







# Federated Learning Using Variable Local Training for Brain Tumor Segmentation

Anup Tuladhar<sup>1,2</sup> , Lakshay Tyagi<sup>3</sup> , Raissa Souza<sup>1,2,4</sup> ,  
and Nils D. Forkert<sup>1,2,5,6</sup> 

<sup>1</sup> Department of Radiology, Cumming School of Medicine,  
University of Calgary, Calgary, AB, Canada  
anup.tuladhar@ucalgary.ca

<sup>2</sup> Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

<sup>3</sup> Department of Chemical Engineering, Indian Institute of Technology Kanpur, Kanpur,  
Uttar Pradesh, India

<sup>4</sup> Biomedical Engineering Program, University of Calgary, Calgary, AB, Canada

<sup>5</sup> Department of Clinical Neurosciences, Cumming School of Medicine, University of Calgary,  
Calgary, AB, Canada

<sup>6</sup> Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada

**Abstract.** The potential for deep learning to improve medical image analysis is often stymied by the difficulty in acquiring and collecting sufficient data to train models. One major barrier to data acquisition is the private and sensitive nature of the data in question, as concerns about patient privacy, among others, make data sharing between institutions difficult. Distributed learning avoids the need to share data centrally by training models locally. One approach to distributed learning is federated learning, where models are trained in parallel at local institutions and aggregated together into a global model. The 2021 Federated Tumor Segmentation (FeTS) challenge focuses on federated learning for brain tumor segmentation using magnetic resonance imaging scans collected from a real-world federation of collaborating institutions. We developed a federated training algorithm that uses a combination of variable local epochs in each federated round, a decaying learning rate, and an ensemble weight aggregation function. When testing on unseen validation data our model trained with federated learning achieves very similar performance (average DSC score of 0.674) to a central model trained on pooled data (average DSC score 0.685). When our federated learning algorithm was evaluated on unseen training and testing data, it achieved similar performances on the FeTS challenge leaderboards 1 and 2 (average DSC scores of 0.623 and 0.608, respectively). This federated learning algorithm offers an approach to training deep learning models without the need to share private and sensitive patient data.

**Keywords:** Distributed learning · Federated learning · Convolutional neural network · Medical image analysis · Brain · Tumor segmentation · MICCAI · FeTS · BraTS

# 1 Introduction

Deep neural networks, and convolutional neural networks (CNNs) in particular, have shown promising results for various classification and segmentation tasks in medical imaging [1]. To achieve these high accuracies, deep learning models often need to be trained on a large quantity and variety of training data, *i.e.*, patient data in the case of medical imaging. The standard approach for training deep learning models is centralized learning, where models are trained on a single repository of collected data. However, this centralized training approach is difficult to implement in healthcare, as there is a lack of large clinical databases for model training and optimization. Although large amounts of imaging data are collected daily, concerns about patient privacy, data security, and legal and administrative hurdles make data collection into a single repository both difficult and unfavorable [2–4]. Thus, alternatives to centralized training of deep learning models are needed to create accurate and robust models for healthcare and medical imaging applications, especially in case of rare diseases.

Distributed learning is an alternative to centralized learning that circumvents the need for data sharing and a centralized repository. In distributed learning, model training occurs locally at each site, *e.g.* healthcare center. After local training, the models from each collaborating healthcare institution are sent back to a central server that combines these local models into a single global model. This global model is effectively trained on a larger amount and variety of patient data to ideally achieve accuracies greater than any locally trained model and similar performance compared to a model trained with central learning. Meanwhile, the patient data never leaves the local institution, which retains control over how the data is used. Thus, distributed learning is a promising approach to train deep learning models on sensitive and private patient data.

Federated learning is one approach to the distributed learning problem that aggregates model parameters, most often the gradient updates or model weights, from parallelly trained local models into a global model [6–8]. This global model then makes an inference on new data. This approach is in contrast to distributed learning by ensembling [5], which has local models make independent inferences on data and aggregates their predictions, and distributed learning using a traveling model [9, 10], which trains a single model at one site at a time.

