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.

A Residual-Encoder nnU-Net Baseline for Pre- and Post-Treatment Brain Metastasis Segmentation

Ujjwal Baid[1], Bhakti Baheti[1], and Anant Madabhushi[1,2]

[1] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA
[2] Atlanta Veterans Affairs Medical Center, Atlanta, GA, USA
ujjwalbaid0408@gmail.com

Abstract. Volumetric delineation of brain metastases is needed for treatment planning and objective disease monitoring, but manual annotation is slow and poorly reproducible, and the lesions of greatest clinical interest are often only a few millimetres across. We describe our entry to Task 1 of the BraTS-Lighthouse 2026 challenge, which asks for simultaneous segmentation of the non-enhancing tumour core, the surrounding non-enhancing FLAIR hyperintensity, the enhancing tumour and the resection cavity in pre- and post-treatment multiparametric MRI. The method is deliberately a reproducible baseline rather than a new architecture: a single extra-large residual-encoder 3D U-Net configured by nnU-Net, trained on all 1,296 annotated training studies without a held-out fold, followed by a region-selective filter that removes only resection-cavity components smaller than 5 mm^3. On the official 179-study validation set it reaches lesion-wise Dice scores of 0.677 (enhancing tumour), 0.702 (tumour core), 0.645 (whole tumour) and 0.526 (resection cavity), with normalised surface distances of 0.735, 0.749, 0.666 and 0.434. A per-case analysis isolates two dominant failure modes. Accuracy on the enhancing tumour degrades as lesion volume and burden decrease, and half of all annotated enhancing components fall below the challenge scoring threshold. The resection cavity, present in only 12.9% of studies, is limited almost entirely by false-positive cavity predictions in studies that contain none, which the volume filter halves without altering accuracy where a cavity exists. Inference takes about three seconds per study in a self-contained container. Code, the trained weights and the container recipe are available at `https://github.com/ujjwalbaid0408/BraTS2026_METS`.

Keywords: Brain metastases · Image segmentation · Magnetic resonance imaging · nnU-Net · BraTS challenge.

1 Introduction

Brain metastases are the most frequently encountered intracranial tumours, and their incidence continues to rise as systemic therapies extend survival in patients with solid primary cancers [12]. Clinical practice still assesses them largely

through the longest unidimensional diameter recommended by the Response Assessment in Neuro-Oncology Brain Metastases (RANO-BM) criteria [13]. Volumetric measurement of the lesions and the surrounding oedema is more informative, yet doing it manually across the many small lesions of a typical study is slow and subject to substantial inter-rater variability. Automated segmentation is therefore attractive, and the field has moved from template matching towards deep learning [7, 14], but progress has been hard to compare because most results come from single-centre cohorts.

The Brain Tumor Segmentation (BraTS) challenge series addresses this by providing standardised multi-institutional benchmarks with a common annotation protocol and a common evaluation server [16, 3, 2]. The metastasis arm, BraTS-METS, was introduced in 2023 for pre-treatment studies [17] and has since been extended to the post-treatment setting, in which a resection cavity must also be delineated. Task 1 of the BraTS-Lighthouse 2026 cluster [4] asks for a single algorithm that handles both settings and remains reliable across the mixture of scanners, protocols and resolutions found in routine care.

Our aim was not to propose a new architecture but to establish how far a carefully configured, containerised and reproducible standard pipeline can be taken, and then to diagnose where it fails. We (i) train a single extra-large residual-encoder 3D U-Net configured automatically by nnU-Net [8, 9] on all 1,296 annotated studies, using only challenge data and no pre-trained weights, and report the full configuration and training trace; (ii) apply a region-selective post-processing rule for the resection cavity and show that its benefit comes entirely from suppressing false-positive cavities rather than from refining real ones; (iii) quantify how accuracy depends on lesion size and lesion burden, and relate this to the 27 mm^3 threshold used in scoring; and (iv) package the method as a container that segments a study in about three seconds on a single GPU.

2 Methods

2.1 Data

The BraTS 2026 brain metastasis dataset is a retrospective collection of pre- and post-treatment multiparametric MRI (mpMRI) studies acquired under standard clinical conditions at nine institutions in the United States and Europe. Each study provides four sequences: pre-contrast T1-weighted (T1n), post-contrast T1-weighted (T1c), T2-weighted (T2w) and T2-weighted fluid-attenuated inversion recovery (T2f). Reference annotations were produced by STAPLE fusion of several automated algorithms, manually refined by neuroradiologists following a common protocol and approved by board-certified attending neuroradiologists. Four foreground labels are defined: non-enhancing tumour core (NETC, label 1), surrounding non-enhancing FLAIR hyperintensity (SNFH, label 2), enhancing tumour (ET, label 3) and resection cavity (RC, label 4). Studies come in a mixture of native and SRI24 atlas space, which adds to the heterogeneity of the benchmark. We used the 1,296 annotated training studies for model development and the 179 validation studies, whose annotations are not released, for

leaderboard evaluation through the challenge benchmarking infrastructure [10]. In accordance with the challenge rules we used no external data and no pre-trained weights.

