Brain Metastases Segmentation for BraTS 2026 Task 1: A Multi-Architecture Comparison

Mahdi Islam[1][0009-0009-5978-0471] and Musarrat Tabassum[2]

[1] Independent University, Bangladesh
mahdi@iub.edu.bd
[2] Independent Researcher
tabassummusarrat212@gmail.com

Abstract. Brain metastases are the most common intracranial malignancy, occurring in roughly 30% of patients with primary solid tumors and carrying a median survival near 5.9 months [7, 14]. Automated segmentation is critical for treatment planning and volumetric monitoring, but metastases are frequently small, numerous, and heterogeneous in size within a single patient. We compare a plain nnU-Net baseline, a Residual Encoder Large (ResEncL) variant, region-based training, and a Primus transformer model for BraTS-METS 2026 Task 1, using patient-grouped cross-validation to prevent leakage from the longitudinal UCSD subset. Primus (label-based) is our strongest individual model by aggregate DSC/NSD, achieving 0.710/0.761 (ET), 0.742/0.785 (TC), 0.683/0.689 (WT), and 0.531/0.436 (RC). ResEncL trails Primus on aggregate DSC/NSD but achieves substantially higher lesion-wise F1 (e.g. ET: 0.452 vs. 0.052); a probability-averaging ensemble of the two only partially preserves ResEncL's F1 advantage (ET lesion-wise F1: 0.064). We further report three postprocessing and label-reconstruction pitfalls we believe generalize beyond this challenge. Code is available at github.com/mahdiislam79/BraTS_METS_2026.

Keywords: Brain Metastases · Tumor Segmentation · nnU-Net · Transformer · Data Leakage · Small Lesion Detection

1 Introduction

Brain metastases represent the most prevalent intracranial malignancy, arising in an estimated 30% of patients diagnosed with primary solid cancers including lung, breast, and melanoma [7, 11]. Their development is associated with substantially increased healthcare burden, approximately $13,000 higher per-patient costs when central nervous system spread occurs [12], and a median overall survival of less than 6 months [14]. Stereotactic radiosurgery (SRS) and whole-brain radiation therapy (WBRT) remain primary treatment modalities, both of which demand precise, reproducible lesion delineation to minimize radiation dose to healthy tissue [10].

Accurate volumetric segmentation across serial MRI examinations is a prerequisite for automated treatment-response monitoring, enabling clinicians to detect progression or regression of individual metastatic lesions without inter-observer variability. The longitudinal structure of the BraTS-METS 2026 dataset [9], particularly the UCSD subset, which provides multiple imaging timepoints per patient makes it uniquely suited for developing such monitoring pipelines, provided that data splitting is handled to prevent temporal leakage between training and validation folds.

Recent advances in automated brain tumor segmentation have been driven primarily by nnU-Net [3], a self-configuring framework that remains highly competitive across segmentation benchmarks. Architecture extensions, including Residual Encoder variants [4], region-based training for overlapping tumor subregion definitions [5], and transformer-based models such as SwinUNETR [2] and Primus [13], have demonstrated incremental gains. Multi-architecture ensembling yields the strong reported results in prior BraTS challenges [1,6], by combining complementary failure modes across convolutional and transformer-based models. Persistent challenges remain in small and multifocal lesion detection, handling of the rare resection cavity (RC) class, and robustness to extreme class imbalance between metastatic lesion volume and surrounding brain tissue.

We present our approach to BraTS-METS 2026 Task 1, segmenting pre- and post-treatment brain metastases into four subregions: enhancing tumor (ET), non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), and resection cavity (RC). Our pipeline compares a plain nnU-Net baseline, a Residual Encoder Large (ResEncL) variant, region-based training, and a Primus transformer model, and combines the two strongest individual models (ResEncL and Primus, both label-based) into a probability-averaging ensemble. Beyond this architectural comparison, we report three findings we believe are useful to other implementers building metastasis segmentation pipelines: (i) a patient-grouped cross-validation scheme that prevents the temporal data leakage inherent to the longitudinal UCSD subset of BraTS-METS; (ii) an analysis of a postprocessing failure mode morphological opening silently destroying small lesions near the challenge's scoring threshold that we argue generalizes to other small-lesion segmentation pipelines; and (iii) identification of a labeling-order assumption in region-based nnU-Net reconstruction that can silently corrupt output classes under non-trivial containment hierarchies, together with a correction procedure requiring no retraining.

