Uncertainty-Guided Federated Learning for Robust Multi-Institutional MRI-Based Brain Metastasis Segmentation

Jingtong Zhao[1], Eugene Vaios[1], Zhenyu Yang[2], Chendong Ni[2], John Ginn[1], Deshan Yang[1], Kyle Lafata[1], Dominic LaBella[1], Calabrese[1], Evan Calabrese[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, particularly for small lesions. Although deep learning methods have shown strong potential for automated BM segmentation, model performance is often unstable when trained on limited single-institution datasets, and generalization is challenged by cross-site differences in imaging protocols and patient populations[2]. Meanwhile, patient data privacy concerns limit centralized multi-institutional training, restricting the ability to build robust models from diverse datasets.

Purpose

To develop and evaluate a federated learning (FL) framework for BM segmentation that integrates an uncertainty score into a novel FL objective, improving segmentation robustness and potentially performance when training on limited-size datasets in heterogeneous multi-institutional settings.

Materials/Methods

Two BM datasets (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). At each institution, a customized UNet++ was trained for BM segmentation using test-time augmentation (TTA) to generate multiple predictions per case. These predictions were ensembled to produce a final segmentation and a pixel-wise variation map that quantifies prediction dispersion (uncertainty).

Using federated averaging[3] (Fig.1), the two client UNet++ models were aggregated into a global model. The FL objective was modified by adding an uncertainty-guided penalty term that suppresses high variation-map intensities within the ground-truth BM regions, thereby encouraging stable and confident segmentations. Three models were compared: two local institutional models (M_S,n=90; M_L, n=459) and the global FL model M_{FL}. Performance was assessed using Dice coefficient, precision, recall(i.e., sensitivity) and F1-score. An uncertainty score ([0,1]) was additionally computed to quantify segmentation robustness.