The composition of the training set explains much of what follows. Enhancing tumour is present in 1,250 of the 1,296 studies and a resection cavity in only 167, that is 12.9%. The annotations contain 12,058 spatially connected enhancing components, a median of 4 per study, with a skewed distribution reaching 403 components in one study. Only 6,006 of them, just under half, exceed the $27\,\mathrm{mm}^3$ volume above which the challenge computes segmentation metrics, so much of the annotated lesion population sits at or below the scale at which the task is scored.

2.2 Preprocessing and network configuration

We used nnU-Net v2 in its `3d_fullres` configuration with the extra-large residual-encoder planner `nnUNetPlannerResEncXL` [9], and left the derived plan unchanged. The planner resampled all studies to the dataset median voxel spacing of $1.00 \times 0.898 \times 0.859\,\mathrm{mm}^3$, using third-order spline interpolation for images and linear interpolation for label maps, and applied per-channel z-score normalisation computed inside the non-zero (brain) mask. The four sequences enter as four input channels in the fixed order T1n, T1c, T2w, T2f.

The architecture is a residual-encoder 3D U-Net with six resolution stages, 32/64/128/256/320/320 features per stage, 1/3/4/6/6/6 residual blocks per encoder stage and one convolution per decoder stage, all with $3 \times 3 \times 3$ kernels, instance normalisation and leaky ReLU activations. The planner selected a patch size of $160 \times 224 \times 192$ voxels with batch size 4, which covers a large fraction of the brain in one forward pass and so gives wide spatial context when deciding whether an enhancing focus is a metastasis or a vessel. Deep supervision is enabled at all decoder resolutions.

2.3 Training

The model was trained with the default nnU-Net recipe: stochastic gradient descent with Nesterov momentum 0.99, initial learning rate 0.01 decayed polynomially as $\mathrm{lr}_e = \mathrm{lr}_0(1 - e/E)^{0.9}$, the sum of Dice and cross-entropy losses computed per sample, 250 iterations per epoch, mixed-precision arithmetic and the standard spatial and intensity augmentation pipeline, with 33% of training patches forced to contain foreground. Following the strategy of the winning 2025 entry [5] we trained on fold `all`, that is on all 1,296 annotated studies rather than a four-fifths split, so as to expose the model to every one of the rare resection-cavity and small-lesion studies.

Training ran on a single NVIDIA A100-SXM4 40 GB GPU with a median epoch time of 588 s on shared cluster hardware, about 200 GPU-hours in total. The 1,000-epoch schedule reached epoch 982 before the compute allocation expired and attempts to resume terminated with GPU out-of-memory errors,

so the deployed weights are the checkpoint with the best exponentially-moving-average pseudo-Dice, recorded at epoch 934 with a value of 0.870. The learning rate had by then decayed to 8.7×10^{-4} and both loss and pseudo-Dice had been flat for more than a hundred epochs (Fig. 1), so we do not expect the missing epochs to be material.

2.4 Inference, post-processing and containerisation

Inference uses nnU-Net sliding-window prediction with Gaussian patch weighting and 50% overlap. We deliberately disabled mirroring test-time augmentation and used a single model rather than an ensemble, so the reported numbers reflect one network evaluated once per study. Predictions are resampled back to the original grid and written as a single 4-label mask.

Post-processing is one region-selective rule: connected components of the resection-cavity label smaller than $5\,\mathrm{mm}^3$ are removed, while the enhancing tumour, non-enhancing core and FLAIR hyperintensity labels are left untouched. The asymmetry is intentional. The 2025 challenge report found that applying a $5\,\mathrm{mm}^3$ filter to all labels slightly reduced Dice for the enhancing tumour, which is the clinically most important region and the one most likely to contain genuinely tiny components [5], so we restrict the filter to the label for which small components are almost always spurious. Section 3 quantifies what the rule actually buys.