The inaugural Federated Tumor Segmentation (FeTS) [11] challenge at the 2021 International Conference on Medical Image Computing & Computer Assisted Intervention (MICCAI) is the first challenge for federated learning in medical imaging [12]. The overall aim of the challenge is the construction and evaluation of CNN models for brain tumor segmentation using multi-institutional clinical datasets whilst, importantly, avoiding data sharing and data pooling. The 2021 FeTS challenge is composed of two tasks:

1. Task 1, Federated Training, is focused on methods for federated weight aggregation from locally-trained models using a pre-defined segmentation algorithm.
2. Task 2, Federated Evaluation, is focused on developing segmentation algorithms that generalize to unseen samples and collaborating institutions not involved in training.

In this paper, we focus on Task 1, Federated Training. We implemented federated learning using variable local training, where the number of epochs per federated training round and learning rate vary over the course of federated training. Additionally, we investigated multiple weight aggregation functions and implemented an ensemble aggregation function. We compared our federated learning model to a model trained using central learning, where all the data is pooled together and trained at a single location.

## 2 Methods

### 2.1 Challenge Dataset

The 2021 FeTS challenge uses clinically acquired multi-institutional, multi-parametric magnetic resonance imaging (MRI) images from the 2020 RSNA-MICCAI Brain Tumor Segmentation (BraTS) challenge [13–16]. The challenge dataset is comprised of 341 pre-processed patient brain imaging samples and accompanying manual segmentations from 17 contributing institutions. Each patient sample consisted of four MRI modalities: native T1-weighted, post-contrast T1-weighted, T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR) volumes. The scans were acquired across multiple sites with different protocols and scanners, and were provided preprocessed for this challenge. Manual segmentations of the tumor tissue were performed by four annotators using a standardized annotation protocol and were verified by expert neuro-radiologists [15, 16]. Three tissue class segmentations were available for each sample: contrast enhanced tumor (ET), tumor core (TC), and whole tumor (WT).

The challenge dataset was partitioned for federated learning based on their source institution. As institutions vary significantly on their size and patient population, the data partitioning is heterogenous and reflective of a real-world situation. Two partitions of the data were specified by the organizers:

1. Partition 1 is the real data distribution from the 17 participating institutions, with a median institution size of 8 samples, minimum size of 3 samples and maximum size of 129 samples (Fig. 1A).
2. Partition 2 is an artificial data distribution based on partition 1 that splits the five largest institutions based on median whole tumor size across all institutions. The median institution size of Partition 2 is 11 samples, the minimum institution size is 3 samples and the maximum institution size is 65 samples (Fig. 1B).

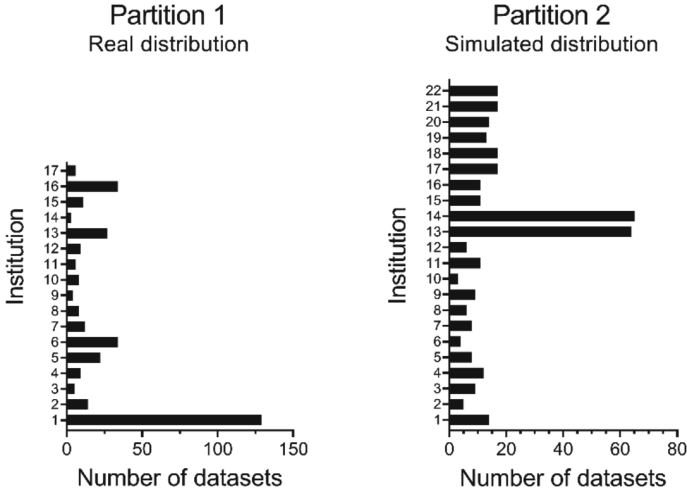


Fig. 1. Challenge data distribution for data partition 1, the real institution distribution, and partition 2, the simulated institution distribution.

For model training and validation, the 341 challenge samples were randomly separated into a training set (80%) and validation set (20%) at an institutional level. In other words, at each institution, 80% of the data was assigned to the training set and 20% of the data was assigned to the validation set.