2 Method

2.1 Dataset and Preprocessing

We used the BraTS-METS 2025 Lighthouse dataset [8][3], reused for BraTS 2026 Task 1, comprising 1,496 training cases from nine contributing institutions: Duke, NCI, Missouri, WashU, Yale, UCSF, Northwestern, UCSD, and Ulm [9]. Of these,

[3] Synapse ID: `syn74274097`.

1,296 cases carry expert annotations and were used for supervised training; the 200 Ulm cases are unannotated and were excluded. Four MRI modalities are provided: native T1 (T1n), contrast-enhanced T1 (T1c), T2-weighted (T2w, non-mandatory from 2025 onward), and T2-weighted FLAIR (T2f). Ground truth labels follow a five-class scheme: 0 = background, 1 = NETC, 2 = SNFH, 3 = ET, 4 = RC. RC (label 4) is present in approximately 13% of cases and is retained as a separate class distinct from the whole tumor (WT) hierarchy.

Two cases contained out-of-scheme segmentation labels. One was resolved by applying the official corrected-labels overlay; the other (129 voxels affected) was not covered by that correction and was excluded from training. Excluding this single unresolved case, all remaining 1,295 annotated cases were eligible for conversion; a further, separate set of cases was excluded at conversion time for missing one or more of the four required modalities (T2w in particular is non-mandatory and absent for several cases), following a conservative complete-modality inclusion criterion. Twenty-eight cases with all-zero segmentation masks confirmed as legitimate no-visible-disease UCSD cases per the dataset description were retained as valid negative training examples rather than treated as missing annotations.

Data were converted to nnU-Net raw format using standard nnU-Net fingerprinting and automatic preprocessing. For the region-based training variant, a second version of the dataset was constructed with overlapping region labels in place of the mutually exclusive per-voxel classes used for the other models.

2.2 Cross-Validation Splits and Leakage Prevention

The UCSD subset has a longitudinal structure where each patient contributes multiple imaging timepoints, encoded within the case identifier. Naive random splitting or modulo-based fold assignment allows the same patient to appear in both training and validation within a fold, inflating estimated performance. We instead grouped cases by patient identifier and applied grouped 5-fold cross-validation (`GroupKFold`). The resulting fold assignment was reused identically across all model variants to ensure fair cross-model comparison.

2.3 Architectures

We trained and evaluated four model configurations.

Plain nnU-Net (Baseline). Standard 3D full-resolution nnU-Net v2 [3] with default architecture and plans. We also included an earlier, shorter-trained (200-epoch) checkpoint that predates the postprocessing correction described in Section 2.6 and differs from the 1000-epoch baseline in both training length and postprocessing.

ResEncL. Residual Encoder Large variant [4], trained on the label-based dataset with training configuration otherwise identical to the baseline.

Region-Based Training + ResEncL. To handle overlapping region definitions ($\mathrm{WT} = \{\mathrm{NETC}, \mathrm{SNFH}, \mathrm{ET}\}$; $\mathrm{TC} = \{\mathrm{NETC}, \mathrm{ET}\}$), we adopted a region-

based training scheme [5] with overlapping region labels in place of mutually exclusive classes, combined with the ResEncL architecture.
Primus. A Primus (PrimusV3S) transformer model [13], integrated as an nnU-Net v2 trainer class, trained in both label-based and region-based configurations. The label-based configuration achieved the strongest aggregate DSC/NSD among individual models (Section 3); reconstruction of discrete labels from the region-based configuration's trained probabilities is described in Section 2.5.

2.4 Training Configurations

All nnU-Net-based models (baseline, ResEncL, and Primus) used nnU-Net's self-configured preprocessing and default optimization protocol, with patch size and batch size determined automatically per configuration via nnU-Net's fingerprinting step, and deep supervision enabled by default. Plain nnU-Net and ResEncL used the default `nnUNetTrainer` (SGD, Nesterov momentum 0.99, initial LR 0.01 with polynomial decay, per-sample rather than batch Dice); Primus used its dedicated `nnUNet_PrimusV3S_Trainer` with that trainer's default settings. Inference used sliding-window prediction with mirroring test-time augmentation, with region probabilities binarized at 0.5. The plain nnU-Net baseline was trained for 1,000 epochs across 5 folds on an NVIDIA A100 GPU; the earlier, shorter-trained checkpoint referenced above was trained for 200 epochs. ResEncL, region-based ResEncL, and both Primus configurations were each trained for 5 folds on NVIDIA H100 SXM GPUs, with folds trained in parallel. All models were trained and evaluated on the identical patient-grouped 5-fold splits described above.

