

A Comparative Analysis of Federated and Centralized Learning Strategies for Segmenting Acute and Chronic Ischemic Brain Lesions

Wouter De Smet^{1*}, Liesbeth Van Dam¹, Pieter Janssens²

¹ Department of Neuroradiology and Federated Learning, Faculty of Medicine, KU Leuven, Leuven, Belgium.

² Department of Ischemic Stroke Imaging, Faculty of Medicine, Ghent University, Ghent, Belgium.

*E-mail ✉ wouter.desmet@kuleuven.be

Abstract

This study aimed to assess how well federated learning (FL) performs compared with standard centralized learning (CL) for automatic segmentation of acute cerebral infarcts and white matter hyperintensities. We conducted a retrospective multi-center analysis involving 13,546 diffusion-weighted images (DWI) collected from 10 hospitals and 8,421 fluid-attenuated inversion recovery (FLAIR) images from 9 hospitals. These supported two separate tasks: acute lesion segmentation (Task I) and chronic lesion segmentation (Task II). Training utilized data from 9 sites for Task I and 3 sites for Task II, while external testing was performed on held-out data from 1 site (Task I) and 6 sites (Task II). During FL, the central server merged client updates every four communication rounds, applying FedYogi for Task I and FedAvg for Task II. A batch-clipping strategy was also examined in the FL framework. Segmentation accuracy was measured using the Dice similarity coefficient (DSC). Participants in the training cohorts had mean ages of 68.1 ± 12.8 years (Task I) and 67.4 ± 13.0 years (Task II), with males representing 51.5% and 60.4% of the groups, respectively. In Task I, the FL model trained for 360 epochs with batch clipping achieved a mean DSC of 0.754 ± 0.183 . This outperformed a matched CL model trained for the same number of epochs (DSC: 0.691 ± 0.229 ; $P < 0.001$) and showed no significant difference from the strongest CL model trained for 940 epochs (DSC: 0.755 ± 0.207 ; $P = 0.701$). For Task II, the FL model with clipping, the FL model without clipping, and the CL model all produced statistically similar results after 48 epochs (DSCs: 0.761 ± 0.299 , 0.751 ± 0.304 , and 0.744 ± 0.304). Few-shot FL variants performed substantially worse. Batch clipping also halved training time for Task II (from 3.5 hours to 1.75 hours). A side-by-side evaluation of centralized and federated approaches under equivalent conditions shows that federated learning is a practical and effective approach for medical image segmentation.

Keywords: Federated learning, Image segmentation, Ischemic brain lesion, Machine learning

Introduction

Extensive prior research has applied machine learning (ML) methods to segment various medical images, often reporting strong results [1-3]. Most of these efforts depend on pre-processed public datasets. However, deploying such models in everyday clinical settings is hindered by performance drops and domain shifts when facing new data [4]. Overcoming these limitations

requires training on more diverse, real-world datasets from different hospitals and scanner manufacturers. Unfortunately, legal and privacy regulations make it difficult for institutions to share raw patient images with external parties for centralized model development.

Federated learning (FL) [5, 6] offers an alternative by allowing model training across distributed data sources without moving the actual images. In this setup, each participating institution (client) trains a local copy of the model on its private data, then sends only the updated model weights to a central server. The server aggregates these weights to create an improved global model and distributes it back to the clients for the next training round. This process repeats while keeping all patient data securely on-site. As a result, FL enables collaborative model improvement across multiple hospitals without

Access this article online

<https://smerpub.com/>

Received: 12 November 2024; Accepted: 19 December 2024

Copyright CC BY-NC-SA 4.0

How to cite this article: De Smet W, Van Dam L, Janssens P. A Comparative Analysis of Federated and Centralized Learning Strategies for Segmenting Acute and Chronic Ischemic Brain Lesions. J Med Sci Interdiscip Res. 2024;4(2):195-205. <https://doi.org/10.51847/FmQ0cJp7G>

compromising data privacy, making it especially attractive for healthcare applications where centralized data sharing is restricted.

Nevertheless, some researchers question whether FL can match the performance of traditional centralized learning (CL) [7]. A key concern is the statistical heterogeneity (non-i.i.d. nature) of data across institutions [8], which may include differences in image quality, scanner parameters, patient demographics, and disease patterns. Such variations can complicate model convergence during aggregation. Frequent communication between clients and the server also adds practical challenges for some facilities. Although low-communication variants like one-shot or few-shot FL are under investigation [9], they require further validation.

Existing studies have mostly compared different FL algorithms using public benchmarks such as CIFAR-10 or BraTS [10, 11]. The current work provides a rigorous head-to-head comparison between cross-silo FL and high-performance industry CL models using a large, heterogeneous clinical brain MRI dataset. We also examined FL with regular communication rounds versus minimal-communication (one-shot) strategies. This represents one of the largest-scale FL experiments in medical imaging to date. In addition, we introduce a batch clipping technique that limits the number of samples each client processes per epoch, substantially accelerating training with little to no loss in final accuracy.

Materials and Methods

The study utilized two retrospective datasets: one for acute ischemic lesions (Task I) and one for chronic

lesions (Task II). Both datasets had been previously employed to develop the CL models evaluated here. To ensure a direct and fair comparison, we replicated the exact network architecture from the commercial CL software for the FL experiments. All data pre-processing, splitting, and evaluation procedures were kept identical between the FL and CL arms. Written informed consent was obtained from all participants or their legal representatives. The study adhered to the ethical principles of the 1964 Declaration of Helsinki and received Institutional Review Board approval from Dongguk University Ilsan Hospital.