### 2.2 Testing on Unseen Data

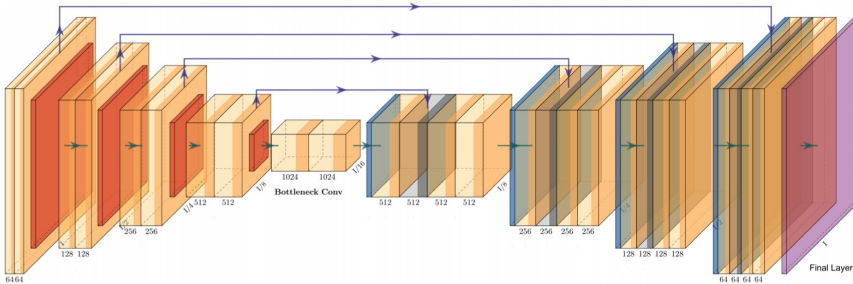
The 2021 FeTS challenge had two types of testing datasets. The first is the “Challenge Validation Set”, which we will refer to as the visible test set, which provides imaging data but not the ground truth segmentations. The inferences made on the visible test set are evaluated by the organizers through the online evaluation platform, which compares the submitted model segmentations to the withheld ground truth segmentations. As per the challenge guidelines, participants must train the model on partition 2 before making inferences on the visible test set in order to obtain valid results.

The second is the “Challenge Test Set”, which we will refer to as the hidden test set in the following. Both, the imaging data and segmentations on a set of unseen data, are withheld from participants. Submitted training algorithms for Task 1 will be trained on a hidden training partition, also withheld from participants, and the resultant model will be evaluated on the hidden test set. Both, training and testing for the “Challenge Test Set”, were done by challenge organizers, who then provided results to participants.

Due to a complication in the 2021 FeTS challenge, two evaluations were done on the “Challenge Test Set”, referred to as “Leaderboards”. The first evaluation, “Leaderboard 1”, used challenge code before the initial submission deadline. The second evaluation, “Leaderboard 2”, used an updated version of challenge code after an error was found in the “Leaderboard 1” challenge code. Our submission described in this paper was developed for “Leaderboard 1” and used “as-is” for “Leaderboard 2”.

### 2.3 Segmentation CNN

As Task 1 of the 2021 FeTS is focused on methods for federated training and weight aggregation methods from locally trained models, the architecture was pre-specified. More precisely, a 3D U-Net with residual connections was used [12] (Fig. 2).



**Fig. 2.** The segmentation U-Net architecture specified for Task 1. Figure taken from [12].

### 2.4 Federated Training

The challenge uses the central parameter server approach to federated training, as opposed to decentralized federated learning. In the central parameter server approach, a centralized aggregation server controls federated training rounds and performs local model aggregation. Briefly stated, at the beginning of each federated round, the server distributes the current model weights to all participating institutions for local training. After local training, contributing institutions send the updated model to the server, which combines the model weights together to produce an updated model, completing one federated round.

In each federated training, we trained local models in parallel at all institutions. We investigated modifications to the learning rate, epochs in a federated round, and methods for aggregating the weights from local models.

### 2.5 Evaluation

The trained models are evaluated on their segmentation performance using the Dice similarity coefficient (DSC) and 95<sup>th</sup> percentile Hausdorff distance (95%HD). The DSC measures the overlap between segmentations, *e.g.* model segmentations and human annotations, where 0 indicates no overlap and 1 indicates perfect overlap. The Hausdorff distance measures the distance between two point sets, *i.e.* edges of the model segmentation and edges of human annotation edges; the 95%HD uses the 95<sup>th</sup> percentile of Hausdorff distances, reducing the impact of segmentations with noisy edges. The DSC and 95%HD are calculated for each type of segmentation (ET, TC, and WT). Additionally, the metrics are also reported for the average of the three segmentations, *e.g.*  $DSC_{AVG} = 1/3(DSC_{WT} + DSC_{TC} + DSC_{ET})$ .

We report results primarily on split 2, the artificial data distribution, as this was the data partition designated by the challenge to be used for evaluation on the visible testing set (“Challenge Validation”).

We also report results from “Challenge Test Set” provided to us by challenge organizers. These are results from our final algorithm trained on a hidden training data partition and tested on a hidden test set. We report results for both “Leaderboard 1” and “Leaderboard 2”.

