This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Segmentation of Pre- and Post-treatment Brain Metastases Using an Ensemble of MRI Foundation Models

Doye Kim[1⋆], Yeon Gyu Han[2⋆], and Dongheon Lee[3,4†]

[1] Department of Radiology, Seoul National University Hospital, Seoul, Republic of Korea
[2] Department of Biomedical Engineering, Chungnam National University College of Medicine, Daejeon, Republic of Korea
[3] Department of Radiology, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Republic of Korea
[4] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea
the.keramik05@gmail.com; dusrb37@gmail.com; dhlee.jubilee@gmail.com

Abstract. Accurate volumetric assessment of brain metastases is critical for treatment planning and response evaluation, yet manual delineation is time-consuming and subject to inter-observer variability, particularly for small and multifocal lesions. In this work, we address the problem of automated brain metastasis detection and multi-class segmentation, including post-treatment resection cavities, using the nnU-Net framework. We integrate two MRI foundation encoders—Triad (convolutional, used for encoder initialization) and BrainIAC (vision-transformer, used as a frozen auxiliary feature extractor)—through the transfer interface each architecture permits. Representation probing of both frozen encoders guided the design of class-specific fusion weights, and confidence-aware connected-component filtering was applied as post-processing. Our experiments on the BraTS-METS 2026 dataset demonstrate improved segmentation across all four evaluation regions relative to a plain nnU-Net baseline, reaching lesion-wise DSC of 0.703, 0.560, 0.727 and 0.686 for ET, RC, TC and WT. An ablation shows that the contributions are region-dependent: RC case oversampling produced the largest resection-cavity gain, with further increments from the BrainIAC branch and RC-specific fusion, whereas confidence-aware filtering drives the gains in ET, TC and WT.

Keywords: Brain metastases · Brain MRI · Foundation model · Semantic Segmentation · Pre-processing · Post-processing

1 Introduction

Brain metastases are the most common malignant intracranial tumors in adults, occurring in 10–20% of patients with malignant disease [1]. Accurate volumetric

⋆ Equal contribution. † Corresponding author.

delineation supports stereotactic-radiosurgery planning, longitudinal response assessment and quantification of intracranial disease burden, and volumetric assessment has shown greater consistency than diameter-based measurement [10]. Manual delineation is time-consuming, subject to inter-observer variability, and particularly difficult in multifocal disease and small lesions [12]. Post-treatment imaging may additionally contain resection cavities and heterogeneous treatment-related changes. Reproducible automated segmentation is therefore of clinical importance.

nnU-Net [5] has become established as a framework providing strong baseline performance across diverse medical image segmentation tasks without task-specific architectural design, and the winning solution of the BraTS-METS 2025 challenge [2] was obtained by configuring nnU-Net variants differing in loss function, model capacity, training schedule and post-processing. We therefore take a configured nnU-Net as the starting point of this work. Two aspects of the task remain difficult for such a baseline. Detection performance depends on lesion size: a systematic review places pooled lesion-wise DSC at 79% with lesion-wise sensitivity at 87% [12], a baseline nnU-Net reached 74.8% lesion-wise sensitivity for metastases under 3 mm in external testing against 90.8% across all sizes [10], and the analysis of the BraTS-METS 2023 challenge identifies false negatives for small lesions as a common error among the leading teams [8]. Separately, the resection cavity occurs only in post-treatment cases and is therefore the rarest of the four evaluation regions.

Foundation models pretrained on large unlabelled MRI corpora offer complementary prior information beyond the labelled challenge cohort. Triad [11] is a 3D MRI encoder pretrained by masked image modelling and reported to improve segmentation across downstream datasets when used to initialize nnU-Net. BrainIAC [9] is a brain-specific encoder pretrained by contrastive self-supervised learning on 32,015 multiparametric brain MRIs, with gains reported particularly under limited supervision: a mean DSC of 0.72 using 10% of the training data against 0.33 for training from scratch, narrowing to 0.79 versus 0.70 at full data availability.

