A Federated Learning Scheme based on Deep Ensemble Learning for MRI-based Multi-Institutional Study of Brain Metastasis Segmentation

Jingtong Zhao[1], Chendong Ni[2], Zhenyu Yang[2], Eugene Vaios[1], John Ginn[1], Deshan Yang[1], Ke Lu[1], Dominic LaBella[1], Evan Calabrese[1], Kyle Lafata[1], Chunhao Wang[1]
(1)Duke University, Durham, NC
(2)Duke Kunshan University, Jiangsu, China

Introduction

Brain metastases (BMs) are the most prevalent CNS malignancy with an incidence rate up to 30% among cancer patients[1]. Stereotactic radiosurgery (SRS) is an established primary treatment for brain metastasis. Accurate BM segmentation is essential for SRS planning and longitudinal outcome assessment; however, manual contouring remains time-consuming and prone to inter-observer variability. Deep learning (DL) segmentation can reduce clinical workload, yet performance and reliability are often constrained by (1) limited local dataset size, (2) heterogeneous imaging protocols across institutions, and (3) data-sharing restrictions that prevent training on an aggregated multi-institutional cohort[2].

Purpose

To propose a federated learning (FL) framework incorporating a novel deep ensemble strategy for multi-institutional BM segmentation, improving performance in limited local datasets while preserving privacy by avoiding large-scale data transfer for retraining.

Materials/Methods

Two BM cohorts (n=90/n=459) with post-contrast T1w MRI and expert BM contours were collected from two institutions and split into training/validation/test sets (7:1:2), respectively. A 3D spherical image transformation based on inhomogeneous lattice scaling in spherical coordinates was designed and applied to each MR volume using 27 equally spaced spherical centers across the field-of-view (FOV), generating locoregional views with varying anatomical detail. A customized UNet++ served as the segmentation backbone and was trained on the transformed images, with ground-truth contours undergoing the same transforms. For each case, 27 segmentations were generated, inversely mapped back to the original Cartesian grid, and fused via a learned ensemble inference strategy to produce the final segmentation.

Fig.1 summarizes the proposed 2-client FL implementation. Trained models from both institutions, i.e., M_S and M_L, were sent to the central aggregation server without any image data transfer. FedAve algorithm[3] was implemented to derive a global model that weighs both local models, which was then redistributed to the local clients. In each communication round, each client performed local training on its private dataset (with model selection guided by the local validation set) and uploaded the updated model to the aggregation server for the next round of global aggregation and redistribution. Leveraging our computation power, 10 communication rounds were studied. In each communication round, each client performed local training on its private dataset (with model selection guided by the local validation set) and uploaded the updated model to the aggregation server for the next round of global aggregation and redistribution. For Institution-S and Institution-L test reporting of M_{FL}, the relative weights of M_S and M_L were set inversely proportional to their dataset sizes to maintain each institution's model characteristics when reporting their test results.

Four models were compared: (1) a baseline model trained on Institution-S (n=90) original data (M_B); (2) a local ensemble model at Institution-S (M_S); (3)a local ensemble model at Institution-L (M_L, n=459); and (4) the proposed FL using M_S and M_L as clients (M_{FL}). Segmentation performance was evaluated using Dice coefficient, precision, recall (sensitivity), and F1-score.