### 3 Results and Discussion

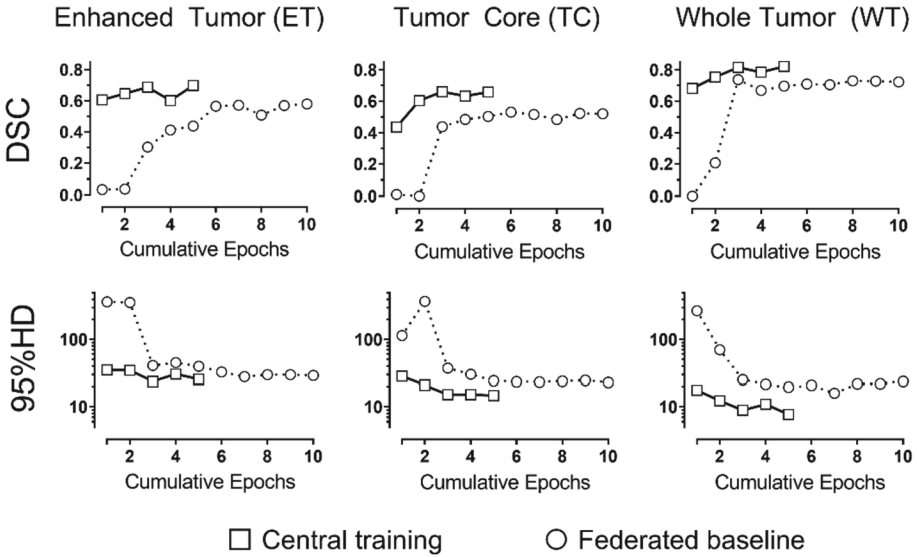
We first established baseline federated training performance using a constant learning rate ( $5e-3$ ) over 10 federated rounds, 1 epoch of local training per federated round, and federated averaging using the mean of weights of locally trained weights (Fig. 3). This was compared against central learning, where all training samples were pooled into one institution. The central model was trained with the same learning rate ( $5e-3$ ) for 5 epochs (Fig. 3). Overall, our baseline federated model performed worse than the central model, with a lower DSC ( $DSC_{AVG} = 0.607$  for federated baseline vs.  $DSC_{AVG} = 0.726$  for central training) and higher 95%HD ( $95\%HD_{AVG} = 25.17$  for federated baseline vs.  $95\%HD_{AVG} = 15.95$  for central training). Furthermore, federated learning required more cumulative epochs than central training, as the initial rounds of federated training showed poor performance due to the low amounts of data at local institutions.

We investigated changes in learning rate, epochs per round, and weight functions aggregation independently, varying the respective values for one hyper-parameter (*e.g.* learning rate) while keeping the others (*e.g.* epochs per round and weight aggregation) at baseline values.

#### 3.1 Learning Rate

First, we investigated learning rate decay using three types of functions: linear decay, exponential decay, and polynomial decay (Fig. 4A). For all three functions, the initial learning rate was  $5e-3$  and the final learning rate at the end of federated training was  $1e-6$ . Local models were trained for 1 epoch per round for 10 rounds, and the mean weight aggregation function was used.

All models trained with decaying learning rates performed better than the constant learning rate used in the federated baseline ( $5e-3$ ). Models trained with federate learning performed poorly when learning rates were higher in the initial training rounds (*i.e.* in the linear and polynomial decay functions) (Fig. 4). While linear and polynomial decay had similar 95%HD scores, linear decay resulted in overall better DSC. Thus, our final model used a linear decay function.



**Fig. 3.** Baseline federated learning model showed overall poorer performance on the validation data against a central learning model trained on all available at a single site.