How these encoders should be attached is not given a priori. Their architectures differ, and each admits a different interface: The released Triad PlainConvUNet checkpoint is topology-compatible with the nnU-Net encoder, whereas the released BrainIAC ViT checkpoint is not; we therefore used direct encoder initialization for Triad and frozen auxiliary-feature integration for BrainIAC. Which regions benefit from which interface is likewise unknown for this task. The segmentation validation of BrainIAC was performed on adult glioma reformulated as single-channel binary segmentation from FLAIR [9], and prior comparisons of brain MRI foundation encoders on BraTS2023-Mets addressed the pre-treatment setting only [13]; neither covers four-region pre- and post-treatment segmentation or the resection cavity.

This paper describes the pipeline submitted to Task 1 of the BraTS-METS 2026 Challenge. Both encoders were integrated into a common nnU-Net framework through the interface each admits (Sect. 3.1), and both frozen representa-

tions were probed to determine which regions each encoder informs (Sect. 3.2). The probing outcome guided a region-specific fusion of the two branches, and the fused scores were refined by size- and confidence-aware connected-component filtering (Sect. 3.3). We report the resulting performance against the nnU-Net baseline together with an ablation over the pipeline.

2 BraTS-METS 2026 Dataset

The dataset was provided by the BraTS-METS 2026 Challenge [6–8] and comprises pre- and post-treatment brain metastasis mpMRI collected retrospectively at nine institutions. The training set contains 1,296 cases from 810 patients with released labels and the validation set 179 cases from 166 patients with labels withheld. One training case was excluded due to a label value outside the four-class schema, leaving 1,295 cases for supervised training.

Each case comprises T1W, T1C, T2W and FLAIR. T2W has not been mandatory since 2025; no missing T2W file was encountered by our pipeline. Images are provided in subject-native space or co-registered to SRI24: 967 of the 1,295 training cases (74.7%) are native and 328 (25.3%) SRI24, whereas the validation set is split at 91 native (50.8%) and 88 SRI24 (49.2%).

Reference labels were produced by STAPLE fusion, neuroradiologist refinement and attending-level approval. Four regions are annotated: non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), enhancing tumor (ET) and resection cavity (RC). RC occurs only in post-treatment cases and is the rarest region.

Evaluation uses four derived regions: ET $= \{3\}$, tumor core TC $= \{1, 3\}$, whole tumor WT $= \{1, 2, 3\}$ and RC $= \{4\}$; RC is scored independently and is not part of WT. Fig. 1 shows the lesion distribution: most cases contain fewer than five ET lesions, while 7.7% contain twenty or more; lesion volumes span five orders of magnitude with a median of 68 voxels.

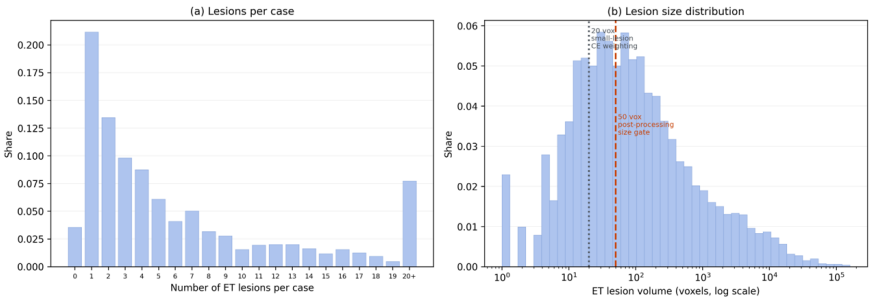

Fig. 1. Distribution of ET lesions in the training set (1,295 cases, 10,112 lesions). (a) Lesions per case; the rightmost bin aggregates ≥20. (b) Lesion volume (log scale); median 68, maximum 160,216 voxels. Dotted line: 20-voxel loss-weighting threshold. Dashed line: 50-voxel post-processing gate.

3 Methods

We adopt the nnU-Net [5] 3D full-resolution configuration and build two variants differing in how a pretrained MRI representation is transferred into it (Fig. 2): Model A initializes the nnU-Net encoder from Triad and fine-tunes it, while Model B supplies frozen BrainIAC features as auxiliary input channels. Both share identical plans: target spacing $1.0 \times 0.8984 \times 0.8594$ mm, per-channel z-score normalization with brain masking, patch size $112 \times 160 \times 128$, batch size 2, channels ordered T1C, T1W, FLAIR, T2W.

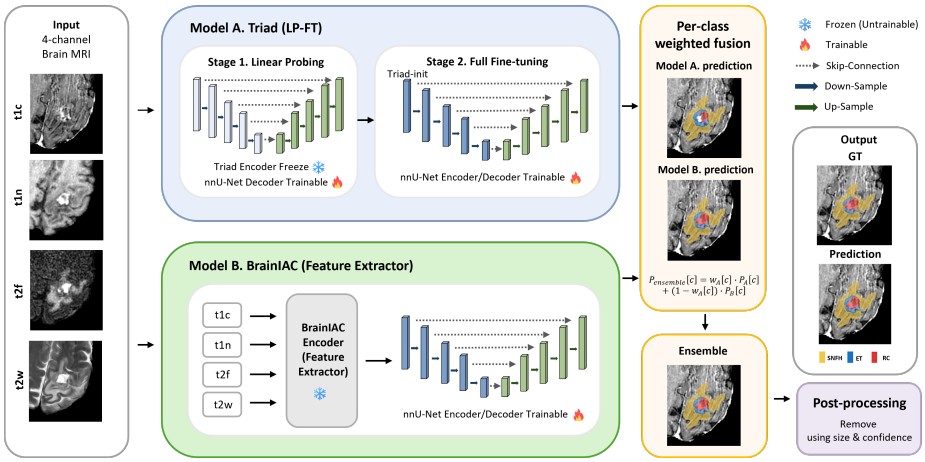

Fig. 2. Pipeline overview. Model A initializes the nnU-Net encoder from Triad and trains in two stages. Model B concatenates frozen BrainIAC features (averaged over modalities, projected to four channels) with the original four channels, forming an eight-channel input. Outputs are fused per class and refined by confidence-aware filtering. The example is a post-treatment case containing a resection cavity.

3.1 Brain MRI Foundation Encoders

Two publicly released self-supervised MRI encoders were used. Filenames and pretraining provenance are listed in the code repository; the container copies were verified to match. Neither checkpoint contains a segmentation head.

Triad [11] was pretrained by masked image modelling on 3D MRI volumes spanning brain, breast and prostate. BrainIAC [9] was pretrained by contrastive learning on 32,015 multiparametric brain MRIs using a 3D ViT-B backbone at 96^3 resolution. The two differ in scope, objective and architecture, and the architectural difference determines the available transfer interface.

Model A: Encoder Initialization. The Triad PlainConvUNet checkpoint matched the nnU-Net encoder topology. All six stages were loaded (32, 64, 128, 256, 320, 320 channels) without modifying the plans. The single-channel first convolution was replicated across four modality channels and divided by four. The decoder was randomly initialized. Training followed a two-stage linear-probing then fine-tuning (LP-FT) schedule: 100 epochs encoder-frozen, then 900 epochs fully trainable. LoRA [4] (rank 8, $\alpha\,8$, stages 3–5) and full fine-tuning were trained for comparison.

Model B: Auxiliary Features. BrainIAC was used fully frozen. Each modality channel is z-scored per patch, resampled to 96^3 and passed through the ViT-B backbone; the class token is discarded and the 216 patch tokens reshaped to $F_m \in \mathbb{R}^{768\times6\times6\times6}$. The four grids are averaged, $\bar{F} = \frac{1}{4}\sum_m F_m$, reduced to four channels by a trainable $1\times1\times1$ convolution, upsampled trilinearly to the patch size and concatenated with the original four channels (eight total). This convolution is the only added module in the auxiliary feature path. The first nnU-Net convolution was widened from four to eight input channels, and the network was trained end to end.

3.2 Representation Probing

The two encoders differ in pretraining corpus and objective, so the regions they inform need not coincide. We probed both frozen representations before designing the fusion.

Linear Probing. A linear classifier was trained on frozen case-level features to predict, for each label and each of two size strata ($<$20 and $\geq$20 mm^3), whether the case contains at least one lesion in that stratum. The strata are non-exclusive. Each label–stratum pair is an independent binary task scored by AUROC (area under the receiver operating characteristic) on the training folds.

Dense Readout. A decoder block was attached to the frozen backbone and trained alone, scored by Dice similarity coefficient (DSC), following the protocol of [9]. MRI-CORE [3] was included as a third encoder in the probing analysis only.

3.3 Training and Inference

Class Imbalance. Both branches use RC case oversampling: cases containing RC receive a sampling weight of 3.0. Model A additionally applies size-dependent voxel weighting to the cross-entropy term: voxels in foreground connected components of at most 20 voxels receive weight 3.0, while all others receive weight 1.0. Model B retains the standard Dice-plus-cross-entropy objective without this weighting. Size-aware voxel weighting has been used previously for this task [10].

Score Fusion. Model A (five folds) and Model B (two folds) each produce softmax probabilities; fold outputs are averaged within each model and combined per class:

$$q_c = w_c\, p_c^A + (1 - w_c)\, p_c^B, \tag{1}$$

with $w_c = 0.50$ for background, NETC, SNFH and ET, and $w_{\mathrm{RC}} = 0.15$. Because each pair of weights sums to one, the fused value remains on a probability scale per class, but the vector does not sum to one across classes; no renormalization is applied. The RC-specific weight was motivated by probing and calibrated by grid search on the two Model B folds. Folds 1 and 3 were selected on internal cross-validation; the fold count was fixed at two by the inference time budget.

Post-processing. The argmax label map is decomposed into 26-connectivity connected components per class. Each component is assigned a confidence score: the mean of its per-voxel maximum fused value. A component is removed only if its size is below 50 voxels *and* its confidence is below a class threshold (0.6 for NETC, SNFH, ET; 0.5 for RC). No component is removed on size alone. Fig. 3 shows that false positives concentrate at small sizes and are more prevalent than true positives in the low-confidence range; however, the two distributions overlap substantially, which is why both criteria are required jointly. RC shows greater overlap, and its lower threshold limits true-positive removal. The size gate was selected from candidates 20, 35, 50, 100 as the value at which DSC saturated. This approach follows [10].

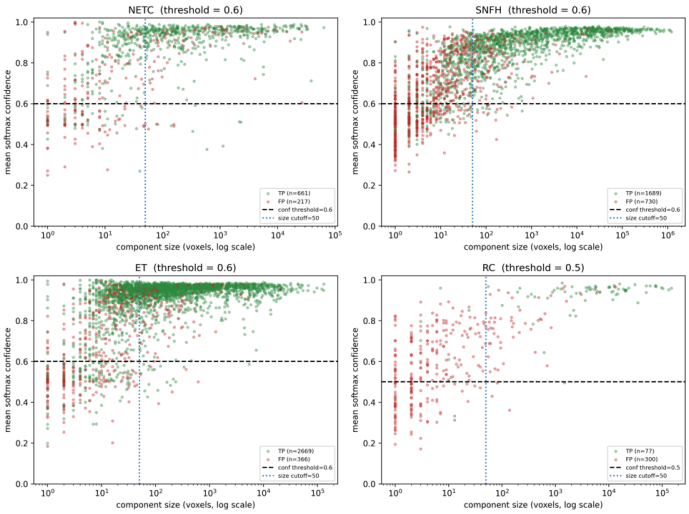

Fig. 3. Component size vs. mean confidence for true and false positives on the calibration folds, per foreground class. (TP: $\geq$10% reference coverage; FP: zero reference overlap). Dotted: 50-voxel gate. Dashed: confidence threshold (0.6 for NETC, SNFH, ET; 0.5 for RC). A component is removed only below both lines.

4 Experiments and Results

Training Details. All training was performed on a single NVIDIA H200 GPU using nnU-Net v2.7.0. Optimization: SGD, Nesterov momentum 0.99, learning rate 10^{-2}, weight decay 3×10^{-5}, polynomial decay (exponent 0.9); default augmentation. Model A: 100+900 epochs × 5 folds; Model B: 1,000 epochs × 2 folds; total ≈128 GPU-hours. Final-epoch checkpoints, seed 12345, no TTA; inference 19.6 s per case. Synapse team: Bio-Intel.

Code: `https://github.com/thekeramik05/MICCAI-METS_challenge_bi-intel`.

Metrics: lesion-wise DSC and normalized surface distance (NSD) (FP/FN contribute zero); small-instance F1 for lesions <27 mm^3. All validation values are from the challenge evaluation server. Comparisons are on mean metrics and are not intended to reproduce the challenge's final ranking procedure.

Fine-tuning Comparison. Table 1 compares three strategies for Model A. LoRA was lowest in all regions. LP-FT was retained on the leaderboard.

Table 1. Fine-tuning strategies for Model A on validation set.

	ET		RC		TC		WT	
	DSC	NSD	DSC	NSD	DSC	NSD	DSC	NSD
Full FT	.672	.731	.559	.469	.701	.748	.657	.679
LoRA	.647	.707	.457	.361	.677	.723	.638	.657
LP-FT	.672	.733	.521	.438	.703	.751	.666	.687

DSC: Dice similarity coefficient. NSD: normalized surface distance. ET: enhancing tumor. RC: resection cavity. TC: tumor core. WT: whole tumor.

Representation Analysis. Table 2 reports probing results averaged across labels. Case-level AUROC was 0.62–0.64 below 20 mm^3 and 0.81–0.84 above, across all three encoders. Dense readout gave mean DSC ≈0.25. When examined per label, the clearest difference appeared for RC: BrainIAC AUROC was 0.94 for RC presence vs. 0.60–0.70 for other labels, while dense-readout DSC for RC was $< 10^{-4}$. This motivated the class-dependent fusion weight of Sect. 3.3.

Pipeline Ablation. Table 3 reports a component-wise analysis of the pipeline on the validation set. RC oversampling raised RC DSC from 0.447 to 0.532. Triad initialization gave 0.672, 0.521, 0.703 and 0.666 for ET, RC, TC and WT; BrainIAC features gave 0.674, 0.551, 0.703 and 0.663. Class-wise fusion changed DSC by +0.004, +0.009, +0.001, −0.002 relative to the stronger branch per region. Confidence filtering added 0.025, 0.000, 0.023 and 0.022 for ET, RC, TC and WT.

Table 2. Representation probing on training folds. AUROC for size-specific lesion presence (averaged across labels) and mean DSC of the frozen-encoder dense readout.

Encoder	AUROC ($<20\,\mathrm{mm}^3$)	AUROC ($\geq 20\,\mathrm{mm}^3$)	Dense DSC
Triad [11]	0.671	0.825	0.246
BrainIAC [9]	0.660	0.798	0.247
MRI-CORE [3]	0.637	0.791	–

AUROC: area under the receiver operating characteristic.

The final pipeline reached DSC 0.703, 0.560, 0.727 and 0.686 and NSD 0.765, 0.485, 0.775 and 0.706 (Table 3, last row), improving over the baseline by 0.032, 0.113, 0.030 and 0.032 DSC. On the other hand, training set yielded the following performance: 0.734, 0.592, 0.761, 0.770 for ET, RC, TC and WT.

Table 3. Pipeline ablation on validation set.

	ET		RC		TC		WT	
	DSC	NSD	DSC	NSD	DSC	NSD	DSC	NSD
nnU-Net baseline	.671	.733	.447	.350	.697	.746	.654	.676
+ RC oversamp.	.678	.740	.532	.444	.707	.756	.669	.689
Model A (Triad [11])	.672	.733	.521	.438	.703	.751	.666	.687
Model B (BrainIAC [9])	.674	.735	.551	.476	.703	.751	.663	.685
+ fusion	.678	.739	.560	.485	.704	.752	.664	.684
+ conf. filtering	.703	.765	.560	.485	.727	.775	.686	.706

ET: enhancing tumor. RC: resection cavity. TC: tumor core. WT: whole tumor. DSC: Dice similarity coefficient. NSD: normalized surface distance.

Small-instance Detection. Table 4 reports small-instance F1 for lesions below 27 mm^3. Confidence filtering raised F1 from 0.412 to 0.426 (ET), 0.421 to 0.433 (TC) and 0.314 to 0.326 (WT), while reducing ET false positives from 1.00 to 0.587 per case (true positives: from 1.813 to 1.800). In the labelled training set, 62 of 253 RC components (24.5%) fall below 27 mm^3 (median 0.60 mm^3); on the validation server none were detected. Fig. 4 shows predictions for four training cases.

Table 4. Small-instance F1 for lesions $<27\,\mathrm{mm}^3$ on validation set.

	ET	TC	WT	RC
Fusion	0.412	0.421	0.314	0.000
+ conf. filtering	0.426	0.433	0.326	0.000

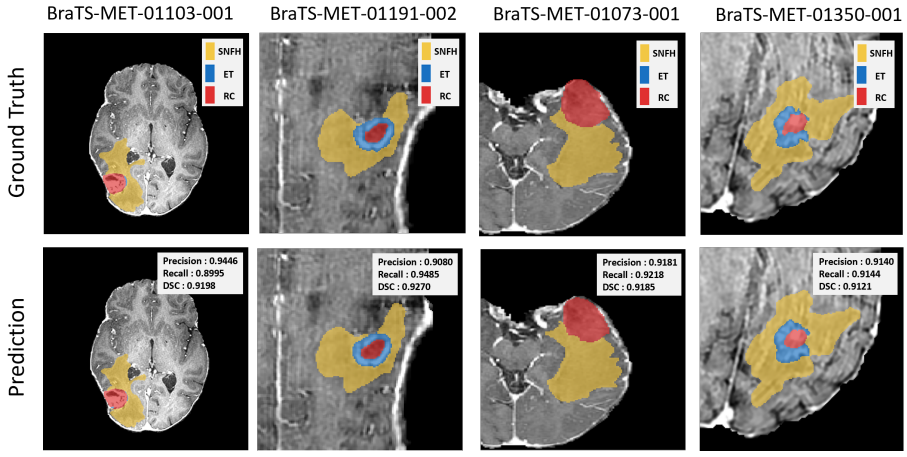

Fig. 4. Ground truth and predictions for four post-treatment training cases containing resection cavities, with case-level voxel-wise precision, recall and DSC (distinct from lesion-wise metrics in Table 3).

5 Discussion

Main Findings. The complete pipeline improved lesion-wise DSC in all four regions over the plain nnU-Net baseline. The ablation shows region-dependent mechanisms: RC case oversampling produced the largest RC increase, from 0.447 to 0.532, with further increments from the BrainIAC branch and class-wise fusion (to 0.560), whereas confidence-aware filtering drove most of the ET, TC and WT improvement.

Fusion Design. BrainIAC showed the strongest case-level signal for RC (AUROC 0.94 vs. 0.60–0.70 for other labels), which guided the RC-specific fusion weight. However, RC occurs only in post-treatment cases, so this separability may partly reflect treatment status; the analysis cannot distinguish the two. Dense readout for RC was $< 10^{-4}$ DSC, indicating that case-level separability did not transfer to dense segmentation under the evaluated interface. In training-set cross-validation, 87.8% of undetected small RC components co-occurred with an RC prediction elsewhere in the same case, indicating a localization failure rather than absence of model response to the cavity. Combining these results, these results show that RC segmentation is not yet reliable enough for clinical use.

Limitations. Only point estimates ar reported; the fusion-stage improvements of 0.001–0.009 DSC did not reach statistical significance under paired bootstrap testing. Moreover, Fusion weights and filtering thresholds were calibrated on two Model B folds, and fine-tuning strategy was selected on the leaderboard; the

hidden test set may therefore yield lower performance. The final system combines seven models against a five-model baseline, so the reported improvement does not fully disentangle foundation-encoder pretraining from ensemble size and training budget. Cross-validation splits were case-level, so the same patient may appear in different folds.;however, excluding such csases changed DSC by less than 0.004 in all regions with no consistent direction The BrainIAC branch was not controlled for input dimensionality separately from the pretrained representation.; a single-fold comparison against randomly initialized features did not reach statistical significance on the 41 available RC cases.

Conclusion. We integrated two MRI foundation encoders into nnU-Net through the interface each architecture admits, used probing to guide region-specific fusion, and applied confidence-aware filtering. The pipeline improved all four regions of BraTS-METS 2026 Task 1, with the largest gain for the rare resection cavity. Foundation-encoder transfer and task-specific optimization made complementary, region-dependent contributions.

Acknowledgments. Data used in this publication were obtained as part of the BraTS 2026 Challenge project through Synapse ID (syn74274097). This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]. This research was also supported by the Advanced GPU Utilization Support Program funded by the Government of the Republic of Korea (Ministry of Science and ICT).

Disclosure of Interests. The authors have no competing interests to declare.

Disclosure of AI Assistance Large language model-based tools were used to assist with language editing, under the authors' supervision. All scientific content, analysis, and conclusions are the responsibility of the authors.

References

1. Achrol, A.S., Rennert, R.C., Anders, C., et al.: Brain metastases. Nature Reviews Disease Primers **5**(1), 5 (2019). https://doi.org/10.1038/s41572-018-0055-y
2. Bancerek, M., Rudzki, P., Nalepa, J.: Segmentation of pre and post-treatment brain metastases using nnu-nets. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. LNCS, vol. 16376, pp. 161–172. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16365-3_15
3. Dong, H., Chen, Y., Gu, H., et al.: MRI-CORE: A foundation model for magnetic resonance imaging (2025). https://doi.org/10.48550/arXiv.2506.12186
4. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9

5. Isensee, F., Jaeger, P.F., Kohl, S.A.A., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
6. Karargyris, A., Umeton, R., Sheller, M.J., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence **5**(7), 799–810 (Jul 2023). https://doi.org/10.1038/s42256-023-00652-2
7. Maleki, N., Amiruddin, R., Moawad, A.W., et al.: Analysis of the miccai brain tumor segmentation – metastases (brats-mets) 2025 lighthouse challenge: Brain metastasis segmentation on pre- and post-treatment mri (2025), https://arxiv.org/abs/2504.12527
8. Moawad, A.W., Janas, A., Baid, U., et al.: The brain tumor segmentation – metastases (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI (Dec 2024), https://arxiv.org/abs/2306.00838, arXiv:2306.00838v3, PMID: 37396600, PMCID: PMC10312806
9. Tak, D., Garomsa, B.A., Zapaishchykova, A., et al.: A generalizable foundation model for analysis of human brain MRI. Nature Neuroscience **29**(4), 945–956 (Apr 2026). https://doi.org/10.1038/s41593-026-02202-6
10. Topff, L., Petrychenko, L., Jain, N., et al.: A data-centric approach to deep learning for brain metastasis analysis at MRI. Radiology **315**(3), e242416 (2025). https://doi.org/10.1148/radiol.242416
11. Wang, S., Safari, M., Li, Q., et al.: Vision foundation model for 3d magnetic resonance imaging segmentation, classification, and registration. Medical Image Analysis **110**, 103992 (2026). https://doi.org/10.1016/j.media.2026.103992, https://www.sciencedirect.com/science/article/pii/S1361841526000617
12. Wang, T.W., Hsu, M.S., Lee, W.K., et al.: Brain metastasis tumor segmentation and detection using deep learning algorithms: A systematic review and meta-analysis. Radiotherapy and Oncology **190**, 110007 (2024). https://doi.org/10.1016/j.radonc.2023.110007, https://www.sciencedirect.com/science/article/pii/S0167814023093143
13. Wu, Y., Wang, S., Li, Y., et al.: BrainDINO: A brain MRI foundation model for generalizable clinical representation learning (2026). https://doi.org/10.48550/arXiv.2604.27277