2.5 Region-Order Assumption in Region-Based Label Reconstruction

Region-based nnU-Net reconstructs a discrete label map from overlapping region predictions by processing regions in a declared order, with each region overwriting prior assignments where they overlap. This assumes the declared order matches the regions' anatomical containment hierarchy, outermost to innermost. Our labeling scheme violates this under naive ascending order: label 1 (NETC) is innermost, while label 2 (SNFH) is outermost, the reverse of what ascending order assumes. As a result, the whole-tumor region was assigned label 1 instead of label 2, inflating NETC with misclassified SNFH voxels and collapsing the tumor-core (TC) score. We identified this via a voxel-count comparison against the label-based model's output, and corrected it by reordering region processing to place the outer region first, overwritten by the inner tumor-core region, then the remaining independent classes.

2.6 Postprocessing

Initial postprocessing applied morphological opening per class, which caused lesion-wise F1 to collapse to near zero (0.01–0.02) despite reasonable DSC. The

root cause was that the erosion step eliminates connected components near the 27 mm^3 scoring threshold before dilation can restore them. We removed morphological opening and replaced it with connected-component filtering based directly on physical volume: components below 27 mm^3 are discarded. This fix alone recovered lesion-wise F1 from ∼0.01 to ∼0.33–0.41 across ET/TC/WT. For the ensemble, per-class hole-filling is additionally applied after the volume-threshold step. We highlighted this failure mode because morphological opening is a common default in segmentation postprocessing pipelines, and its interaction with small-lesion scoring thresholds is rarely reported explicitly in prior BraTS work.

3 Results

Table 1 reports lesion-wise DSC and NSD, per class, for all trained configurations on the official validation leaderboard. Table 2 reports lesion-wise F1, taken directly from the official scorer's small-instance F1 output for each submission.

Table 1. Lesion-wise DSC / NSD on the validation set, by class.

Model	ET	TC	WT	RC
Plain nnU-Net, 200 ep	0.646 / 0.698	0.666 / 0.683	0.609 / 0.597	0.406 / 0.279
Plain nnU-Net, 1000 ep	0.664 / 0.727	0.688 / 0.737	0.652 / 0.675	0.478 / 0.376
ResEncL, label-based	0.693 / 0.755	0.708 / 0.757	0.676 / 0.699	0.415 / 0.345
Primus, label-based	0.710 / 0.761	**0.742 / 0.785**	0.683 / 0.689	0.531 / **0.436**
Primus, region-based	**0.713 / 0.762**	0.400 / 0.433	0.680 / 0.682	**0.566** / 0.376
Ensemble (Primus + ResEncL)	0.711 / 0.758	0.742 / 0.782	**0.685 / 0.691**	0.520 / 0.425

Table 2. Lesion-wise F1 on the validation set, by class.

Model	ET	TC	WT	RC
Plain nnU-Net, 200 ep	0.012	0.012	0.002	0
Plain nnU-Net, 1000 ep	0.414	0.423	0.331	0
ResEncL, label-based	**0.452**	**0.449**	**0.354**	0
Primus, label-based	0.052	0.053	0.044	0
Primus, region-based	0.051	0.044	0.044	0
Ensemble (Primus + ResEncL)	0.064	0.065	0.044	0

Primus (label-based) achieves the highest mean DSC/NSD across classes among individual models, leads on TC, and is the model used in our final ensemble; per-class differences among the top configurations are otherwise narrow, with the ensemble marginally ahead on WT and region-based Primus ahead on RC DSC specifically. Region-based Primus improves ET DSC/NSD (0.713/0.762 vs. 0.710/0.761) but underperforms on TC and RC NSD, a trade-off we consider a current limitation of the region-based approach. On lesion-wise F1 (Table 2), both Primus configurations perform similarly poorly (0.044–0.053) and are substantially outperformed by ResEncL (0.354–0.452) on ET, TC, and WT; all models score zero on RC (discussed below).