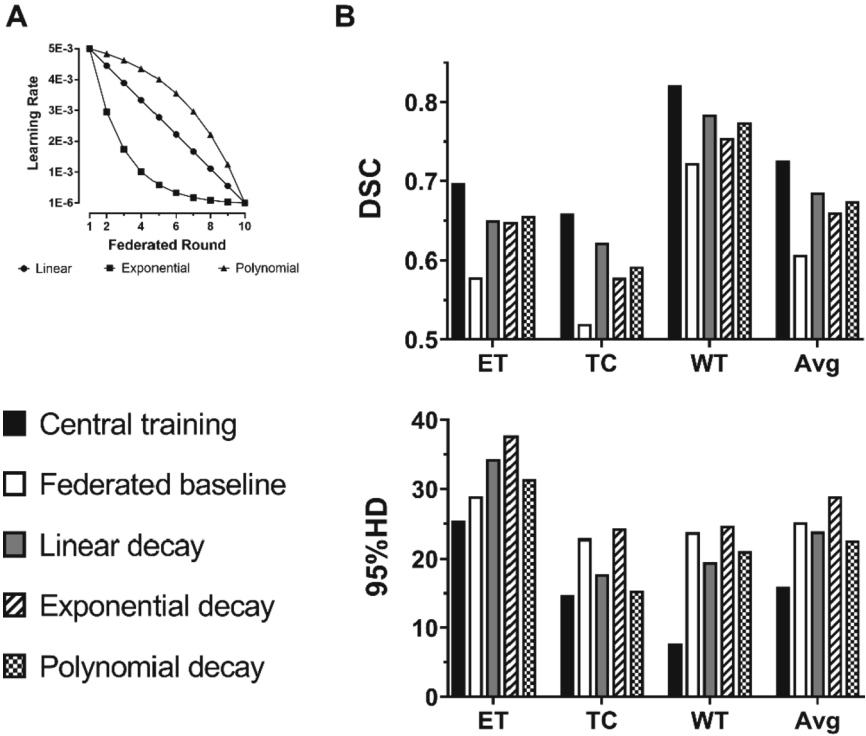
### 3.2 Epochs Per Round

Next, we investigated both constant and variable epochs per round (EpR) throughout the course of federated training. For constant EpR, we used 0.5, 1, and 2 as parameter values. To control for the total number of cumulative epochs across all federated training, we fixed the total epochs to 10 and varied the number of rounds (*i.e.* 20 rounds for 0.5 EpR, 10 rounds for 1 EpR, and 5 rounds for 2 EpR). The learning rate was set to  $5e-3$  and the mean weight aggregation function was used.

Overall, the 1 EpR parameter used in the baseline model had the best performance on the validation data ( $DSC_{AVG} = 0.607$ ;  $95\%HD_{AVG} = 25.17$ ). Although 0.5 EpR and 2 EpR had better 95%HD scores ( $95\%HD_{AVG} = 18.23$  for 0.5 EpR;  $95\%HD_{AVG} = 23.97$ ), the DSC was noticeably worse ( $DSC_{AVG} = 0.578$  for 0.5 EpR;  $DSC_{AVG} = 0.523$  for 2 EpR).

For variable EpR, we fixed the total number of rounds to 10, and varied the number of epochs in each round. We investigated increasing the number of EpR and increasing and decreasing the number of EpR in a pyramid-like shape (low numbers of EpR at the initial and final rounds, and high numbers of EpR in the middle rounds) (Fig. 5A). Increasing the number of EpR over the course of federated training resulted in substantial degraded performance, with lower DSC and higher 95%HD scores (Fig. 5B). This is likely due to overfitting to local model data at the latter stages of training due to the high number of epochs on small local datasets.

The pyramid-like changes in EpR, with high amounts of local rounds in the middle of federated training and low amounts of local rounds in the beginning and end of federated training (Fig. 5A) resulted in improved DSC scores compared to the federated baseline model. This was more noticeable in the “big pyramid”, which had a maximum number



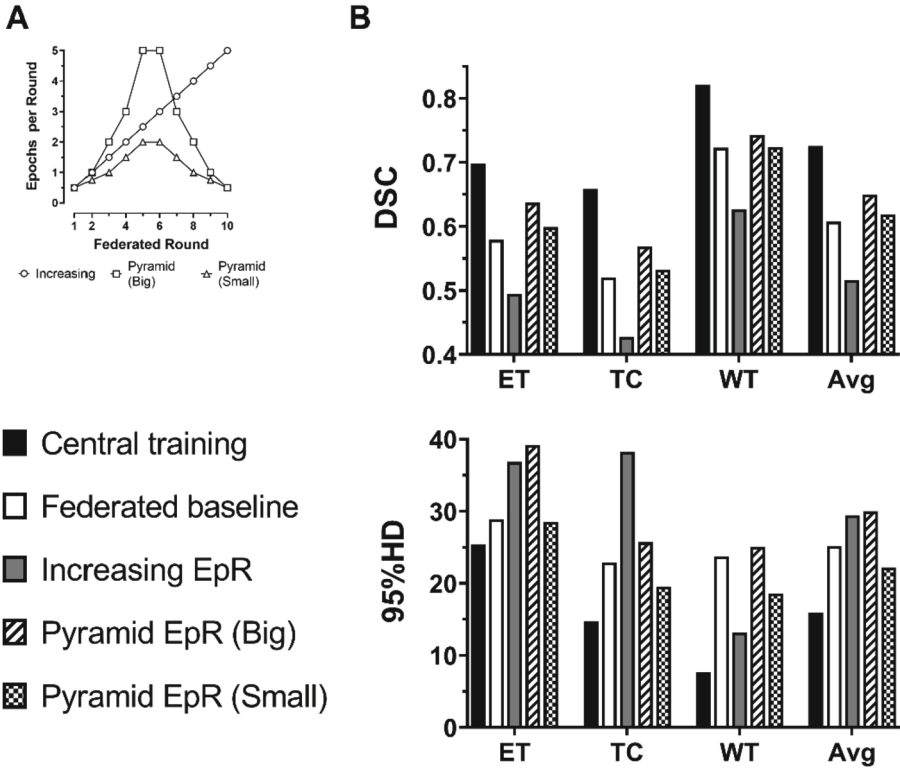
**Fig. 4.** Federated learning using decaying learning rates. (A) Three types of learning rate decay functions were investigated. (B) DSC and 95%HD scores for validation data on federated learning with decaying learning rates were compared against the central training and federated baseline models.