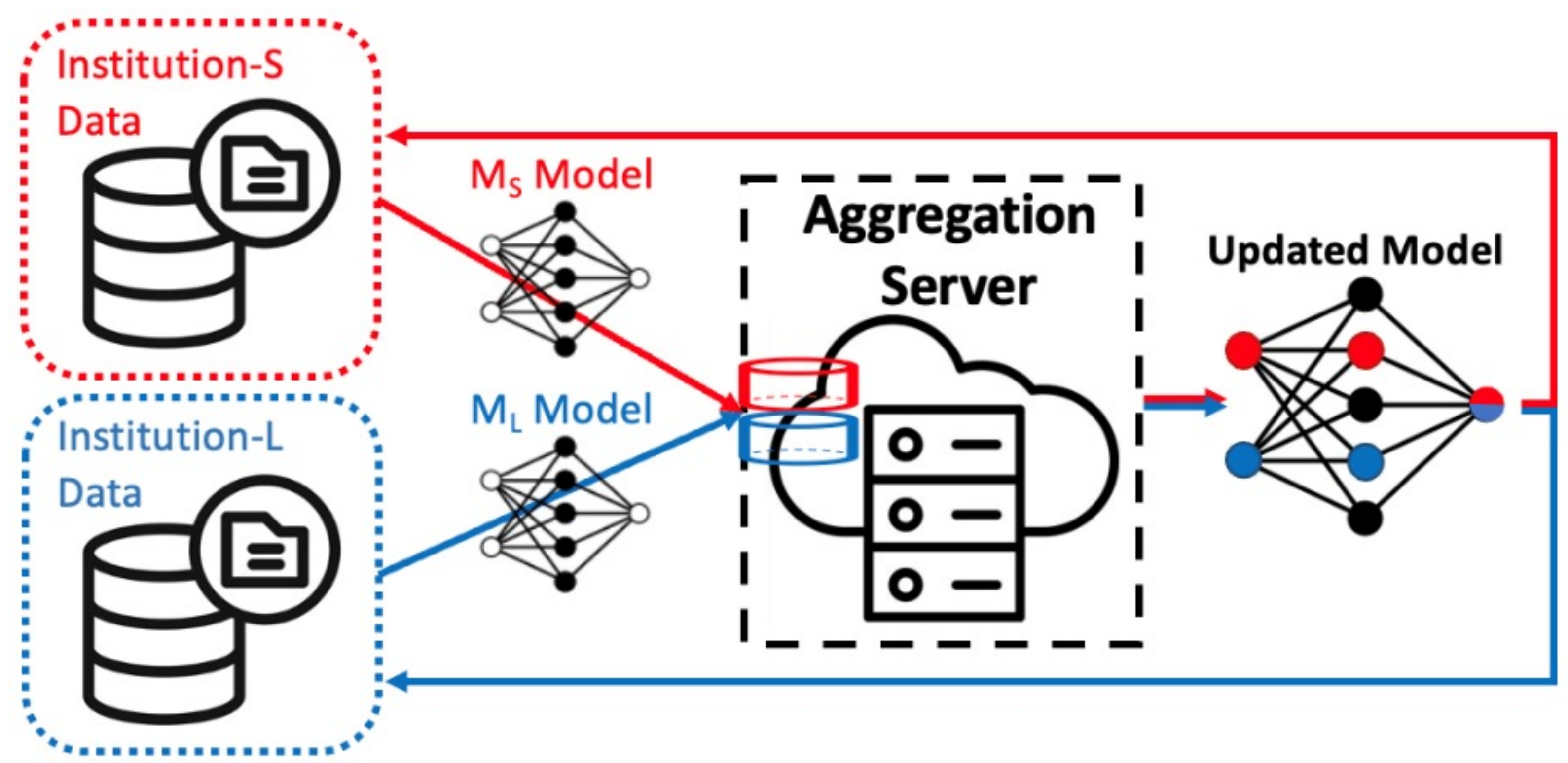

Fig 1. A brief pathway of the proposed FL framework as a 2-client design, allowing up to 10 rounds of communications between aggregation server and local clients.

Results

Fig. 2 shows an example of a BM patient from Institution-S. M_S results missed the lateral BM with false-positive results from blood vessel signals (blue circle). In contrast, M_{FL} correctly identified the lateral BM (red arrow) and eliminated the false-positive region.