Dataset and brain image analysis

Acute ischemic lesion dataset

From May 2011 to November 2013, 14,740 consecutive patients with ischemic stroke or transient ischemic attack were enrolled at 10 centers within one week of symptom onset (**Figure 1**). After excluding 1,194 cases due to MRI contraindications ($n = 688$), poor or unavailable DWI ($n = 321$), or registration errors ($n = 185$), the final training and validation set comprised 13,546 subjects from 9 institutions, with external testing performed on data from the remaining institution. A previous CL study [12] had analyzed 13,597 subjects across all 10 sites. DWI scans were acquired using 1.5-T ($n = 9,084$) or 3.0-T ($n = 2,923$) scanners with b-values of 0 and 1,000 s/mm², TR 2,400–9,000 ms, TE 50–99 ms, voxel size $1 \times 1 \times 3$ –5 mm³, interslice gap 0–2 mm, and slice thickness 3–7 mm. Expert researchers manually delineated ischemic lesions on the DWIs using dedicated in-house software under the supervision of a senior vascular neurologist, following standardized protocols described in earlier publications [13–15].

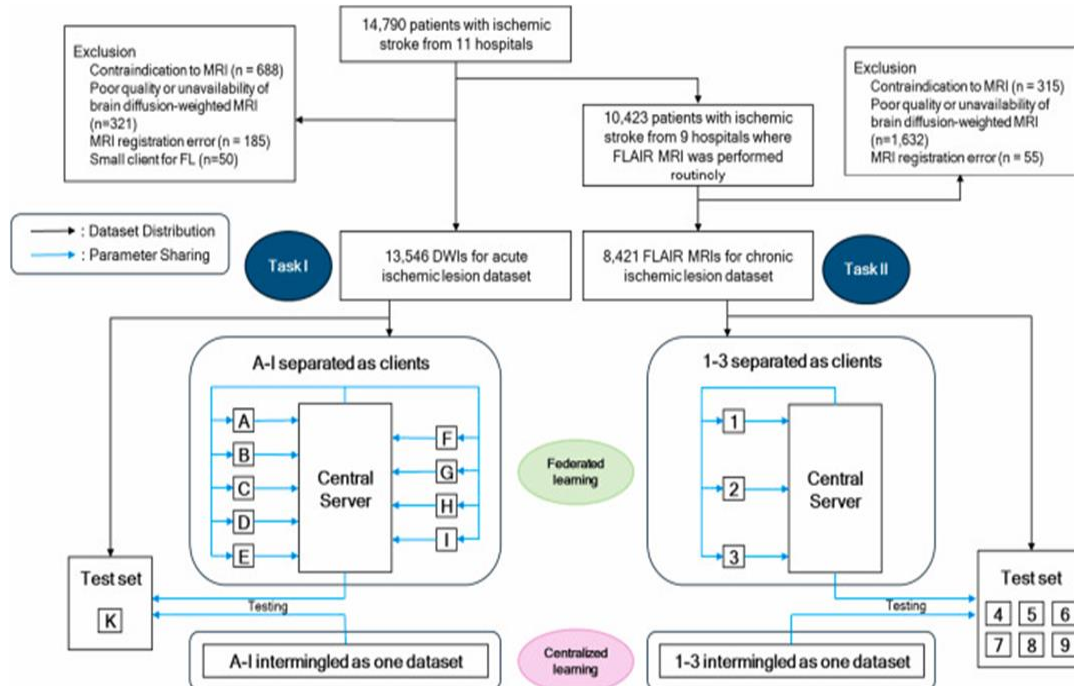


Figure 1. Flow diagram and overall design of the study for Tasks I and II. Blue arrows represent the circulation of model parameters: distributed from the central server to each client for local training and then returned to the server for aggregation.

Chronic ischemic lesion dataset

From May 2011 to November 2013, 10,423 consecutive patients admitted to 9 participating centers within seven days of symptom onset were enrolled. After exclusions for MRI contraindications ($n = 315$), insufficient or low-quality DWI/FLAIR images ($n = 1,632$), and registration failures ($n = 55$), 8,421 subjects remained. These were divided such that data from 3 institutions were used for training and data from 6 institutions for testing. All cases originated from the prior centralized learning investigation [16]. FLAIR imaging was conducted on 1.5-T ($n = 6,583$) or 3.0-T ($n = 1,803$) scanners with the following parameters: echo time (TE) 76–187 ms, repetition time (TR) 6,000–11,568 ms, voxel size $1 \times 1 \times 3\text{--}7\text{ mm}^3$, interslice spacing 0.3–1.0 mm, slice thickness 3–7 mm, field of view (FOV) 175–280 mm, and matrix size 256–768 (row). As described in previous reports [15, 17, 18], six trained researchers manually outlined regions of high signal intensity on FLAIR images, corresponding to white matter lesions, under the direct oversight of an experienced vascular neurologist. In cases where chronic white matter hyperintensities on FLAIR overlapped or bordered acute infarcts on DWI, lesion boundaries were delineated by referencing the contralateral hemisphere (unaffected by the acute infarct), taking advantage of the

typical symmetry in morphology and spatial distribution of white matter hyperintensities between hemispheres [15, 17, 18].