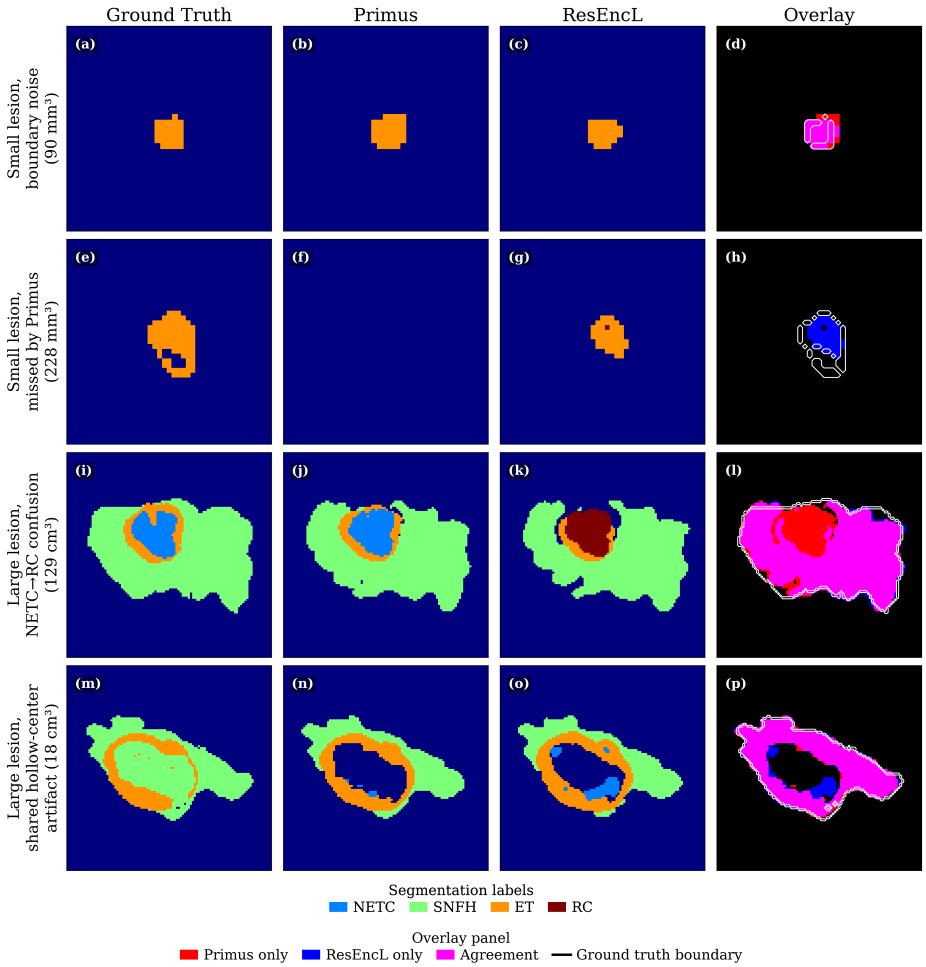

Fig. 1. Qualitative comparison of ResEncL and Primus segmentations against ground truth for four representative cases, illustrating (row 1) boundary-level disagreement on a small, correctly-detected lesion; (row 2) a complete detection failure by Primus on a small lesion correctly identified by ResEncL; (row 3) a case-specific substitution of necrotic tumor core (NETC) with resection cavity (RC) by ResEncL; and (row 4) a hollow-center prediction artifact shared by both models on a ring-shaped lesion. Columns show ground truth, Primus prediction, ResEncL prediction, and an overlay of the whole-tumor region (NETC ∪ SNFH ∪ ET) where red indicates Primus-only, blue indicates ResEncL-only, magenta indicates agreement, and the white contour marks the ground-truth boundary.

Lesion-wise F1 reveals a starkly different pattern from DSC/NSD. ResEncL achieves the strongest lesion-wise F1 across ET/TC/WT (0.354–0.452), moder-ately exceeding even the nnU-Net baseline (0.331–0.423), while Primus lags far behind both (0.044–0.053) despite leading on aggregate overlap metrics. This

dissociation between overlap-based accuracy and small-lesion detection is the central empirical finding of this comparison: a model's DSC/NSD ranking is not predictive of its small-instance detection ranking, and the two must be evaluated independently. We attribute Primus's small-instance weakness to its transformer backbone's coarse 3D tokenization, which operates at a spatial scale close to that of the smallest scored lesions, while ResEncL's fully convolutional design preserves fine-grained spatial resolution better suited to sub-token-scale structures. This complementary strength profile suggests genuine architectural diversity between the two models, making them promising candidates for ensembling.

Figure 1 illustrates this dissociation qualitatively across four representative cases spanning small and large lesion scales. For small lesions, model disagreement is not uniform: in one case (90 mm^3), both models correctly localize and roughly match the lesion's extent, with the discrepancy confined to a small number of boundary voxels. In another case (228 mm^3), however, Primus fails to detect the lesion entirely while ResEncL correctly identifies it, indicating a genuine detection failure rather than boundary-level noise, consistent with the quantitative lesion-wise F1 gap above.

For large lesions, model behavior diverges differently. In one case (129 cm^3), ResEncL substitutes a substantial portion of the necrotic tumor core (NETC) with resection cavity (RC), a clinically distinct structure, while Primus's prediction closely matches the ground-truth NETC/ET layering; we observe this substitution in this case only and report it as a case-specific failure mode rather than a systematic tendency. A separate case (18 cm^3) illustrates a distinct, shared limitation: both models under-predict the interior enhancing/necrotic composition of a ring-shaped lesion, producing a hollow region not present in the ground truth.

4 Discussion

The most practically impactful finding is that morphological postprocessing meant to clean up predictions can suppress detection of the small lesions the evaluation penalizes most heavily. The lesion-wise F1 metric is highly sensitive to whether small components survive postprocessing, so practitioners should check postprocessing effects on the small-lesion subset specifically, not just aggregate DSC.

A related pattern emerges from comparing Primus and ResEncL under identical postprocessing: Primus leads on DSC/NSD, but ResEncL's lesion-wise F1 is roughly an order of magnitude higher (e.g. ET: 0.052 vs. 0.452). Strong overlap metrics do not guarantee good small-lesion detection, and the two should be evaluated separately. Since ResEncL achieved strong DSC without losing lesion-wise F1, the two models appeared complementary, motivating our ensemble. In practice, the ensemble's lesion-wise F1 (e.g. ET: 0.064) stayed close to Primus's rather than ResEncL's, despite weighting intended to protect ResEncL's contribution. This suggests probability-averaging ensembles do not automatically preserve a component model's minority-case strengths: when Primus assigns

near-zero probability to a lesion it misses, averaging with ResEncL's moderate probability often still falls short of the 0.5 decision threshold. Threshold- or size-adaptive ensemble weighting may address this.

ResEncL also shows a distinct weakness on RC (DSC 0.415, the lowest of the individually-trained models), despite sitting between the nnU-Net baseline and Primus elsewhere. RC is a single, well-defined cavity rather than a small multifocal structure, so this is likely a separate, unexplained effect rather than the same cause as Primus's small-lesion gap.

The under-detection pattern (false negatives outweighing false positives) seen in cross-validation points toward foreground oversampling, false-negative-weighted losses such as Tversky loss, or a two-stage cascade as likely next steps. Given Primus's shortfall sits in the same small-lesion regime, we consider a two-stage cascade the highest-priority next experiment.

RC remains a weakness across our pipeline. ResEncL performs worst (DSC 0.415), the nnU-Net baseline partially compensates (0.478), and region-based Primus achieves the highest RC DSC among our models (0.566), though label-based Primus leads RC NSD (0.436); even the best of these falls well short of competitive pipelines. Official small-instance scoring shows zero true positives for small resection-cavity instances ($\sim$1.17 per case, averaged) across every model we submitted – nnU-Net, ResEncL, both Primus configurations, and the ensemble, indicating a small-RC detection failure that is universal across architectures rather than specific to any one model.

Our region-based training pipeline surfaces a labeling-order assumption in nnU-Net's region reconstruction (Section 2.5) likely to recur in other setups where label containment order does not match ascending label values.

Open items. (1) NSD tolerance in our offline evaluator is unverified against the official scoring script. (2) The lesion matching rule is unconfirmed against the official evaluator. (3) Missing-T2w validation cases are excluded at inference, which may cause silent score drag.

5 Conclusion

We compared four segmentation configurations for BraTS-METS 2026 Task 1 under patient-grouped cross-validation, finding that Primus and ResEncL showed complementary strongholds: Primus led on aggregate DSC/NSD, while ResEncL led substantially on lesion-wise F1, with neither model unambiguously best across both. Our probability-averaging ensemble only partially captured this complementarity, suggesting a more targeted ensembling strategy e.g. size- or threshold-adaptive weighting could recover more of ResEncL's small-lesion strength without sacrificing Primus's overlap accuracy. Future work includes such adaptive ensembling, a two-stage cascade for the small-lesion regime, and anatomical hierarchy enforcement for RC detection.

Acknowledgments Data used in this publication were obtained as part of the Challenge project. 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.

References

1. Astaraki, M., Moghadam, F.S., Toma-Dasu, I.: Brain tissue context for enhancing brain tumor segmentation: A contribution to brats 2025. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 112–126. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-16365-3_10
2. Hatamizadeh, A., et al.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021). Lecture Notes in Computer Science, vol. 12962. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08999-2_22
3. 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. Nature Methods **18**, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
4. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation (2024), https://arxiv.org/abs/2404.09556
5. Isensee, F., et al.: nnU-Net for Brain Tumor Segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2020). Lecture Notes in Computer Science, vol. 12659. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72087-2_11
6. Krikorian, W., Purwar, A.: Automated segmentation for the brain tumor segmentation (brats) metastases 2025 challenge using multi-architectural deep learning. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 173–181. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-16365-3_16
7. Lamba, N., Wen, P.Y., Aizer, A.A.: Epidemiology of brain metastases and leptomeningeal disease. Neuro-Oncology **23**(9), 1447–1456 (2021). https://doi.org/10.1093/neuonc/noab101
8. Maleki, N., Amiruddin, R., Moawad, A.W., Yordanov, N., Gkampenis, A., Fehringer, P., Umeh, F., Chukwurah, C., Memon, F., Petrovic, B., 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/2504.12527v3
9. Moawad, A.W., et al.: The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri (2024), https://arxiv.org/abs/2306.00838
10. Nahed, B.V., Alvarez-Breckenridge, C., Brastianos, P.K., Shih, H., Sloan, A., Ammirati, M., Kuo, J.S., Ryken, T.C., Kalkanis, S.N., Olson, J.J.: Congress of neurological surgeons systematic review and evidence-based guidelines on the role of surgery in the management of adults with metastatic brain tumors. Neurosurgery **84**(3), E152–E155 (2019). https://doi.org/10.1093/neuros/nyy542
11. Ostrom, Q.T., Wright, C.H., Barnholtz-Sloan, J.S.: Brain metastases: Epidemiology. In: Handbook of Clinical Neurology, vol. 149, pp. 27–42. Elsevier (2018). https://doi.org/10.1016/B978-0-12-811161-1.00002-5
12. Shim, Y.B., Byun, J.Y., Lee, J.Y., Lee, E.K., Park, M.H.: Economic burden of brain metastases in non-small cell lung cancer patients in South Korea: A retrospective cohort study using nationwide claims data. PLOS ONE **17**(9), e0274876 (2022). https://doi.org/10.1371/journal.pone.0274876
13. Wald, T., Roy, S., Isensee, F., Ulrich, C., Ziegler, S., Trofimova, D., Stock, R., Baumgartner, M., Köhler, G., Maier-Hein, K.: Primus: Enforcing attention usage for 3d medical image segmentation (2026), https://arxiv.org/abs/2503.01835
14. Yri, O.E., Astrup, G.L., Karlsson, A.T., Van Helvoirt, R., Hjermstad, M.J., Husby, K.M., Loge, J.H., Lund, J.Å., Lundeby, T., Paulsen, Ø., Skovlund, E., Taran, M.I., Winther, R.R., Aass, N., Kaasa, S.: Survival and quality of life after first-time diagnosis of brain metastases: A multicenter, prospective, observational study. The Lancet Regional Health - Europe **49**, 101181 (2025). https://doi.org/10.1016/j.lanepe.2024.101181