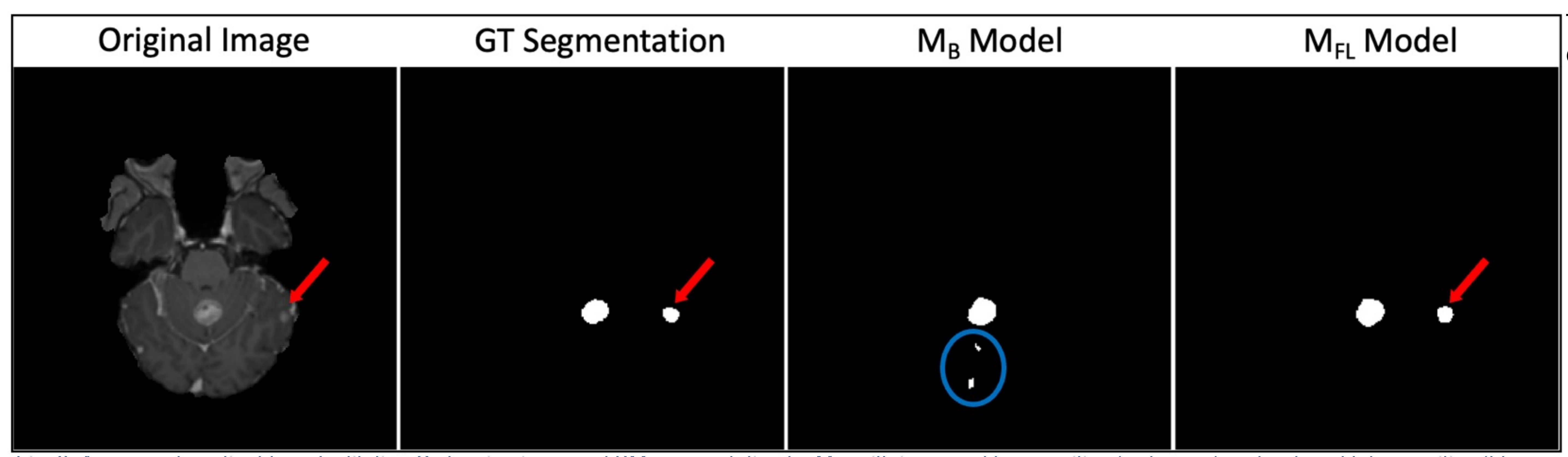

Fig. 2. An example patient from Institution-S showing improved BM segmentation by M_{FL} with improved true-positive (red arrow) and reduced false-positive (blue circle) results.

Table 1 summarizes the segmentation results. MB showed very limited BM segmentation performance, whereas MBL (the 'MB equivalent' at Institution-L, omitted in Abstract due to limit) achieved stronger performance using the large dataset. In contrast, MS and ML improved all evaluation metrics compared to their baseline models (particularly #1 vs #2), highlighting the generalizable benefit of the deep ensemble strategy enabled by the spherical coordinate-based image transformation. MFL further improved the BM segmentation at Institution-S (#5), achieving higher precision and recall by increasing true positives while reducing false negatives. Although ML, trained on the large cohort, achieved strong segmentation accuracy, MFL preserved this performance without meaningful dilution, maintaining high accuracy on Institution-L test cases. Taking together, our results demonstrate the effectiveness of deep ensemble strategy (#1 vs #2, #3 vs #4) and the advantage of ensemble-enabled FL (#2 vs #5, #4 vs #6) in the multi-institutional BM study under inhomogeneous data availability.

#	Model Name	DICE	Precision	Recall	F1
1	M_B	0.193	0.134	0.506	0.184
2	M_S	0.583	0.807	0.727	0.714
3	M_{BL}	0.756	0.674	0.891	0.715
4	M_L	0.767	0.811	0.870	0.820
5	**M_{FL} on M_S test set**	0.641	0.770	0.806	0.724
6	M_{FL} on M_L test set	0.778	0.842	0.835	0.813

Table 1. BM segmentation performance results from all investigated models

Conclusions

The proposed ensemble-enabled FL framework leverages models trained on large datasets to regularize and improve BM segmentation in limited-data settings, providing a scalable, privacy-preserving strategy for multi-institutional collaboration, particularly for sites with limited data availability.

References: [1]Sacks et al. Neuro.Clic.NA 2020; [2]Gu et al. Int J Environ Res Public Health 2023; [3]Pati, S. et al. Nat Comm 2022