Image pre-processing

Acute ischemic lesion dataset

DWI scans underwent pre-processing that included Otsu's thresholding and N4 bias field correction [19, 20]. Signal intensities were first scaled to the 0–255 range, followed by z-score normalization. Data augmentation was applied using a combination of slice-wise similarity transformations, simulated MRI bias fields, axial flips, and adjustments to gamma and contrast levels.

Chronic ischemic lesion dataset

FLAIR slices were resized to 256×256 pixels via 2D B-spline interpolation. Slice-level intensity normalization was then performed to map pixel values into the uint8 range (0–255). To harmonize images acquired from different MRI vendors, case-wise histogram equalization with 32 bins was performed, shifting the dominant peak to an intensity of 150. A binary brain mask was created by thresholding at an intensity of 30 and filling small gaps with the `binary_fill_holes` function from `scipy.ndimage`. Finally, Gaussian normalization was

applied within the brain mask region to achieve a mean of 0 and standard deviation of 1, generating input arrays optimized for deep learning model training.

Batch clipping

A substantial imbalance in dataset sizes existed, with the largest participating client holding approximately three times as much data as the smallest. This imbalance caused prolonged idle time for smaller clients while waiting for larger ones to complete local training, reducing overall efficiency. To address this and shorten the total training duration, a batch-clipping strategy was

implemented that sets an upper limit on the number of batches each client processes per epoch. Although some samples might be skipped in a given epoch, random shuffling across many iterations ensures that nearly all data points are eventually utilized during training. The maximum batch counts were set at 200 for Task I, and 1,250 for Task II (**Figure 2**). These thresholds corresponded to roughly 50–55% of the largest client's batch count. Clients with fewer batches than the specified limit continued to use their complete datasets each epoch without modification. Both tasks were evaluated with and without this batch-clipping approach to assess its impact.

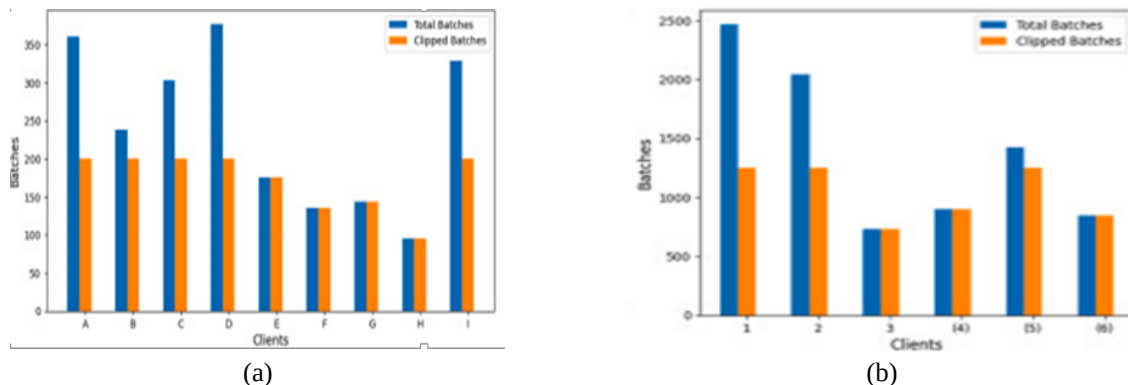


Figure 2. Total and clipped batch counts per client for Tasks I and II. Blue bars represent the full batch count from each client's entire dataset, while red bars indicate the reduced number after clipping. Clients stop local training once they reach the clipping threshold within an epoch. Clipping limits were set at 200 batches for Task I and 1,250 batches for Task II. Institutions with fewer batches than the threshold were not affected and continued using their complete data each epoch.

Implementation details

The FL models developed for both tasks were benchmarked against an established industry-grade CL model. Model architecture, training duration (epochs), hyperparameter settings, optimizers, and loss functions were aligned as closely as possible with those used in earlier studies. To limit communication costs between the central server and clients, local training was performed for four epochs per communication round before parameter aggregation. Following each aggregation step, clients evaluated the updated global model using their local validation data. All experiments were executed on three RTX A6000 GPUs, utilizing the Flower framework [21] for federated learning.

Two aggregation strategies were tested: FedAvg [5] and FedYogi [22]. FedAvg computes a simple coordinate-wise average of the client models received. FedYogi applies the Yogi optimizer [23] at the server level for improved update dynamics and serves as an enhanced

version of FedAvg [24]. These methods were selected because they function as accessible baselines for more complex approaches. Demonstrating strong performance with such straightforward FL implementations would provide robust evidence for the practicality of federated learning in medical imaging. Importantly, neither method modified the local client-side training procedure from standard CL, preserving compatibility with the batch clipping technique.

For Task I, acute ischemic lesions were segmented on DWI scans using the STU-Net architecture [25]. The dataset was randomly split into training and validation sets at a 4:1 ratio. External testing included 2,777 DWI scans from a single institution, which were excluded from the training set. Since the reference CL model required 1,000 epochs for full convergence, FedYogi was chosen as the aggregation method for its anticipated advantage in longer training scenarios. Both FL and CL implementations used PyTorch [26]. During intermediate

evaluations of test Dice similarity coefficients (DSCs) across rounds used, only the first 100 scans from the external test set were used due to computational constraints; final reported DSCs were computed on the complete external test set.