of local epochs of 5, compared to the “small pyramid”, which had a maximum number of local epochs of 2. The 95%HD was best with the “small pyramid” variable EpR. However, given the substantial improvement in DSC score, our final model used the “large pyramid” variable EpR scheme.

### 3.3 Weight Aggregation Function

Finally, we investigated different functions for aggregating weights. The general approach used to aggregate weights was federated averaging, where matching weights from locally trained models with identical architectures are averaged together across collaborating institutions. The baseline weight averaging function used was an arithmetic mean. We investigated three additional averaging functions: weighted arithmetic mean, median, and geometric median. The weighted arithmetic mean used the number of training examples at a given institution as a weighting factor when calculating the arithmetic mean of weights. Additionally, we investigated ensembling the three weight averaging methods, so that the federated averaged weight is a combination of the weight estimate from the weighted arithmetic mean (50%), median (10%), and geometric median (40%); the





**Fig. 5.** Variable Epochs per Round (EpR) over the course of federated training. (A) Three types of variable EpR round were investigated. (B) DSC and 95%HD scores on validation data for federated learning with variable EpR were compared against the central training and federated baseline models.

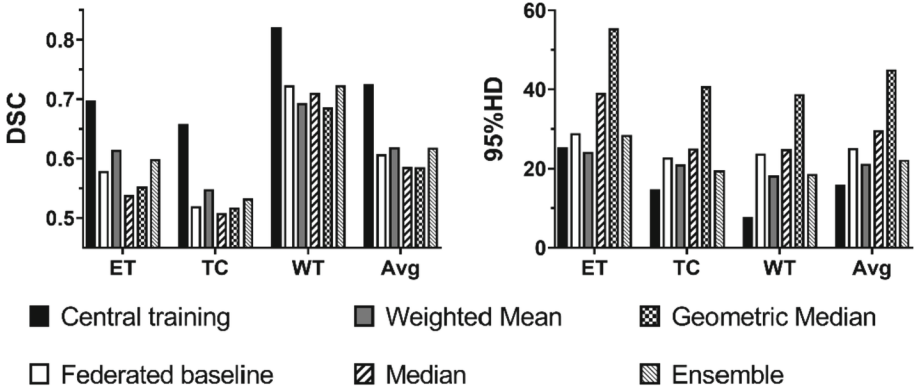
ensemble weights were determined experimentally. Models were trained for 10 rounds at 1 epoch per round and the learning rate was set at  $5e-3$ .

Of the individual weight aggregation functions, the weighted mean consistently performed the best across the regions of interest (ET, TC, WT) and metrics (DSC, 95%HD) (Fig. 6). The median and geometric median were comparable to each other, but generally worse than the baseline federated model that used the unweighted mean.

Compared to the federated baseline, the ensemble also significantly improved performance. On average, the ensemble was comparable to the weighted mean function, and improved DSC scores for WT segmentation. Given this, our final model used the ensemble weight aggregation function.

