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.

Region-Aligned Training and Target-Specific 2D–3D nnU-Net Fusion for BraTS 2026 Brain Metastasis Segmentation

Amirmahdi Mojtahedzadeh[1,2], Oluwatobi Iyanuoluwa Akinmuleya[4], Jeremiah Oluwatomi Itodo Daniel[3], Maxwell Adinavi[5]

[1] Faculty of Medicine, Semmelweis University, Budapest, Hungary
[2] Center of Neurosurgery and Neurointervention, Semmelweis University, Budapest, Hungary
[3] College of Health Sciences, Obafemi Awolowo University, Ile-Ife, Nigeria
[4] Shenyang Medical College, Liaoning, China
[5] Harbin Institute of Technology, Shenzhen, China
amirmahdi.mojtahedzadeh@stud.semmelweis.hu

Abstract. Automated brain-metastasis segmentation is complicated by lesion multiplicity, small volumes, heterogeneous MRI, postoperative anatomy, and overlapping evaluation regions. We evaluated a two-phase BraTS 2026 Task 1 framework using organizer-provided data. Excluding one mask with an uncorrected invalid label left 1,295 development cases; official validation comprised 179 unlabeled cases. Five-fold 2D and 3D full-resolution nnU-Net models were trained with mutually exclusive class labels and then overlapping whole-tumor, tumor-core, enhancing-tumor, and resection-cavity targets using identical images, plans, and partitions. On the 1,295-case development cohort, pooled five-fold out-of-fold RC50 mean DSC, NSD, and small-lesion F1 were 0.5605, 0.5788, and 0.3064; on the 179-case official validation cohort, the corresponding means were 0.6456, 0.6555, and 0.3153. Within region-based out-of-fold analysis, RC50 versus global 50/50 fusion changed mean DSC by +0.0050, mean NSD by −0.0010, and mean small-lesion F1 by +0.0323. Relative to the class-based equal-fusion system, observed mean differences were +0.0588 in DSC, +0.0451 in NSD, and +0.0659 in small-lesion F1. Because target formulation and fusion both changed, the official comparison does not isolate either component. The method used no external data, test-time augmentation, or component postprocessing and was packaged offline.

Keywords: Brain metastases, Medical image segmentation, Magnetic resonance imaging, nnU-Net, Region-based training, Small-lesion detection, BraTS

1 Introduction

Brain metastases are a common manifestation of systemic malignancy and frequently require serial magnetic resonance imaging (MRI) for diagnosis, treatment planning, response assessment, and post-treatment surveillance. The Response Assessment in

Neuro-Oncology Brain Metastases criteria standardized conventional response evaluation through measurements of selected target lesions [1]. Automated volumetric delineation may provide complementary information when lesions are irregular, multifocal, or altered by surgery, radiotherapy, or systemic treatment.

Manual contouring is time-consuming, particularly in patients with numerous lesions. It is also affected by differences in lesion identification, boundary placement, interpretation of treatment-related enhancement, perilesional edema, and postoperative cavities. Automated segmentation may reduce this burden and provide reproducible volumetric measurements, but brain metastases vary substantially in size, morphology, contrast, anatomical location, and treatment status. Small enhancing metastases may occupy only a few voxels and can resemble vessels, normal enhancement, hemorrhage, or imaging artifacts.

BraTS-METS challenge reports have highlighted that brain metastasis analysis is simultaneously a semantic-segmentation and lesion-detection problem [2, 3]. Performance deteriorates as lesion volume decreases, and a method can obtain favorable overlap for a large lesion while missing several smaller but clinically relevant metastases [2, 3]. The 2025 challenge extended the task to pre- and post-treatment imaging, introducing additional variability from resection cavities and treatment-related changes [2]. Task 1 of BraTS 2026 continues this setting and evaluates whole tumor (WT), tumor core (TC), enhancing tumor (ET), and resection cavity (RC) in heterogeneous multi-institutional MRI [4].