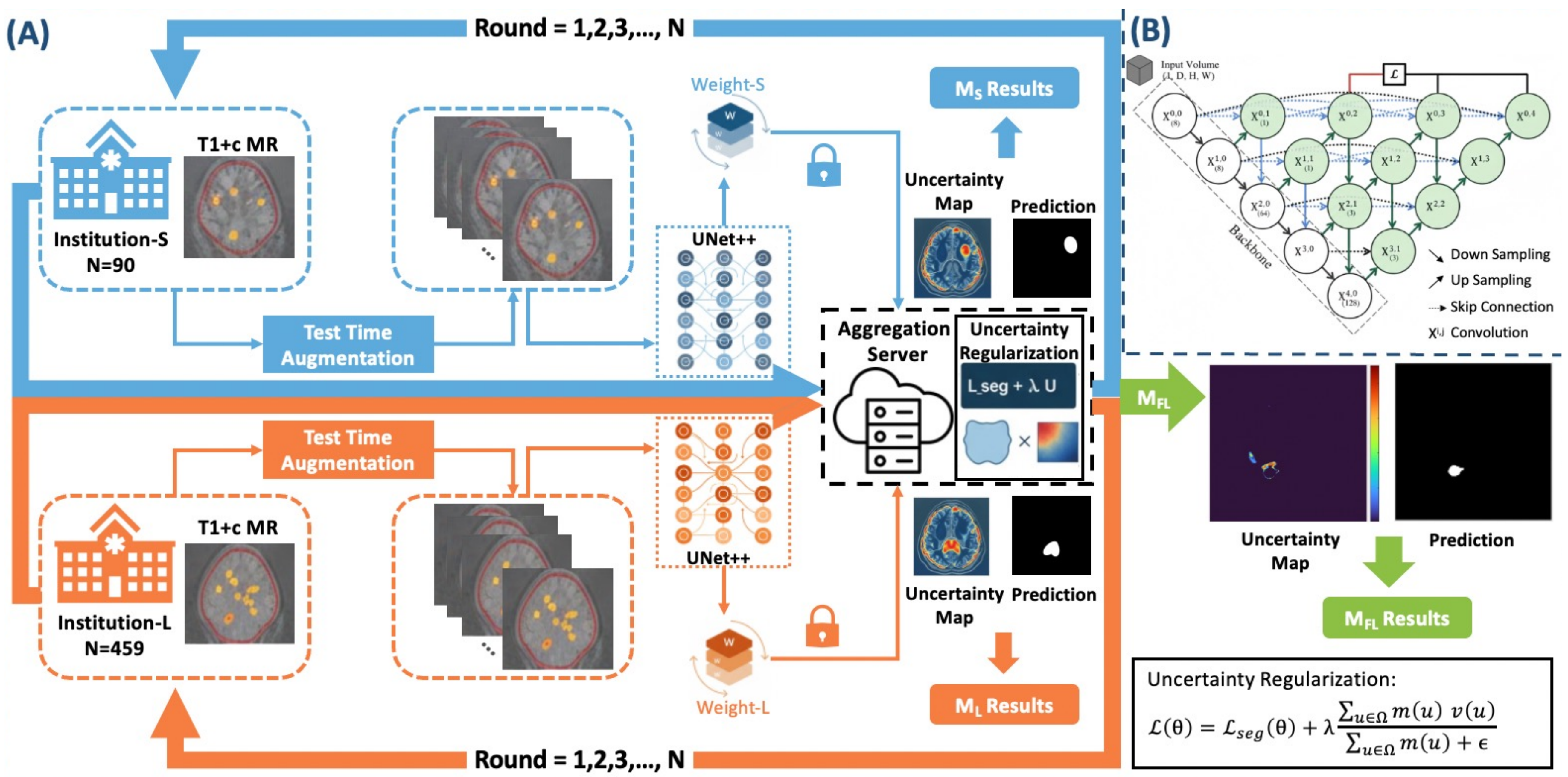

Fig 1. An overall study design of this work. (A) The proposed FL scheme with uncertainty regularization; (B) UNet++ architecture for local segmentation model

Results

Fig. 2 shows an example of a challenging BM segmentation case by M_S with both false-positive and false-negative segmentation results. In contrast, M_{FL} accurately identified all 4 BMs. The uncertainty map exhibited the expected pattern, with higher uncertainty along BM boundaries ('low confidence but thinking hard') and very low uncertainty within the core of the larger BM ('high confidence') in the zoomed view. Notably, low-uncertainty signals outside the ground-truth BM regions were correctly classified as negative, illustrating how M_{FL} leverages uncertainty guidance to improve segmentation robustness.

Table 1 summarizes the segmentation results. M_S showed very limited BM segmentation performance, whereas M_{FL} achieved a marked improvement, particularly through false-positive reduction, reflected by a slightly reduced recall but a substantially higher F1-score. M_L achieved satisfactory performance using Institution-L's large dataset; however, M_{FL} showed reduced performance on Institution-L's test set, as expected because averaging M_L with an underperforming M_S can dilute performance. Taken together, these results suggest that the proposed M_{FL} design can effectively improve BM segmentation when local dataset size is limited. The uncertainty scores for M_{FL} (Institution-1 test: 0.683, Institution-2 test: 0.810) were higher than local models (M_S=0.655 and M_S=0.705), suggesting the enhanced segmentation robustness for both large and small datasets.

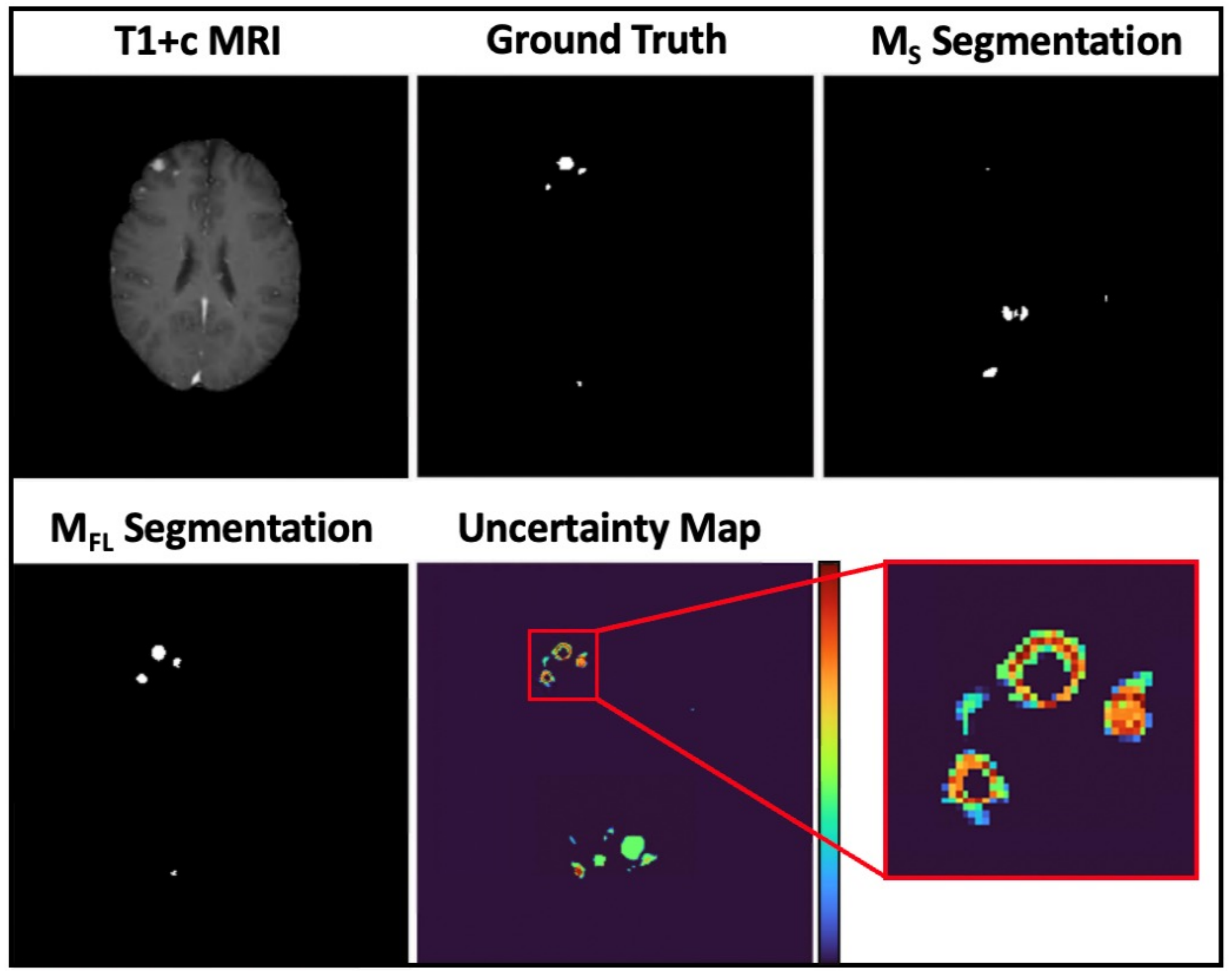

Fig. 2. An example patient from Institution-S showing improved BM segmentation by MFL with the associated uncertainty map

	DICE	Precision	Recall	F1-score
M_S	0.123	0.055	0.864	0.094
M_L	0.763	0.737	0.901	0.781
M_{FL} on M_S test set	0.537	0.643	0.755	0.612
M_{FL} on M_L test set	0.347	0.337	0.757	0.408

Table 1. BM segmentation performances from three studied models

Conclusions

An uncertainty-guided FL framework was developed to enhance robustness and improve BM segmentation performance in limited-data settings. This scalable, privacy-preserving approach enables multi-institutional collaboration with integrated uncertainty benchmarking.

References: [1]Sacks et al. Neuro.Clic.NA 2020; [2]Rauschecker, A. M. et al. Radiology: Artificial Intelligence 2022; [3]Pati, S. et al. Nat Comm 2022