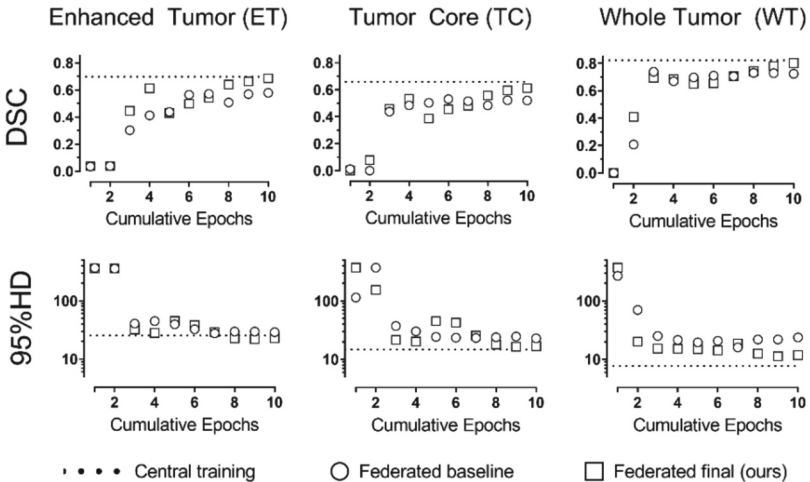
### 3.4 Final Model Results on “Challenge Validation Set”

Our final model was trained using 10 federated rounds with a variable number of epochs per round (“Big Pyramid”, Fig. 5A), using a linearly decaying learning rate (from  $5e-3$  to  $1e-6$ , Fig. 4A), and used an ensemble of weighted arithmetic mean, median and



**Fig. 6.** Federated learning using alternate weight aggregation functions. The federated baseline used an unweighted mean to average model weights. The weighted mean used the number of local training examples at a given institution as a weighting factor when calculating the mean. DSC and 95%HD scores are reported on the validation data.

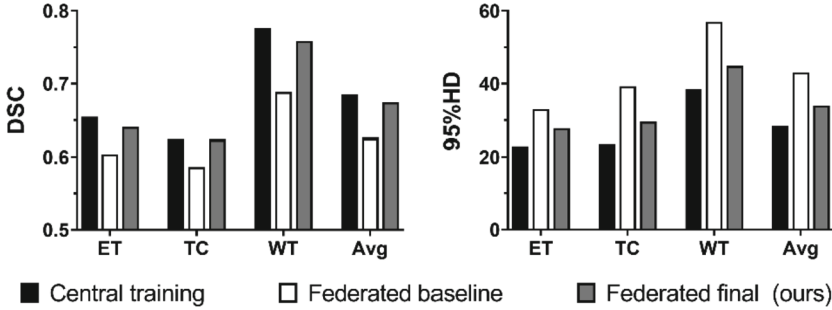
geometric median (50%, 10% and 40%, respectively) (Fig. 6). Overall, the model performed significantly better than the federated baseline model (Fig. 7). In some cases, such as  $DSC_{ET}$ ,  $DSC_{WT}$ ,  $95\%HD_{ET}$ , our final federated model matched central model performance.



**Fig. 7.** Our finalized federated learning model using a “pyramid-shaped” variable number of epochs per round, a linearly decaying learning rate, and an ensemble of weight aggregation functions. DSC and 95%HD scores are reported on the validation data.

Finally, we evaluated our finalized federated training algorithm on unseen data using the visible testing set, also known as the “Challenge Validation” set. Our final federated model showed marked improvements compared to the federated baseline on this dataset

as well (Fig. 8). Similarly, our federated model showed comparable performance to the central model. Notably, the average DSC score for our federated model (0.674) is only 0.011 less than the central model (0.685).

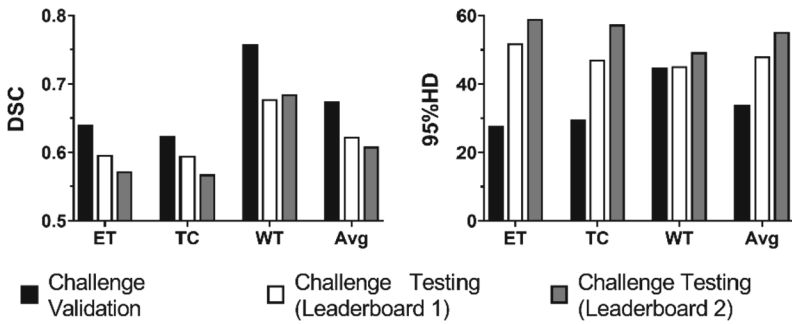


**Fig. 8.** Our finalized federated model performance (DSC and 95%HD scores) on the visible testing set, also known as the “Challenge Validation set”.