The distinction between the supplied annotation labels and the evaluated regions is methodologically important. The reference map contains mutually exclusive labels for nonenhancing tumor core, surrounding nonenhancing FLAIR hyperintensity, enhancing tumor, and resection cavity. In contrast, WT and TC are overlapping unions of these labels. Consequently, a single ET error can influence WT, TC, and ET simultaneously. Validation guidelines emphasize that target representation and metrics should reflect the structural and clinical properties of the task [5, 6].

nnU-Net is a self-configuring biomedical segmentation framework that adapts preprocessing, network topology, patch size, batch size, training, and inference to the dataset [7]. Carefully configured convolutional segmentation systems remain highly competitive under rigorous validation protocols [8]. The evaluated nnU-Net configurations belong to the U-Net encoder–decoder family [7, 9]. Importantly, nnU-Net supports region-based training for BraTS-type tasks in which the evaluated structures differ from mutually exclusive integer labels. Overlapping regions are learned directly and converted back to an integer segmentation using a predefined overwrite order [10]. Related BraTS work has investigated small-lesion-aware volumetric segmentation and 3D transformer-based U-Net variants [11, 12], while the original multimodal BraTS benchmark established the broader evaluation lineage [13].

Using identical cases, preprocessing, plans, and partitions, we compared the complete region-based RC50 system with the class-based equal-fusion system on official validation, examined fusion candidates in region-based out-of-fold predictions, and functionally verified challenge packaging. A matched official region-based equal-fusion arm and direct 2D-versus-3D disagreement analysis were unavailable. The study followed the organizer's Task 1 specification [14].

2 Methods

2.1 Data, Labels, and Quality Control

The method was developed for Task 1 of the MICCAI BraTS 2026 Challenge [14] using only organizer-provided imaging data and no externally pretrained segmentation weights. Organizer-issued corrected masks replaced matching originals. Of 1,296 nominal development cases, BraTS-MET-01094-002 was excluded because its mask contained invalid label 6 and no matching official correction was available when the dataset was frozen; 1,295 cases remained. The official validation cohort contained 179 cases without public reference annotations.

Four MRI channels were used: post-contrast T1-weighted MRI (T1C), pre-contrast T1-weighted MRI (T1N), T2-weighted FLAIR (T2F), and T2-weighted MRI (T2W), mapped to nnU-Net channels 0000–0003, respectively. The reference labels were background (0), nonenhancing tumor core (NETC, 1), surrounding nonenhancing FLAIR hyperintensity (SNFH, 2), enhancing tumor (ET, 3), and resection cavity (RC, 4). The challenge-oriented regions were:

$$\mathrm{WT} = \{1,2,3\},\ \mathrm{TC} = \{1,3\},\ \mathrm{ET} = \{3\},\ \mathrm{RC} = \{4\}. \tag{1}$$

The preprocessing audit required valid identifiers, four readable MRI modalities, one readable segmentation, compatible within-case geometry, and labels in {0, 1, 2, 3, 4}; no repair was performed beyond organizer corrections. The prediction audit verified complete coverage, unique/readable NIfTI files, matching geometry, labels 0–4, and a flat output directory without assuming uniform dimensions or orientation.

2.2 Preprocessing and Model Training

Preprocessing and model training used nnU-Net v2.8.1. Dataset fingerprinting and planning automatically selected normalization, target spacing, patch dimensions, batch size, and network depth. No additional registration, skull stripping, handcrafted intensity normalization, or external anatomical preprocessing was introduced. Predictions were resampled to the original image geometry during export.

Both 2D and 3D full-resolution configurations used the standard `nnUNetTrainer`. Each fold was trained for 1,000 epochs using stochastic gradient descent with an initial learning rate of 0.01, momentum of 0.99, Nesterov acceleration, weight decay of (3×10^{-5}), polynomial learning-rate decay, deep supervision, mixed-precision training, and standard nnU-Net spatial and intensity augmentation. The final checkpoint was used for inference.