For Task II, chronic ischemic lesions (white matter hyperintensities) were segmented on FLAIR images with a U-Net model [27]. The dataset partitioning mirrored the original CL study (3:1:1 ratio for training, validation, and internal testing). However, the internal test set was not used; instead, external validation was performed on data from six additional institutions, with slice counts ranging from 8,422 to 53,344 per site. FedAvg was selected for aggregation in Task II to demonstrate effective convergence without extensive hyperparameter optimization. Both FL and CL models were implemented in TensorFlow [28]. For efficiency during epoch-wise evaluations (epochs 2–12), only the first 2,000 slices of the test sets were analyzed; final DSCs were calculated using the full test data.

Supplementary experiments for Task II involved reassigning three institutions originally in the external test set to the training client pool. This was done to evaluate FL robustness under different client configurations. These three institutions (**Figure 1**) were excluded from the primary FL results but included in the additional analysis. All other experimental parameters remained consistent with the main Task II setup.

Statistical analysis

Demographic and imaging features between the training/validation and external test cohorts were compared using appropriate tests: t-tests, rank-sum tests, ANOVA, or Kruskal-Wallis tests for continuous variables, and the chi-square test for categorical variables.

Model performance in both tasks was quantified with the Dice similarity coefficient (DSC). All reported DSCs represent mean and standard deviation values for the specified datasets. Primary comparisons between FL and CL models employed paired t-tests per client. For Task II external test data, one-way ANOVA was used to compare DSCs across clients. Statistical significance was set at $P < 0.05$. Analyses were conducted with the `scipy.stats` library (version 1.10.1) in Python (version 3.8.10, The Python Software Foundation).

Results and Discussion

Patient demographics

In Task I, mean ages (SD) were 68.1 (12.8) years for the training/validation set and 68.2 (12.4) years for the external test set. External test cohorts ranged in age from 67.5 to 70.0 years. All examined variables differed significantly across datasets, confirming substantial heterogeneity.

Main test results

For Task I, FedYogi-based FL models trained for 90 rounds (360 epochs) achieved DSCs of 0.754 ± 0.183 (with batch clipping) and 0.762 ± 0.173 (without clipping) (**Table 1**). These outperformed the CL model trained for the same 360 epochs (DSC 0.691 ± 0.229 ; $P < 0.001$) and were comparable to the best CL model after 940 epochs (DSC 0.755 ± 0.207 ; $P = 0.701$). When extended to 250 rounds (1,000 epochs), both FL variants (with and without clipping) reached a DSC of 0.776 ± 0.170 and 0.776 ± 0.169 , respectively, surpassing the 940-epoch CL model ($P < 0.001$). In Task II, after 48 epochs, the FL model with clipping (0.751 ± 0.304), the FL model without clipping (0.761 ± 0.299), and the CL model (0.744 ± 0.304) showed similar performance for chronic lesion segmentation.

Table 1. Comparison of segmenting performance between federated and centralized learning.

Empty cell		Acute ischemic lesion							Chronic ischemic lesion					
Model		STU-Net							U-Net					
Number of clients for training		9							3					
Aggregation method		FL, FedYogi							CL		FL, FedAvg		CL	
Batch Clipping		Yes			No				N/A		Yes		No	N/A
Rounds ^a		33	90	250	1	90	250	N/A	N/A	12	2	12	N/A	
Epochs		132	360	1000	132	360	1000	360	940	48	48	48	48	
Avg. DSC ^b		0.717	0.754	0.776	0.142	0.762	0.776	0.691	0.755	0.751	0.581	0.761	0.744	
SD ^b		0.199	0.183	0.170	0.184	0.173	0.169	0.229	0.207	0.304	0.437	0.299	0.304	

Abbreviations: FL = federated learning; CL = centralized learning; N/A = not available; DSC = Dice similarity coefficient; SD = standard deviation.

^a After each round, the training results were sent to the server, aggregated using FedYogi or FedAvg, and returned to each client.

^b Average DSC and standard deviation were calculated using DSC per patient.

For comparison, we performed additional experiments with few-shot federated learning, where parameter aggregation happened only once or twice following prolonged independent local training at each site. In Task I, conducting a single FedYogi aggregation after 132 epochs of local training yielded a poorly performing model (DSC: 0.142 ± 0.184 ; $P < 0.001$ compared to the standard FL approach after 37 rounds or 148 epochs). For Task II, a two-round FedAvg setup was examined, with each round comprising 24 epochs (for a total of 48). This configuration produced significantly worse outcomes than the standard protocol using four epochs per round (DSC: 0.581 ± 0.437 ; $P < 0.001$).

Task I convergence

Figures 3a and 3b display the evolution of training and validation loss curves throughout Task I. The centralized model, along with all FL clients that applied batch

clipping, showed stable, successful convergence, although they settled at varying loss values. Upon reaching 1,000 epochs, the CL model recorded a training loss of 0.121 (**Figure 3a**). In contrast, individual FL clients achieved training losses ranging from 0.113 to 0.182, yielding a weighted average of 0.144. Regarding validation loss (**Figure 3b**), the CL model attained 0.126, while FL clients ranged from 0.098 to 0.176, with an overall weighted average of 0.138.

Figure 3c tracks the test DSC performance of FL and CL models across training epochs, supplemented by a smoothed version of the curves for easier interpretation. Despite FL models consistently achieving higher average losses than the CL counterpart during training and validation, they surpassed CL performance on the external test set after roughly 200 epochs and ultimately achieved a higher final DSC value.

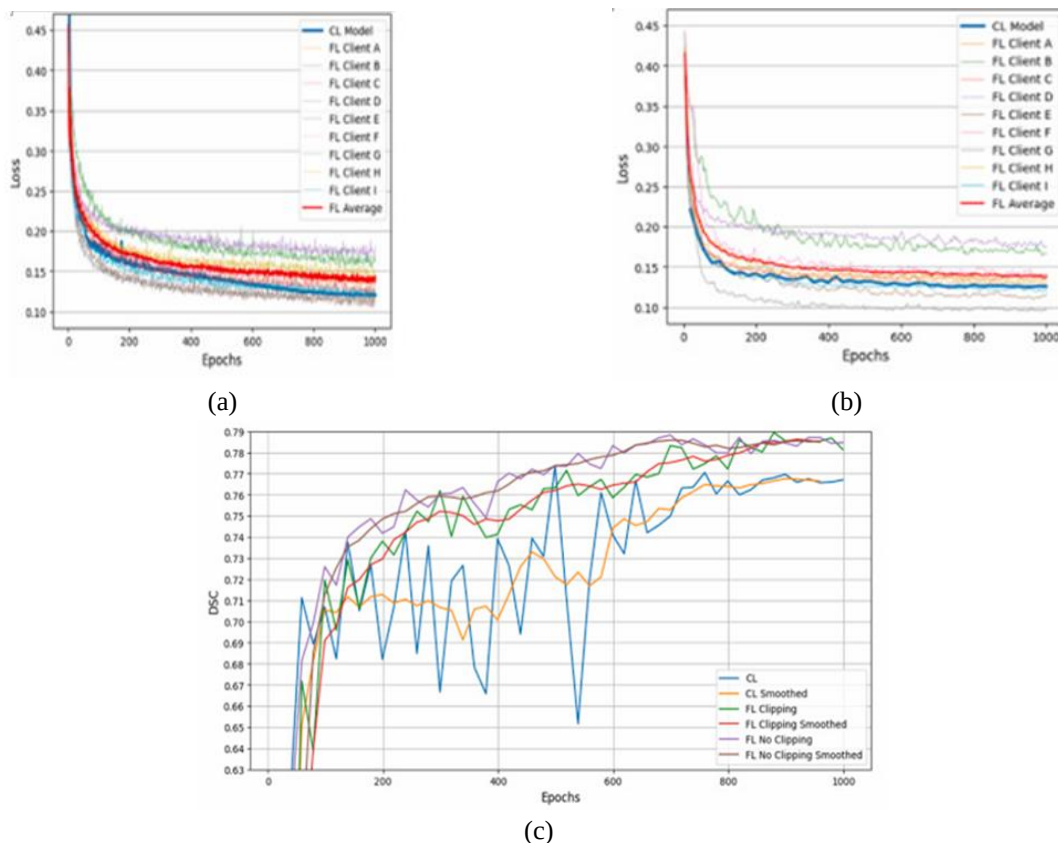


Figure 3. Overview of STU-Net outcomes in Task I, including loss trends and test DSC values: (a) Training loss progression plotted for the centralized model and batch-clipped FL clients 0 through 8 over the full 1,000 epochs;

(b) Validation loss progression for the centralized model and batch-clipped FL clients 0–8 throughout 1,000 epochs, with evaluations conducted every 4 epochs immediately after each server aggregation;

(c) Changes in test

DSC for the centralized model and the combined FL models, measured using the initial 100 images from the external test set across up to 1,000 epochs. A smoothed line, created by convolving with the prior 5 data points, is shown to aid direct comparison.

Task II convergence

Figures 4a and 4b display validation DSC curves for the individual training clients participating in Task II under both batch-clipped and non-clipped FL conditions. Observed test DSC values fell below validation levels, which is attributable to the test data originating from entirely separate institutions that may possess differing image distributions. **Figure 4c** tracks the test DSC development for the FL model using clipping, the FL model without clipping, and the standard CL model. By the conclusion of 48 epochs (equivalent to 12 aggregation

rounds), all three configurations stabilized near a DSC of 0.75.

Table 2 lists the mean test DSC achieved by each external test client. Despite noticeable differences in performance among these clients, a one-way ANOVA of the per-client average DSCs found no statistically significant difference ($P = 0.650$). The complete training process for 48 epochs required roughly 3.5 hours for the FL model without batch clipping, compared to only 1.75 hours with batch clipping.

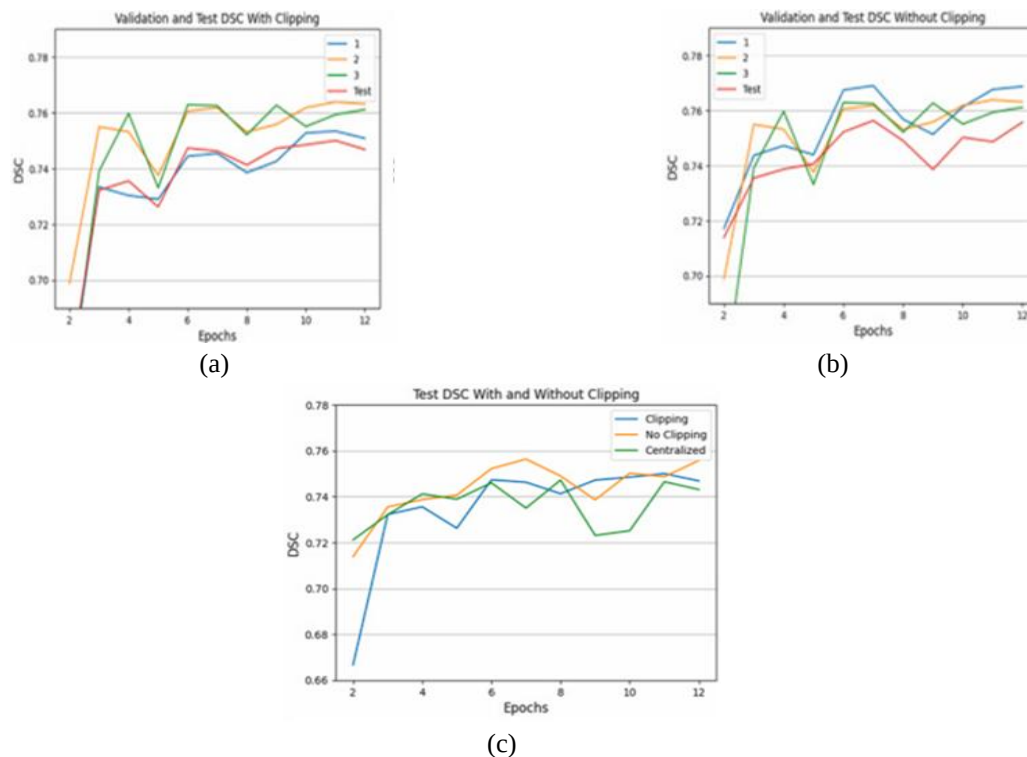


Figure 4. Validation and test DSC results for Task II: (a) Validation DSC curves for clients 1–3 along with the test DSC of the aggregated global model across rounds 2–12 when using batch clipping; (b) Validation DSC curves for clients 1–3 along with the test DSC of the aggregated global model across rounds 2–12 when batch clipping was not applied; (c) Direct comparison of test DSC performance among the models with clipping, without clipping, and the centralized approach. All test DSC values displayed in this figure were computed using the initial 2,000 slices from the test set. Due to FedAvg producing near-zero performance (around 0.1) during the initial aggregation, all Task II graphs begin from the second round onward.

Table 2. Task II average Dice similarity coefficient for each client in external test datasets.

Client ^a	Clipping	No Clipping	Centralized
	12 Round	12 Round	48 Epoch

4	0.735	0.737	0.724
5	0.704	0.711	0.704
6	0.743	0.757	0.739
7	0.719	0.727	0.710
8	0.763	0.774	0.755
9	0.779	0.790	0.772
Total	0.751	0.761	0.744
SD^b	0.0254	0.0271	0.0243

Abbreviation: SD = standard deviation.

^a Each client represents a separate institution.

^b Standard deviation was calculated using DSC per center.

Additional experiments: task II training with six clients

To more thoroughly evaluate the effects of training variability and rule out dependence on any particular choice of participating sites, we performed a follow-up experiment for Task II. A one-way ANOVA examining the mean Dice Similarity Coefficient (DSC) values across the external test centers indicated no statistically significant differences ($P = 0.846$). These results highlight the FL model's robustness to substantial changes in the training cohort and demonstrate that its success does not depend on a specific set of clients.

The second supplementary experiment examined the impact of applying batch clipping within a centralized learning (CL) framework. To this end, we merged the datasets from the FL batch-clipped training runs to create a single, centralized dataset. Test DSCs were computed using the same protocol as the primary Task II experiments. The findings revealed that CL with batch clipping achieved inferior results compared to FL with batch clipping, exhibiting slower convergence and lower final DSC values.

Client 5, which achieved the strongest standalone performance, reached a DSC of 0.749—comparable to the FL models—while the other five clients failed to exceed 0.73 DSC after 48 epochs. This suggests that most institutions would have strong incentives to join FL collaborations rather than train models in isolation. The volume of data available to each client was a major determinant of standalone training success, except for Client 5.

This investigation evaluated FL and CL approaches for segmenting acute and chronic ischemic brain lesions, confirming the feasibility of FL for this clinical task. We assessed both a data-heavy scenario (Task A) and a more resource-light setting (Task B) to illustrate the broad applicability of FL across different scales. We further showed that one-shot or few-shot communication strategies produced weaker FL performance than

sustained, regular communication. Finally, we provided empirical evidence that batch clipping is an effective technique for shortening training duration while preserving model robustness.

The BraTS dataset, a widely used multi-center benchmark for brain image segmentation, has featured prominently in prior medical FL research. Pati *et al.* [3] leveraged BraTS along with additional datasets to develop a leading FL model; however, their training began from a “public initial model” pre-trained on BraTS rather than from random initialization. Li *et al.* [10] similarly employed BraTS to model non-i.i.d. distributions of medical data and to explore differential privacy. The considerably larger scale of our datasets (approximately 10,000 images) demonstrates that FL remains effective even on bigger data volumes and when starting from randomly initialized weights.

Our results indicate that successful FL requires ongoing communication between clients and the central server, rather than relying on a single or two aggregation rounds after extended local training. Although a two-round aggregation strategy produced acceptable outcomes for Task II, it lacked generalization to unseen data and underperformed relative to models that communicated every four epochs. FedYogi and FedAvg proved unable to coherently aggregate all model parameters within a single step, resulting in suboptimal performance after a limited number of communication rounds.

Batch clipping successfully halved training time compared with unclipped models. In FL, where training proceeds in parallel across institutions, the total runtime is determined by the slowest client per round rather than by the overall data volume. Consequently, capping the maximum number of batches processed per epoch substantially shortens the overall process. Notably, increasing the number of clients in Task II did not meaningfully extend FL training time, underscoring the benefits of parallelization. In contrast, CL training time

scaled linearly with dataset size, reaching approximately 6.5 hours in our experiments. CL also benefited less from batch clipping than FL: although wall-clock time decreased markedly, the corresponding reduction in effective training data had a more pronounced negative impact on model quality.

We attribute the success of batch clipping to the aggregation mechanisms used (FedYogi and FedAvg). These methods compute weighted averages of client updates, accounting for the number of batches each client processes per round. As a result, even when some clients see fewer examples per epoch, the overall gradient magnitude remains appropriately balanced because contributions from other clients receive higher weight. Although not tested here, we expect batch clipping to be compatible with alternative aggregation rules, such as KRUM [29] and median aggregation, as well as privacy-enhancing techniques, such as Differential Privacy [30]. It is important to note that our batch clipping implementation deviates from standard theoretical formulations. Conventional aggregation typically weights client contributions by dataset size rather than by update count. Batch clipping effectively treats every client as having at most a fixed number of data points per round (randomly sampled). This leads to heterogeneous numbers of updates per sample depending on whether the sample belongs to a larger or smaller client's dataset. Consequently, our empirical results diverge somewhat from pure theory, and it remains challenging to determine whether this discrepancy harms long-term FL training. In the experiments conducted, no performance degradation attributable to batch clipping was observed; nonetheless, the technique warrants additional theoretical and empirical examination in future work. One straightforward mitigation is to adjust client selection probabilities (instead of uniform sampling) so that each client receives an equal number of updates during training.

In Task I, the FL model surprisingly outperformed the CL model on external test data. This outcome contrasts with the typical finding that FL performance is bounded below that of CL [31]. This reversal was not observed in the training or validation losses, where the average FL loss remained higher than the CL loss, even though some clients achieved lower local losses. Possible explanations include the variance-reducing effect of averaging client updates, consistent with the “bucketing” concept proposed by Karimireddy *et al.* [32]. Another plausible factor is that the smaller clients in the training set

happened to better represent the test distribution. In CL, contributions from these smaller but better-matched datasets may be diluted by data from larger institutions. Although FL still performs weighted aggregation based on dataset size, individual small clients can fully influence the model with their well-aligned data, potentially boosting generalization on the test set.

Regarding the third supplementary study, Client 5 outperformed the FL model that included its data. We do not interpret this as evidence that Client 5's data possessed superior generalization properties. When Client 5 was part of the test set for an FL model trained only on Clients 1–3, it exhibited substantially lower DSC, indicating a markedly different data distribution. A reasonable interpretation is that Client 5 possesses greater internal heterogeneity than other clients. This heterogeneity may have been detrimental when it served as unseen test data but beneficial when the client trained independently.

Several limitations should be acknowledged. All experiments were conducted in a simulated environment, with the primary goal of adapting an existing CL model to FL. Consequently, data were uniformly pre-processed and maintained as high-quality pixel-level annotations. Real-world deployments would likely involve less idealized conditions [33]. Minor design choices were also made, such as reducing the batch size from 12 to 4 in Task I simulations due to GPU memory limits, and selecting four epochs per communication round arbitrarily. These decisions are expected to have negligible influence on the core findings but are disclosed for transparency. Additionally, Task I experiments were not repeated across multiple random seeds or varied training partitions due to computational constraints. Although we observed statistically comparable or superior performance for FL versus CL in Task I, further large-scale validation is encouraged.