### 3.5 Final Model Results on “Challenge Test Set”

Our final model was evaluated by challenge organizers on “Leaderboard 1” and “Leaderboard 2”. The training algorithm and hyperparameters used were developed using the codebase for “Leaderboard 1” and used “as-is” for “Leaderboard 2”. Models were trained using a hidden training partition and evaluated on a hidden testing partition; both were only visible to challenge organizers.

Overall, our final federated training algorithm generalized well to unseen training and testing data, the “Challenge Test set”, with comparable performance in both “Leaderboard 1” and “Leaderboard 2” (Fig. 9). Although some performance drop is to be expected when generalizing to unseen training and testing data, this overall drop in generalization to unseen data was not substantial. The average DSC score on unseen data for our approach only dropped by 8% in “Leaderboard 1” (0.623) and 10% in “Leaderboard 2” (0.608) compared to that predicted by the performance on the “Challenge Validation set” (0.674). Notably, both DSC and 95%HD scores on “Leaderboard 2” were comparable to that on “Leaderboard 1”, despite the fact that no model tuning was done for “Leaderboard 2”. The model communication costs, which were calculated by the challenge organizers using a standardized approach, were 0.773 for “Leaderboard 1” and 0.715 for “Leaderboard 2”.



**Fig. 9.** Our finalized federated model performance (DSC and 95%HD scores) on the “Challenge Validation” and “Challenge Testing” sets (Leaderboard 1 and Leaderboard 2).

**Acknowledgements.** This work was supported by the Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship (AT) and Discovery Grant (NDF), MITACS Globlink Research Internship (LT), University of Calgary BME Research Scholarship (RS), Canada Research Chairs program (NDF), the River Fund at Calgary Foundation (NDF), and the Canadian Institutes of Health Research (CIHR).

**Conflicts of Interest.** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- Lo Vercio, L., et al.: Supervised machine learning tools: a tutorial for clinicians. *Journal of Neural Engineering* **17**(6), 062001 (Oct 9 2020). <https://doi.org/10.1088/1741-2552/abbff2>
- Hinton, G.: Deep learning—a technology with the potential to transform health care. *JAMA - Journal of the American Medical Association*, **320**(11), pp. 1101–1102. American Medical Association (Sep 18 2018). <https://doi.org/10.1001/jama.2018.11100>
- Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**(6), 305–311 (Jun 2020). <https://doi.org/10.1038/s42256-020-0186-1>
- MacEachern, S.J., Forkert, N.D.: Machine learning for precision medicine. *Genome* **64**(4), 416–425 (2021). <https://doi.org/10.1139/gen-2020-0131>. Epub 2020 Oct 22 PMID: 33091314 Apr
- Tuladhar, A., Gill, S., Ismail, Z., Forkert, N.D.: Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by en-sembling. *J. Biomed. Inform.* **106**, 103424 (2020). <https://doi.org/10.1016/j.jbi.2020.103424>. Jun.
- McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-Efficient Learning of Deep Networks from Decentralized Data. *Arxiv* (2016)
- Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* **10**(2), 1–19 (2019). <https://doi.org/10.1145/3298981>. Jan.
- Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* **2**(6), 305–311 (2020). <https://doi.org/10.1038/s42256-020-0186-1>

9. Chang, K., et al.: Distributed deep learning networks among institutions for medical imaging. *J. Am. Med. Informatics Assoc.* **25**(8), 945–954 (2018). <https://doi.org/10.1093/jamia/ocy017>. Aug.
10. Remedios, S.W., et al.: Distributed deep learning across multisite datasets for generalized CT hemorrhage segmentation. *Med. Phys.* **47**(1), 89–98 (2020). <https://doi.org/10.1002/mp.13880>. Jan.
11. Reina, G.A., et al.: OpenFL: An open-source framework for Federated Learning. arXiv preprint [arXiv:2105.06413](https://arxiv.org/abs/2105.06413) (2021)
12. Pati, S., et al.: The Federated Tumor Segmentation (FeTS) Challenge. arXiv preprint [arXiv:2105.05874](https://arxiv.org/abs/2105.05874) (2021)
13. Sheller, M.J., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Nat. Sci. Rep.* **10**, 12598 (2020). <https://doi.org/10.1038/s41598-020-69250-1>
14. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Nature Scientific Data* **4**, 170117 (2017). <https://doi.org/10.1038/SDATA.2017.117>
15. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017). <https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q>
16. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive (2017). <https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF>