Menze. Fedcostwavg:

a new averaging for better federated learning. In *International MICCAI Brainlesion Workshop*, pages 383–391. Springer, 2021.

Leon Mächler, Ivan Ezhov, Suprosanna Shit, and Johannes C Paetzold. Fedpidavg: A pid controller inspired aggregation method for federated learning. In *International MICCAI Brainlesion Workshop*, pages 209–217. Springer, 2022.

Leon Mächler, Gustav Grimberg, Ivan Ezhov, Manuel Nickel, Suprosanna Shit, David Naccache, and Johannes C Paetzold. Fedpid: An aggregation method for federated learning. *arXiv preprint arXiv:2411.02152*, 2024.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024.

Nazanin Maleki, Raisa Amiruddin, Ahmed W Moawad, Nikolay Yordanov, Athanasios Gkampenis, Pascal Fehringer, Fabian Umeh, Crystal Chukwurah, Fatima Memon, Bojan Petrovic, et al. Analysis of the miccai brain tumor segmentation–metastases (brats-mets) 2025 lighthouse challenge: Brain metastasis segmentation on pre-and post-treatment mri. *arXiv preprint arXiv:2504.12527*, 2025.

Richard McKinley, Raphael Meier, and Roland Wiest. Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 456–465. Springer, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Raghav Mehta, Angelos Filos, Ujjwal Baid, Chiharu Sako, Richard McKinley, Michael Rebsamen, Katrin Dätwyler, Raphael Meier, Piotr Radojewski, Gowtham Krishnan Murugesan, et al. Qu-brats: Miccai brats 2020 challenge on quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. *The journal of machine learning for biomedical imaging*, 2022: https–www, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Rachit Saluja, Nader Ashraf, Nazanin Maleki, Leon Jekel, Nikolay Yordanov, Pascal Fehringer, et al. The brain tumor segmentation-metastases (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri. *ArXiv*, pages arXiv–2306, 2024.

Quinn T Ostrom, Haley Gittleman, Jordonna Fulop, Max Liu, Rachel Blanda, Courtney Kromer, Yingli Wolinsky, Carol Kruchko, and Jill S Barnholtz-Sloan. Cbtrus statistical report: primary brain and central nervous system tumors diagnosed in the united states in 2008-2012. *Neuro-oncology*, 17(suppl_4):iv1–iv62, 2015.

Sarthak Pati, Ashish Singh, Saima Rathore, Aimilia Gastounioti, Mark Bergman, Phuc Ngo, Sung Min Ha, Dimitrios Bounias, James Minock, Grayson Murphy, et al. The cancer imaging phenomics toolkit (captk): Technical overview. In *International MICCAI Brainlesion Workshop*, pages 380–394. Springer, 2019.

Sarthak Pati, Ruchika Verma, Hamed Akbari, Michel Bilello, Virginia B Hill, Chiharu Sako, Ramon Correa, Niha Beig, Ludovic Venet, Siddhesh Thakur, et al. Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the ivy glioblastoma atlas project (ivy gap) dataset. *Medical Physics*, 2020.

Sarthak Pati, Siddhesh P Thakur, Megh Bhalerao, Ujjwal Baid, Caleb Grenko, Brandon Edwards, Micah Sheller, Jose Agraz, Bhakti Baheti, Vishnu Bashyam, et al. Gandlf: A generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. *arXiv preprint arXiv:2103.01006*, 2021.

Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G. Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, Chiharu Sako, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Philipp Vollmuth, Gianluca Brugnara, Chandrakanth J. Preetha, Felix Sahm, Klaus Maier-Hein, Maximilian Zenk, Martin Bendszus, Wolfgang Wick, Evan Calabrese, Jeffrey Rudie, Javier Villanueva-Meyer, Soonmee Cha, Madhura Ingalhalikar, Manali Jadhav, Umang Pandey, Jitender Saini, John Garrett, Matthew Larson, Robert Jeraj, Stuart Currie, Russell Frood, Kavi Fatania, Raymond Y. Huang, Ken Chang, Carmen Balaña Quintero, Jaume Capellades, Josep Puig, Johannes Trenkler, Josef Pichler, Georg Necker, Andreas Haunschmidt, Stephan Meckel, Gaurav Shukla, Spencer Liem, Gregory S. Alexander, Joseph Lombardo, Joshua D. Palmer, Adam E. Flanders, Adam P. Dicker, Haris I. Sair, Craig K. Jones, Archana Venkataraman, Meirui Jiang, Tiffany Y. So, Cheng Chen, Pheng Ann Heng, Qi Dou, Michal Kozubek, Filip Lux, Jan Michálek,

Petr Matula, Miloš Keřkovský, Tereza Kopřivová, Marek Dostál, Václav Vybíhal, Michael A. Vogelbaum, J. Ross Mitchell, Joaquim Farinhas, Joseph A. Maldjian, Chandan Ganesh Bangalore Yogananda, Marco C. Pinho, Divya Reddy, James Holcomb, Benjamin C. Wagner, Benjamin M. Ellingson, Timothy F. Cloughesy, Catalina Raymond, Talia Oughourlian, Akifumi Hagiwara, Chencai Wang, Minh-Son To, Sargam Bhardwaj, Chee Chong, Marc Agzarian, Alexandre Xavier Falcão, Samuel B. Martins, Bernardo C. A. Teixeira, Flávia Sprenger, David Menotti, Diego R. Lucio, Pamela LaMontagne, Daniel Marcus, Benedikt Wiestler, Florian Kofler, Ivan Ezhov, Marie Metz, Rajan Jain, Matthew Lee, Yvonne W. Lui, Richard McKinley, Johannes Slotboom, Piotr Radojewski, Raphael Meier, Roland Wiest, Derrick Murcia, Eric Fu, Rourke Haas, John Thompson, David Ryan Ormond, Chaitra Badve, Andrew E. Sloan, Vachan Vadmal, Kristin Waite, Rivka R. Colen, Linmin Pei, Murat Ak, Ashok Srinivasan, J. Rajiv Bapuraj, Arvind Rao, Nicholas Wang, Ota Yoshiaki, Toshio Moritani, Sevcan Turk, Joonsang Lee, Snehal Prabhudesai, Fanny Morón, Jacob Mandel, Konstantinos Kamnitsas, Ben Glocker, Luke V. M. Dixon, Matthew Williams, Peter Zampakis, Vasileios Panagiotopoulos, Panagiotis Tsiganos, Sotiris Alexiou, Ilias Haliassos, Evangelia I. Zacharaki, Konstantinos Moustakas, Christina Kalogeropoulou, Dimitrios M. Kardamakis, Yoon Seong Choi, Seung-Koo Lee, Jong Hee Chang, Sung Soo Ahn, Bing Luo, Laila Poisson, Ning Wen, Pallavi Tiwari, Ruchika Verma, Rohan Bareja, Ipsa Yadav, Jonathan Chen, Neeraj Kumar, Marion Smits, Sebastian R. van der Voort, Ahmed Alafandi, Fatih Incekara, Maarten M. J. Wijnenga, Georgios Kapsas, Renske Gahrmann, Joost W. Schouten, Hendrikus J. Dubbink, Arnaud J. P. E. Vincent, Martin J. van den Bent, Pim J. French, Stefan Klein, Yading Yuan, Sonam Sharma, Tzu-Chi Tseng, Saba Adabi, Simone P. Niclou, Olivier Keunen, Ann-Christin Hau, Martin Vallières, David Fortin, Martin Lepage, Bennett Landman, Karthik Ramadass, Kaiwen Xu, Silky Chotai, Lola B. Chambless, Akshitkumar Mistry, Reid C. Thompson, Yuriy Gusev, Krithika Bhuvaneshwar, Anousheh Sayah, Camelia Bencheqroun, Anas Belouali, Subha Madhavan, Thomas C. Booth, Alysha Chelliah, Marc Modat, Haris Shuaib, Carmen Dragos, Aly Abayazeed, Kenneth Kolodziej, Michael Hill, Ahmed Abbassy, Shady Gamal, Mahmoud Mekhaimar, Mohamed Qayati, Mauricio Reyes, Ji Eun Park, Jihye Yun, Ho Sung Kim, Abhishek Mahajan, Mark Muzi, Sean Benson, Regina G. H. Beets-Tan, Jonas Teuwen, Alejandro Herrera-Trujillo, Maria Trujillo, William Escobar, Ana Abello, Jose Bernal, Jhon Gómez, Joseph Choi, Stephen Baek, Yusung Kim, Heba Ismael, Bryan Allen, John M. Buatti, Aikaterini Kotrotsou, Hongwei Li, Tobias Weiss, Michael Weller, Andrea Bink, Bertrand Pouymayou, Hassan F. Shaykh, Joel Saltz, Prateek Prasanna, Sampurna Shrestha, Kartik M. Mani, David Payne, Tahsin Kurc, Enrique Pelaez, Heydy Franco-Maldonado, Francis Loayza, Sebastian Quevedo, Pamela Guevara, Esteban Torche, Cristobal Mendoza, Franco Vera, Elvis Ríos, Eduardo López, Sergio A. Velastin, Godwin Ogbole, Mayowa Soneye, Dotun Oyekunle, Olubunmi Odafe-Oyibotha, Babatunde Osobu, Mustapha Shu'aibu, Adeleye Dorcas, Farouk Dako, Amber L. Simpson, Mohammad Hamghalam, Jacob J. Peoples, Ricky Hu, Anh Tran, Danielle Cutler, Fabio Y. Moraes, Michael A. Boss, James Gimpel, Deepak Kattil Veettil, Kendall Schmidt, Brian Bialecki, Sailaja Marella, Cynthia Price, Lisa Cimino, Charles Apgar, Prashant Shah, Bjoern Menze, Jill S. Barnholtz-Sloan, Jason Martin, and Spyridon Bakas. Federated learning enables big data for rare cancer boundary detection. *Nature Communications*, 13(1):7346, December 2022a. ISSN 2041-1723. . URL `https://www.nature.com/articles/s41467-022-33407-5`. Number: 1 Publisher: Nature Publishing Group.

Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):7346, 2022b.

Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah J Sheller, Patrick Foley, G Anthony Reina, Siddhesh Thakur, Chiharu Sako, Michel Bilello, Christos Davatzikos, et al. The federated tumor segmentation (fets) tool: an open-source solution to further solid tumor research. *Physics in Medicine & Biology*, 67(20):204002, 2022c.

Sarthak Pati, Sourav Kumar, Amokh Varma, Brandon Edwards, Charles Lu, Liangqiong Qu, Justin J Wang, Anantharaman Lakshminarayanan, Shih-han Wang, Micah J Sheller, et al. Privacy preservation for federated learning in health care. *Patterns*, 5(7), 2024.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28 (1):31–38, 2022.

Mayank Rathee, Conghao Shen, Sameer Wagh, and Raluca Ada Popa. Elsa: Secure aggregation for federated learning with malicious actors. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1961–1979. IEEE, 2023.

Saima Rathore, Spyridon Bakas, Sarthak Pati, Hamed Akbari, Ratheesh Kalarot, Patmaa Sridharan, Martin Rozycki, Mark Bergman, Birkan Tunc, Ragini Verma, et al.

Brain cancer imaging phenomics toolkit (brain-captk): an interactive platform for quantitative analysis of glioblastoma. In *International MICCAI Brainlesion Workshop*, pages 133–145. Springer, 2017.

G Anthony Reina, Alexey Gruzdev, Patrick Foley, Olga Perepelkina, Mansi Sharma, Igor Davidyuk, Ilya Trushkin, Maksim Radionov, Aleksandr Mokrov, Dmitry Agapov, et al. Openfl: An open-source framework for federated learning. *arXiv preprint arXiv:2105.06413*, 2021.

Annika Reinke, Minu D Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A Emre Kavur, Tim Rädsch, Carole H Sudre, Laura Acion, Michela Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature methods*, 21 (2):182–194, 2024.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Jae Hun Ro, Ananda Theertha Suresh, and Ke Wu. Fedjax: Federated learning simulation with jax. *arXiv preprint arXiv:2108.02117*, 2021.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2005.

Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31 (5):798–819, 2010.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Holger R Roth, Yan Cheng, Yuhong Wen, Isaac Yang, Ziyue Xu, Yuan-Ting Hsieh, Kristopher Kersten, Ahmed Harouni, Can Zhao, Kevin Lu, et al. Nvidia flare: Federated learning from simulation to real-world. *arXiv preprint arXiv:2210.13291*, 2022.

Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 92–104. Springer, 2018.

Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 92–104. Springer, 2019.

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020.

Gaurav Shukla, Gregory S Alexander, Spyridon Bakas, Rahul Nikam, Kiran Talekar, Joshua D Palmer, and Wenyin Shi. Advanced magnetic resonance imaging in glioblastoma: a review. *Chinese clinical oncology*, 6(4):40–40, 2017.

Jinhyun So, Chaoyang He, Chien-Sheng Yang, Songze Li, Qian Yu, Ramy E Ali, Basak Guler, and Salman Avestimehr. Lightsecagg: a lightweight and versatile design for secure aggregation in federated learning. *Proceedings of Machine Learning and Systems*, 4:694–720, 2022.

Siddhesh Thakur, Jimit Doshi, Sarthak Pati, Saima Rathore, Chiharu Sako, Michel Bilello, Sung Min Ha, Gaurav Shukla, Adam Flanders, Aikaterini Kotrotsou, et al. Brain extraction on mri scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *NeuroImage*, 220:117081, 2020.

Javier E Villanueva-Meyer, Spyridon Bakas, Pallavi Tiwari, Janine M Lupo, Evan Calabrese, Christos Davatzikos, Wenya Linda Bi, Marwa Ismail, Hamed Akbari, Philipp Lohmann, et al. Artificial intelligence for response assessment in neuro oncology (ai-rano), part 1: review of current advancements. *The Lancet Oncology*, 25(11): e581–e588, 2024.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10:3152676, 2017.

Qipeng Xie, Siyang Jiang, Linshan Jiang, Yongzhi Huang, Zhihe Zhao, Salabat Khan, Wangchen Dai, Zhe Liu, and Kaishun Wu. Efficiency optimization techniques in privacy-preserving federated learning with homomorphic encryption: A brief survey. *IEEE Internet of Things Journal*, 11(14):24569–24580, 2024.

Guang Yang, Brandon Edwards, Spyridon Bakas, Qi Dou, Daguang Xu, Xiaoxiao Li, and Wanying Wang. Federated learning as a catalyst for digital healthcare innovations. *Patterns*, 5(7), 2024.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Paul A Yushkevich, John Pluta, Hongzhi Wang, Laura EM Wisse, Sandhitsu Das, and David Wolk. Fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 tesla and 7 tesla t2-weighted mri. *Alzheimer's & Dementia*, 7(12):P126–P127, 2016.

Ke Zeng, Spyridon Bakas, Aristeidis Sotiras, Hamed Akbari, Martin Rozycki, Saima Rathore, Sarthak Pati, and Christos Davatzikos. Segmentation of gliomas in pre-operative and post-operative multimodal magnetic resonance imaging volumes based on a hybrid generative-discriminative framework. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 184–194. Springer, 2016.

Maximilian Zenk, Ujjwal Baid, Sarthak Pati, Akis Linardos, Brandon Edwards, Micah Sheller, Patrick Foley, Alejandro Aristizabal, David Zimmerer, Alexey Gruzdev, et al. Towards fair decentralized benchmarking of healthcare ai algorithms with the federated tumor segmentation (fets) challenge. *Nature communications*, 16(1):6274, 2025.

Rui Zhang, Song Guo, Junxiao Wang, Xin Xie, and Dacheng Tao. A survey on gradient inversion: Attacks, defenses and future directions. *arXiv preprint arXiv:2206.07284*, 2022.

Joshua Zhao, Saurabh Bagchi, Salman Avestimehr, Kevin Chan, Somali Chaterji, Dimitris Dimitriadis, Jiacheng Li, Ninghui Li, Arash Nourian, and Holger Roth. The federation strikes back: A survey of federated learning privacy attacks, defenses, applications, and policy landscape. *ACM Computing Surveys*, 57(9):1–37, 2025.

Alexander Ziller, Andrew Trask, Antonio Lopardo, Benjamin Szymkow, Bobby Wagner, Emma Bluemke, Jean-Mickael Nounahon, Jonathan Passerat-Palmbach, Kritika Prakash, Nick Rose, et al. Pysyft: A library for easy federated learning. *Federated Learning Systems: Towards Next-Generation AI*, pages 111–139, 2021.