Conclusion

In summary, this work presents two FL models for brain lesion segmentation that achieve performance on par with CL counterparts, reinforcing the viability of FL for medical image analysis. We also successfully employed batch clipping to enable efficient training. These insights are expected to support future FL research involving larger numbers of collaborating institutions and more heterogeneous datasets.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Adnan M, Kalra S, Cresswell JC, Zafar S, Qayyum A, Razzak I, et al. Federated learning and differential privacy for medical image analysis. *Sci Rep*. 2022;12(1):1953–7.
2. Feng B, Shi J, Huang L, Chen Y, Sun Y, Wang X, et al. Robustly federated learning model for identifying high-risk patients with postoperative gastric cancer recurrence. *Nat Commun*. 2024;15(1):742–8.
3. Pati S, Baid U, Edwards B, Sheller M, Foley P, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. *Nat Commun*. 2022;13(1):7346–52.
4. Yan W, Wang Y, Gu S, Huang L, Yan F, Xia L, et al. The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN. In: Shen D, Liu T, Peters TM, et al., editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham: Springer; 2019. p. 623–31.
5. McMahan B, Moore E, Ramage D, Hampson S, Arcas B, Yang Q. Communication-efficient learning of deep networks from decentralized data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, FL: PMLR; 2017. p. 1273–82.
6. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol*. 2019;10(2):1–19.
7. Zhang Y, Zeng D, Luo J, Liu Y, Xu Z, Wang S, et al. A survey of trustworthy federated learning: issues, solutions, and challenges. *ACM Trans Intell Syst Technol*. 2024;15(3):1–42.
8. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated learning with non-iid data. *arXiv:1806.00582 [cs.LG]*. 2018.
9. Kang M, Chikontwe P, Kim S, Park J, Lee M, Kim J, et al. One-shot federated learning on medical data using knowledge distillation with image synthesis and client model adaptation. In: Greenspan H, et al., editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Cham: Springer; 2023. p. 521–31.
10. Li W, Milletari F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-preserving federated brain tumour segmentation. In: Suk HI, Liu M, Yan P, et al., editors. *Machine Learning in Medical Imaging – MLMI 2019*. Cham: Springer; 2019. p. 133–41.
11. Cho YJ, Wang J, Joshi G. Towards understanding biased client selection in federated learning. In: *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Virtual: PMLR; 2022. p. 10351–75.
12. Noh YG, Ryu WS, Schellingerhout D, Kim H, Lee J, Park S. Deep learning algorithms for automatic segmentation of acute cerebral infarcts on diffusion-weighted images: effects of training data sample size, transfer learning, and data features. *medRxiv*. 2023. doi:10.1101/2023.07.02.23292150
13. Ryu WS, Kang YR, Noh YG, Kim H, Lee J, Park S. Acute infarct segmentation on diffusion-weighted imaging using deep learning algorithm and RAPID MRI. *J Stroke*. 2023;25(3):425–9.
14. Ryu WS, Schellingerhout D, Hong KS, Jeong SW, Kim SH, Lee J. Relation of pre-stroke aspirin use with cerebral infarct volume and functional outcomes. *Ann Neurol*. 2021;90(5):763–76.
15. Ryu WS, Woo SH, Schellingerhout D, Hong KS, Jeong SW, Kim SH. Stroke outcomes are worse with larger leukoaraiosis volumes. *Brain*. 2017;140(1):158–70.
16. Kim H, Ryu WS, Schellingerhout D, Noh YG, Lee J, Park S, et al. Automated segmentation of MRI white matter hyperintensities in 8,421 patients with acute ischemic stroke. *AJNR Am J Neuroradiol*. 2024;45(4):512–9.
17. Ryu WS, Schellingerhout D, Hong KS, Jeong SW, Kim SH, Lee J. White matter hyperintensity load on stroke recurrence and mortality at 1 year after ischemic stroke. *Neurology*. 2019;93(6):e578–89.
18. Ryu WS, Woo SH, Schellingerhout D, Hong KS, Jeong SW, Kim SH. Grading and interpretation of white matter hyperintensities using statistical maps. *Stroke*. 2014;45(12):3567–75.
19. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310–20.

20. Otsu N. A threshold selection method from gray-level histograms. *Automatica*. 1975;11(285–96):23–7.
21. Beutel DJ, Topal T, Mathur A, Qiu X, Parcollet T, Lane ND. Flower: a friendly federated learning research framework. *arXiv:2007.14390 [cs.LG]*. 2020.
22. Reddi S, Charles Z, Zaheer M, Garrett Z, Rush K, Konečný J, et al. Adaptive federated optimization. *arXiv:2003.00295 [cs.LG]*. 2020.
23. Zaheer M, Reddi S, Sachan D, Kale S, Kumar S, Savarese S. Adaptive methods for nonconvex optimization. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Montréal: Curran Associates; 2018. p. 9793–803.
24. Li Q, Diao Y, Chen Q, He B. Federated learning on non-iid data silos: an experimental study. In: *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. Kuala Lumpur: IEEE; 2022. p. 965–78.
25. Huang Z, Wang H, Deng Z, Chen Y, Liu Y, Zhang L, et al. STU-Net: scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv:2304.06716 [eess.IV]*. 2023.
26. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vancouver: Curran Associates; 2019. p. 8024–35.
27. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, et al., editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer; 2015. p. 234–41.
28. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467 [cs.DC]*. 2016.
29. Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Machine learning with adversaries: Byzantine tolerant gradient descent. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Long Beach, CA: Curran Associates; 2017. p. 119–29.
30. Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, et al. Federated learning with differential privacy: algorithms and performance analysis. *IEEE Trans Inf Forensics Secur*. 2020;15:3454–69.
31. Fauzi MA, Yang B, Blobel B. Comparative analysis between individual, centralized, and federated learning for smartwatch based stress detection. *J Pers Med*. 2022;12(10):1684.
32. Karimireddy SP, He L, Jaggi M. Byzantine-robust learning on heterogeneous datasets via bucketing. *arXiv:2006.09365 [cs.LG]*. 2020.
33. Wicaksana J, Yan Z, Zhang D, Huang X, Wu H, Li L, et al. FedMix: mixed supervised federated learning for medical image segmentation. *IEEE Trans Med Imaging*. 2022;42(7):1955–68.