A seeded five-fold cross-validation split was preserved across all experiments. Corresponding class-based and region-based folds therefore contained exactly the same training and validation cases. Five models were trained for each of four configurations: class-based 2D, class-based 3D full resolution, region-based 2D, and region-based 3D full resolution. All 20 folds completed the planned training schedule on NVIDIA A100 80-GB and H100 80-GB GPUs.

Table 1. Dataset, training, fusion, and deployment configuration

Characteristic	Final setting
Development / official validation cases	1,295 / 179
MRI channels	T1C, T1N, T2F, T2W
Configurations	2D and 3D full resolution
Cross-validation	Five folds; shared splits
Epochs per fold	1,000
Class-based objective	Soft Dice + multiclass cross-entropy
Region-based objective	Soft Dice + binary cross-entropy
Final 3D weights, WT/TC/ET/RC	0.65 / 0.50 / 0.75 / 0.50
Probability threshold	(> 0.50) for all regions
Test-time augmentation	Disabled
Component postprocessing	None
Final ensemble	Five 2D + five 3D checkpoints
Deployment	Offline Linux/AMD64 Docker

2.3 Class-Based Baseline and Region-Aligned Reformulation

In the baseline formulation, NETC, SNFH, ET, and RC were treated as mutually exclusive semantic classes. The standard nnU-Net objective combined memory-efficient soft Dice loss and multiclass cross-entropy loss. Five-fold probabilities were averaged separately for the 2D and 3D configurations, and the submitted baseline used equal probability averaging between the two configurations.

The region-aligned formulation replaced the class targets with overlapping WT, TC, ET, and RC outputs. The dataset specification defined WT as labels 1, 2, and 3; TC as labels 1 and 3; ET as label 3; and RC as label 4. The reconstruction order was $([2,1,3,4])$. WT-positive voxels were first assigned label 2, followed by TC-positive voxels assigned label 1, ET-positive voxels assigned label 3, and RC-positive voxels assigned label 4. Later regions therefore overwrote earlier encompassing regions, as documented for nnU-Net region-based training [10]. The overlapping outputs were optimized using a compound soft Dice and binary cross-entropy objective with sigmoid activation.

For either 2D or 3D configuration, fold probabilities were averaged according to:

$$P_r^m(x) = \frac{1}{5}\sum_{k=1}^{5} P_{r,k}^m(x), \quad m \in \{2D, 3D\}. \tag{2}$$

Here, $(P_{r,k}^{m}(x))$ denotes the probability assigned to region (r) at voxel (x) by fold (k) of configuration (m).

2.4 Probability Fusion and Exploratory Candidate Analysis

Equal 2D–3D averaging was treated as a reference rather than an assumed optimum. For each region, the fused probability was defined as:

$$P_r^{\text{fusion}}(x) = \alpha_r P_r^{3D}(x) + (1 - \alpha_r) P_r^{2D}(x). \quad (3)$$

Development-cohort performance was estimated exclusively from pooled five-fold out-of-fold (OOF) predictions. Each of the 1,295 cases was predicted once using the corresponding 2D and 3D models whose training folds excluded that case. The predictions were evaluated with the official implementation using a 27-mm³ small-instance threshold and 0.1 overlap threshold. The exploratory record documented no prespecified scalar rule or tie-breaker. The 12-metric mean was the unweighted mean of DSC, NSD, and small-lesion F1 for WT, TC, ET, and RC; it was descriptive, not the official ranking score. RC50 was carried forward because it had the largest descriptive 12-metric mean among the five candidates, without implying a statistically established optimum.

The target-specific candidates fixed WT/TC/ET 3D weights at 0.65/0.50/0.75 and varied RC from 0.35 to 0.50; thus they do not establish optimal WT/TC/ET weights. All used strict $p > 0.50$, with no documented threshold search. Regions were reconstructed in fixed order without component filtering, volume removal, rescue rules, or anatomical heuristics.

2.5 Inference, Evaluation, and Containerization

Inference used `checkpoint_final.pth` for all ten region-based models. Test-time augmentation was disabled. The final sequence comprised five-fold 3D prediction with probability export, five-fold 2D prediction with probability export, channel-wise RC50 fusion, thresholding, sequential class reconstruction, restoration to source geometry, and export of one segmentation per case.

Official validation was performed through the BraTS 2026 Synapse evaluation queue. We report lesion-wise DSC, NSD, and lesion-detection F1 for the relevant lesion-size categories. Results were interpreted jointly because favorable overlap for a large lesion can coexist with missed small instances [5, 6].

The frozen method was packaged in an offline Linux/AMD64 Docker image based on PyTorch 2.8.0 with CUDA 12.8 and nnU-Net 2.8.1. The container included all ten checkpoints and required dependencies, discovered cases recursively from a read-only `/input` mount, and wrote one flat NIfTI output per case to `/output`. No network access was required. Output verification was executed before packaging and after validation inference.

A public implementation, including inference code, Docker configuration, frozen fusion settings, and reproducibility documentation, is available at https://github.com/BrainExeBrats2026/brain-exe-brats2026-task1. The repository corresponds to the frozen RC50 inference and packaging workflow described in this manuscript and will be updated with the final bibliographic citation when proceedings metadata are issued. Challenge imaging data, reference annotations, hidden-test data, and trained checkpoints are not redistributed.

3 Results

3.1 Exploratory Out-of-Fold Candidate Analysis

Across the 1,295-case labeled development (training) cohort, pooled five-fold OOF RC50 performance was 0.5605 mean DSC, 0.5788 mean NSD, and 0.3064 mean small-lesion F1. These held-out OOF values quantify development-cohort performance without reporting optimistic in-sample training scores. Table 2 summarizes the exploratory region-based comparison. RC50 had the largest 12-metric mean (0.4819). Versus global 50/50, it changed mean DSC by +0.0050 (0.5555 to 0.5605), mean NSD by −0.0010 (0.5798 to 0.5788), small-lesion F1 by +0.0323 (0.2741 to 0.3064), and the 12-metric mean by +0.0121. This isolates fusion, not target formulation. Only pooled aggregates were retained, precluding paired intervals or tests; differences are descriptive.

RC50 OOF DSC values for WT/TC/ET/RC were 0.5851/0.6113/0.6003/0.4454; NSD values were 0.6069/0.6728/0.6722/0.3633; and small-lesion F1 values were 0.3749/0.4198/0.4058/0.0250. These are held-out internal cross-validation estimates, not in-sample performance on cases used to fit a model. The sparse discrete search and absent resampling uncertainty do not establish a smooth or stable optimum.

Table 2. Development-cohort five-fold out-of-fold comparison of region-based probability-fusion candidates

Candidate	3D weights WT/TC/ET/RC	Mean DSC	Mean NSD	Mean small F1	12-metric mean
Global 50/50	0.50/0.50/0.50/0.50	0.5555	0.5798	0.2741	0.4698
RC35	0.65/0.50/0.75/0.35	0.5424	0.5658	0.3082	0.4721
RC40	0.65/0.50/0.75/0.40	0.5430	0.5663	0.3061	0.4718
RC45	0.65/0.50/0.75/0.45	0.5446	0.5679	0.3061	0.4729
RC50	**0.65/0.50/0.75/0.50**	**0.5605**	**0.5788**	**0.3064**	**0.4819**

3.2 Official Validation Performance

On the 179-case official validation cohort, RC50 regional DSC values for WT/TC/ET/RC were 0.6532/0.6975/0.6821/0.5494 (mean, 0.6456); NSD values were 0.6754/0.7472/0.7436/0.4558 (mean, 0.6555); and small-lesion F1 values were 0.3651/0.4483/0.4480/0.0000 (mean, 0.3153). The complete RC50 system had higher values than the class-based equal-fusion system for all eight displayed regional DSC/NSD outcomes. Observed mean differences were +0.0588 in DSC, +0.0451 in NSD, and +0.0659 in small-lesion F1. Because both target formulation and fusion differ, these values describe the combined system change and do not isolate either component.

Table 3. Official-validation comparison of the class-based equal-fusion system and complete region-based RC50 system

Region	Metric	Class-based equal	Final RC50	Change
WT	DSC	0.6234	0.6532	+0.0298
WT	NSD	0.6455	0.6754	+0.0299
WT	Small F1	0.2752	0.3651	+0.0899
TC	DSC	0.6639	0.6975	+0.0336
TC	NSD	0.7154	0.7472	+0.0318
TC	Small F1	0.3660	0.4483	+0.0823
ET	DSC	0.6375	0.6821	+0.0446
ET	NSD	0.7056	0.7436	+0.0380
ET	Small F1	0.3565	0.4480	+0.0915
RC	DSC	0.4224	0.5494	+0.1271
RC	NSD	0.3753	0.4558	+0.0805
RC	Small F1	0.0000	0.0000	0.0000

The largest observed DSC difference was for RC (0.4224 to 0.5494), followed by ET (0.6375 to 0.6821). Small-lesion F1 differences were +0.0899 for WT, +0.0823 for TC, and +0.0915 for ET; small-RC F1 was 0.0000 for both systems. The unweighted descriptive mean across the 12 displayed values changed from 0.4822 to 0.5388 (11.7%), but this was not the official ranking formula and no statistical significance is implied.

RC50 all-instance F1 values for WT/TC/ET/RC were 0.7562/0.7737/0.7667/0.5690; large-instance values were 0.7504/0.7467/0.7441/0.0955. RC small-, all-, and large-instance F1 values were 0.0000/0.5690/0.0955, but unavailable size-category counts preclude mechanistic interpretation. Official-minus-OOF DSC differences were +0.0681/+0.0862/+0.0818/+0.1040 for WT/TC/ET/RC. The disjoint cohorts and

unavailable official case-level data prevented explanation of this pattern. Fig. 1 summarizes the workflow and results.

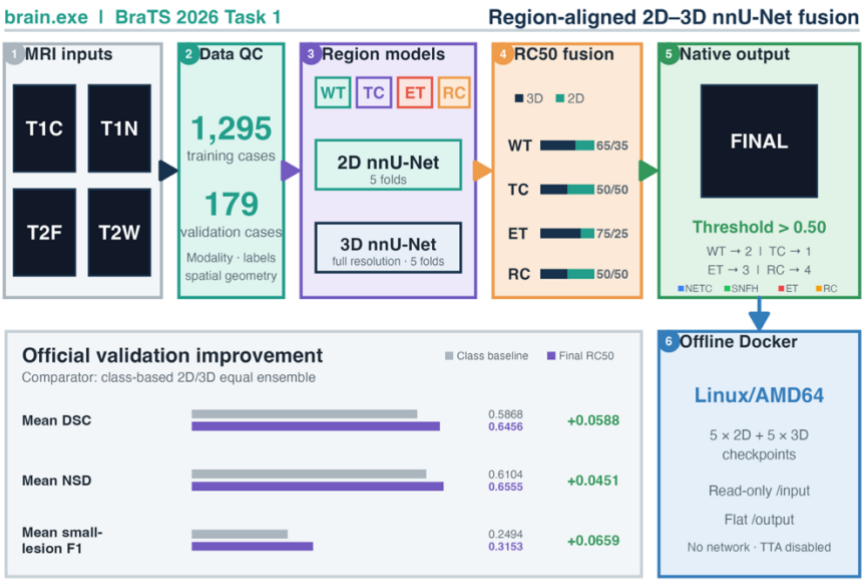

Fig. 1. Workflow and observed official-validation differences for complete region-based RC50 versus the class-based equal-fusion system. The lower-panel label denotes arithmetic differences only; the comparison changes both target formulation and fusion, and no statistical significance is implied. WT whole tumor, TC tumor core, ET enhancing tumor, RC resection cavity, TTA test-time augmentation.

3.3 Output and Container Verification

The final validation artifact contained exactly 179 flat NIfTI files. All expected case identifiers were present, with no duplicates. Every file was readable, matched the source dimensions and affine geometry, and contained only labels 0–4.

The image contained five 2D and five 3D checkpoints and passed no-network/read-only-input audits. With Docker unavailable on the H100 host, the embedded bundle was run natively on one case and matched the frozen submission in geometry, labels, and every voxel. Its 45-s runtime and approximately 1.1-GB peak allocation were functional checks only. A 303-case A10G runtime was not measured; no compliance claim is made pending organizer execution.

4 Discussion

4.1 Principal Findings

The complete RC50 system had higher official-validation DSC and NSD values than the class-based equal-fusion system, with mean differences of +0.0588 in DSC, +0.0451 in NSD, and +0.0659 in small-lesion F1. This compares complete systems: target formulation and fusion both changed.

Region-based training directly optimizes the overlapping evaluation structures [10], providing a rationale for reformulation, but the official comparison cannot assign the difference to it alone. The region-based OOF fusion-only contrast showed a small DSC increase, marginal NSD decrease, and larger small-lesion F1 increase. A matched official comparison under identical fusion is required to isolate target formulation.

4.2 Rationale for 2D–3D Fusion

Different spatial context motivated 2D–3D fusion, but disagreement by lesion size or region was not quantified. The weights neither establish a specific division of labor nor dimensional optimality, particularly given the restricted candidate search.

Illustratively, 3D/2D probabilities of 0.90/0.08 average to 0.49 but yield 0.695 at a 0.75 3D weight; the reverse can occur when 2D is more confident. The OOF aggregates show metric-dependent fusion differences but not their anatomical source.

4.3 Small Lesions and Resection Cavities

Small metastases contribute few voxels and can resemble vessels or artifacts [2, 3, 11]. Complete RC50 showed approximately +0.09 small-lesion F1 differences for WT, TC, and ET versus the class-based equal-fusion system, without a rescue heuristic or volume filter. Validation annotations were unavailable for lesion-level error analysis.

For RC, OOF RC35-to-RC50 differences were +0.0098 in the 12-metric mean, +0.0181 in mean DSC, +0.0130 in mean NSD, and −0.0018 in small-lesion F1. Separately, the +0.1271 official RC DSC difference compares RC50 with the class-based system and cannot be attributed to RC weighting; RC50 uses 50/50 RC fusion. Small-RC F1 remained 0.0000, with size-category denominators unavailable.

4.4 Strengths and Limitations

The training arms used identical images, source annotations, preprocessing, plans, partitions, and training duration; they differed in target representation and the corresponding loss formulation. The final method used no external data, pretrained segmentation weights, component filtering, anatomical heuristic, or extra classifier. Complete OOF coverage and native-space/container verification reduced split and implementation risk.

Limitations include the missing matched official ablation; sparse weight and threshold search; and pooled outputs that precluded fold dispersion, paired intervals, or tests. Unavailable official labels, case metrics, and RC denominators prevented explanation of the official-versus-OOF and RC F1 patterns. Direct 2D/3D disagreement, calibration, dedicated small-lesion sampling, and boundary loss were not evaluated. A one-case H100 check does not establish 303-case A10G compliance. Weights may not transfer, and validation may not predict hidden-test rank.

5 Conclusion

We developed a region-aligned, target-specific 2D–3D nnU-Net fusion framework for BraTS 2026. Complete RC50 had higher observed official-validation values than the class-based equal-fusion system, but the comparison does not separate target reformulation from fusion; OOF fusion effects were small and metric-dependent. Small-RC detection remained unresolved, and a matched official ablation is needed. The organizer-data-only method was packaged and functionally verified for challenge execution.

Acknowledgments. The authors gratefully acknowledge the BraTS "Fast-track-to-MICCAI" Educational Program for providing the computational resources and educational materials that supported this work. The program aims to provide students or early-career trainees, selected through a rigorous selection process, with resources for developing state-of-the-art solutions for brain tumor segmentation and detection on Magnetic Resonance Imaging (MRI) studies. This educational program is led by Mariam S. Aboian MD, PhD (Assistant Professor of Radiology at the Department of Radiology, Children's Hospital of Philadelphia) and is funded by the RSNA Derek Harwood-Nash International Education Scholar Grant. Beyond providing these resources, Dr. Aboian and her research team had no role in the conception of the research idea, algorithm development, data analysis, or preparation of this manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process. During preparation of the revised manuscript, the authors used ChatGPT (OpenAI) for language proofing, assistance with manuscript revision . ChatGPT was not used to generate experimental data, execute the segmentation experiments, or perform the primary evaluation. All figure content was derived from the authors' reported method and results. The authors independently verified all reported numerical values, reviewed and edited all AI-assisted text and visual content, and take full responsibility for the final scientific content.

References

1. Lin, N.U., Lee, E.Q., Aoyama, H., et al.: Response assessment criteria for brain metastases: proposal from the RANO group. *Lancet Oncol.* **16**(6), e270-e278 (2015). https://doi.org/10.1016/S1470-2045(15)70057-4

2. 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. arXiv:2504.12527v3 (2025). https://arxiv.org/abs/2504.12527v3
3. Moawad, A.W., Janas, A., Baid, U., et al.: The Brain Tumor Segmentation-Metastases Challenge 2023: brain metastasis segmentation on pre-treatment MRI. arXiv:2306.00838 (2023). https://doi.org/10.48550/arXiv.2306.00838
4. Bakas, S., Linguraru, M., Kazerooni, A.F., et al.: BraTS 2026 Cluster of Challenges. Zenodo (2026). https://doi.org/10.5281/zenodo.19714728
5. Reinke, A., Tizabi, M.D., Baumgartner, M., et al.: Understanding metric-related pitfalls in image analysis validation. *Nat. Methods* **21**, 182-194 (2024). https://doi.org/10.1038/s41592-023-02150-0
6. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics Reloaded: recommendations for image analysis validation. *Nat. Methods* **21**, 195-212 (2024). https://doi.org/10.1038/s41592-023-02151-z
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203-211 (2021). https://doi.org/10.1038/s41592-020-01008-z
8. Isensee, F., Wald, T., Ulrich, C., et al.: nnU-Net Revisited: a call for rigorous validation in 3D medical image segmentation. In: Linguraru, M.G., et al. (eds.) MICCAI 2024, LNCS, vol. 15009, pp. 488-498. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_47
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015, LNCS, vol. 9351, pp. 234-241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. MIC-DKFZ: Region-based training. nnU-Net documentation. https://github.com/MIC-DKFZ/nnUNet/blob/master/documentation/region_based_training.md, last accessed 2026/07/30
11. Erdur, A.C., Scholz, D., Buchner, J.A., et al.: All sizes matter: improving volumetric brain segmentation on small lesions. In: Baid, U., et al. (eds.) Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation, LNCS, vol. 14669, pp. 177-189. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-76163-8_16
12. Yang, S., Li, X., Mei, J., Chen, J., Xie, C., Zhou, Y.: 3D-TransUNet for brain metastases segmentation in the BraTS2023 Challenge. In: Baid, U., et al. (eds.) Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation, LNCS, vol. 14669, pp. 190-199. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-76163-8_17
13. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
14. BraTS 2026 Organizing Committee: BraTS 2026 Challenge, Task 1—Segmentation of Pre- and Post-Treatment Brain Metastases. Synapse ID syn74274097 (2026). Last accessed 30 July 2026.