The method is packaged as a container implementing the challenge submission contract, reading one folder per subject from `/input` and writing one mask to `/output`. The image bundles the trained weights and pins the training environment (Python 3.11, PyTorch 2.11.0 with CUDA 12.8, nnU-Net v2.7.0, `dynamic-network-architectures` 0.4.3, SimpleITK 2.5.4, NumPy 2.4.3) so the checkpoint loads deterministically. It was validated end to end under Apptainer with an isolated filesystem (`--containall`), reproducing the conditions under which the organisers execute submissions. Segmenting the 179 validation studies took 549 s on a single GPU including format conversion and post-processing, about 3 s per study.

2.5 Evaluation metrics

The challenge evaluates four composite regions: enhancing tumour (ET, label 3), resection cavity (RC, label 4), tumour core (TC, labels 1 and 3) and whole tumour (WT, labels 1, 2 and 3). It computes the Dice similarity coefficient (DSC) and the normalised surface distance (NSD), both bounded above by 1, aggregated lesion-wise over lesions larger than $27\,\mathrm{mm}^3$; undefined metrics are excluded rather than penalised, and ranking follows current recommendations for multi-property validation [18, 15]. For the error analysis of Section 3 we also computed case-wise DSC and NSD (1 mm tolerance) locally, treating a metric as undefined when prediction and reference are both empty.

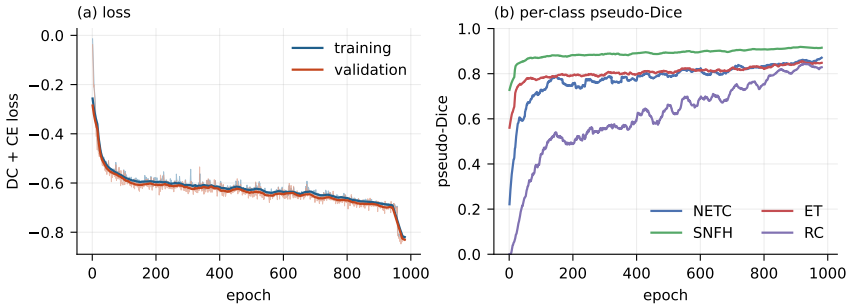

Fig. 1. Training dynamics over the 983 completed epochs. (a) Loss on training and on nnU-Net's internal validation patches; faint lines are per-epoch values, solid lines a 31-epoch moving average. A step near epoch 958 is not accompanied by any change in pseudo-Dice. (b) Per-class pseudo-Dice, computed on foreground-oversampled patches and therefore optimistic relative to case-wise scores; the resection cavity converges far more slowly than the other classes.

3 Results

3.1 Training behaviour

Fig. 1 shows the training trace. The loss falls steeply over the first 50 epochs, then improves slowly and monotonically, with no divergence between training and internal validation curves, as expected when every study is in the training pool. The pseudo-Dice curves separate the labels clearly: FLAIR hyperintensity plateaus near 0.91 and the enhancing tumour and non-enhancing core near 0.85 to 0.87, whereas the cavity needs roughly 400 epochs to become competitive and is still improving late in training, which is the strongest argument for the long default schedule.

3.2 Official validation results

Table 1 (upper block) reports the scores returned by the challenge evaluation server on the 179-study validation set. The tumour core is the strongest region, followed by the enhancing tumour and the whole tumour, with the resection cavity clearly weakest. Normalised surface distance exceeds Dice for every region except the resection cavity, which indicates that boundaries usually fall within tolerance even where volumetric overlap is imperfect, and that the residual error is concentrated in whole lesions missed or spuriously added rather than in systematically misplaced contours.

3.3 Training-set error analysis

Because the validation annotations are not released, all further analysis uses the training set. We re-ran the complete inference pipeline over all 1,296 training studies and computed case-wise metrics locally (Table 1, lower block). These

Table 1. Top: lesion-wise results on the official 179-study validation set, as returned by the challenge evaluation server for the submitted container. Bottom: case-wise metrics (mean ± standard deviation) on the 1,296 training studies, before and after post-processing, where n is the number of studies for which the metric is defined and the last row restricts the resection cavity to the 167 studies in which one is annotated. The training rows are in-sample and serve only for error analysis. Higher is better throughout.

Region	Setting	n	DSC	NSD
Official validation set, lesion-wise				
ET	submitted container	179	0.677	0.735
TC	submitted container	179	0.702	0.749
WT	submitted container	179	0.645	0.666
RC	submitted container	179	0.526	0.434
Training set, case-wise (in-sample)				
ET	raw and filtered	1263	0.789 ± 0.203	0.863 ± 0.198
TC	raw and filtered	1263	0.813 ± 0.210	0.863 ± 0.199
WT	raw and filtered	1269	0.831 ± 0.203	0.827 ± 0.194
RC	raw	320	0.394 ± 0.414	0.388 ± 0.411
RC	filtered	241	0.523 ± 0.400	0.516 ± 0.398
RC	filtered, cavity present	167	0.754 ± 0.235	0.744 ± 0.241

numbers are in-sample and are an upper bound rather than an estimate of generalisation; we use them only to locate failure modes, and they are not comparable with the lesion-wise validation scores above.

The resection-cavity rows are the informative ones. Taken at face value the $5\,\mathrm{mm}^3$ filter appears to raise cavity Dice from 0.394 to 0.523. It does nothing of the sort. On the 167 studies that genuinely contain a cavity the mean Dice is 0.754 both before and after filtering, unchanged to three decimals. The entire apparent gain comes from the 1,129 studies containing no cavity, in which the raw model nevertheless predicted one in 153 cases; the filter removes 79 of those spurious predictions, leaving 74. Each suppressed false positive turns a score of zero into an undefined value that is then excluded from the average, so the aggregate rises while nothing about the segmentation of real cavities changes. Resection-cavity performance in this benchmark is therefore governed by a detection decision rather than by delineation quality, and aggregate cavity scores must be read together with the number of contributing cases.

Read as a study-level detection problem, the same counts give a sensitivity of 0.928 (155 of the 167 cavity-bearing studies overlap the reference cavity), while the filter raises specificity from 0.864 to 0.934, precision from 0.503 to 0.677 and the F_1 score from 0.653 to 0.783, and lowers the false-positive rate from 0.118 to 0.057 spurious cavities per study. Sensitivity is identical before and after filtering. Among the 167 cavity-bearing studies the filter changed case-wise cavity Dice by at most -0.006 and by $+0.003$ at best, leaving the mean unmoved at 0.754, and even in the 14 studies whose annotated cavity is smaller than $100\,\mathrm{mm}^3$ the

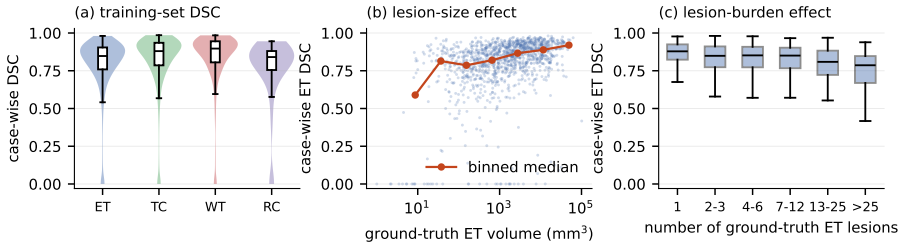

Fig. 2. Case-wise analysis on the 1,296 training studies (in-sample). (a) Dice per region, boxes showing quartiles and median inside the kernel density estimate ($n = 1263$, 1263, 1269 and 167 studies for ET, TC, WT and RC, the cavity being restricted to studies containing one). (b) Enhancing-tumour Dice against total annotated enhancing volume, with binned medians. (c) The same Dice grouped by number of annotated components.

mean moves only from 0.2899 to 0.2901. The 5 mm^3 rule thus suppresses spurious cavities without removing a genuine one, and the 12 cavities the model misses entirely are missed before the filter is applied, not because of it.

Fig. 2 decomposes the enhancing-tumour result. All regions have strongly left-skewed distributions: the median enhancing-tumour Dice is 0.848, well above the mean of 0.789, so the mean is pulled down by a minority of near-zero cases rather than by a broad loss of accuracy. Panel (b) shows where those cases are. Grouping studies by total annotated enhancing volume, median Dice rises from 0.796 below 100 mm^3 ($n = 131$) to 0.800 between 100 and 1,000 mm^3 ($n = 409$), 0.868 between 10^3 and 10^4 mm^3 ($n = 530$) and 0.896 above 10^4 mm^3 ($n = 180$). The corresponding means, 0.659, 0.768, 0.833 and 0.856, fall much faster than the medians towards small volumes, confirming that small-volume studies fail catastrophically when they fail at all. Panel (c) shows a weaker second effect: median Dice falls from 0.879 for studies with one enhancing lesion to 0.786 for studies with more than 25, since the probability of missing at least one lesion grows with the count.

3.4 Relation to the previous challenge edition

Five BraTS-METS 2025 entries have been published [5, 11, 19, 1, 6]. We deliberately do not tabulate their scores against ours. The two editions differ in cohort and, more consequentially, in metric aggregation: the 2026 server aggregates lesion-wise over lesions above 27 mm^3, whereas the 2025 papers report subject-wise aggregation. The numbers are therefore not on a common scale and a side-by-side reading would suggest a precision of comparison that does not exist. We note only the methodological relationship. The 2025 winning entry [5] used the same nnU-Net family and the same fold `all` strategy as ours, but additionally searched over loss functions, model sizes and post-processing variants. That search, rather than any architectural difference, is the main methodological gap between that work and this baseline, and is the most plausible explanation

for its stronger resection-cavity handling. The observation is consistent with the finding that well-configured convolutional baselines remain competitive in 3D medical segmentation [9].

4 Discussion

Small lesions, not contours, set the ceiling. Across every analysis the dominant error is a whole lesion missed or hallucinated rather than a boundary misplaced: surface distance exceeds Dice on three of four validation regions, the Dice distributions are left-skewed with long tails to zero, and mean accuracy on the enhancing tumour degrades far faster than median accuracy as lesion volume falls. Effort spent on boundary-refinement losses is therefore unlikely to move the aggregate much, whereas effort on detection sensitivity for lesions of a few dozen cubic millimetres almost certainly will. That 6,052 of the 12,058 annotated enhancing components fall below the $27\,\mathrm{mm}^3$ scoring threshold also means the metric is forgiving of the lesions hardest to find, so a leaderboard-tuned model is not automatically tuned to clinical need; the 2026 detection leaderboard responds to this gap.

The resection cavity is a detection problem in disguise. Cavity segmentation is accurate when a cavity exists (in-sample Dice 0.754), but the model proposes cavities in cavity-free studies often enough to halve the aggregate. The $5\,\mathrm{mm}^3$ filter removes about half of those false positives at no cost, yet it cannot remove a plausible spurious cavity of realistic size, which is what our remaining 74 false-positive studies contain. A study-level gate, for instance a classifier deciding whether a study is post-treatment before the cavity label is emitted, is the change we expect to yield the largest gain, since the cavity is one of four equally weighted regions yet is present in only one study in eight.

Limitations and future work. Training on fold `all` leaves no held-out estimate of generalisation, so the 179-study validation set is our only unbiased measurement. The error analysis above is consequently in-sample throughout: its absolute values are optimistic, and the false-positive and missed-lesion counts in particular are lower bounds on what the model would produce on unseen data. We rely on it only for the direction and relative size of the effects it exposes, which are large, monotonic and mechanistically plausible, and not for their magnitude. Retaining a fold would have supplied out-of-sample evidence, but at the cost of withholding roughly a fifth of the 167 cavity-bearing and small-lesion studies from training, which for labels this rare we judged the worse trade; confirming the analysis out of fold would require a cross-validation run beyond the compute allocation available for this entry. Training also stopped early at epoch 982, and the deployed weights come from the epoch-934 checkpoint selected on nnU-Net's foreground-oversampled training-patch pseudo-Dice, which is an internal training signal rather than an independent model-selection criterion. Fig. 1(b) shows the cavity class still improving slowly at that point, so a completed or longer

schedule would plausibly add a little further accuracy, most likely on that label. We ran neither test-time augmentation nor ensembling, both typically worth one to two Dice points, and we compared no alternative planner, loss function or post-processing variant, so the reported scores characterise one configuration rather than isolating the contribution of any individual design choice. The lesion-size and lesion-burden effects are resolved at the level of the study, using total annotated enhancing volume and component count, rather than at the level of the individual lesion; they therefore establish that studies dominated by small lesions are the hard ones, but they do not by themselves give a per-lesion detection rate as a function of lesion volume, which would need lesion-to-lesion matching and is the natural next measurement. The analysis also treats connected components of the annotation as lesions, merging touching ones and splitting fragmented ones. Our $5\,\mathrm{mm}^3$ threshold was fixed a priori from the previous edition rather than tuned, and was in place before any submission to the evaluation server, so no unfiltered official score exists against which its effect on unseen data could be quantified. Next steps follow from the analysis: the cavity gate above, test-time augmentation and a small ensemble, a loss emphasising small foreground components such as the top-k cross-entropy variant the 2025 winner evaluated, and semi-supervised use of the unannotated 2026 studies.

References

1. Astaraki, M., Sajadi Moghadam, F., 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. LNCS, vol. 16376, pp. 112–126. Springer (2026). https://doi.org/10.1007/978-3-032-16365-3_10
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021). https://doi.org/10.48550/arXiv.2107.02314
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
4. Bakas, S., Linguraru, M.G., Fathi Kazerooni, A., Farahani, K., Astaraki, M., Maleki, N., et al.: BraTS 2026 cluster of challenges. Zenodo (2026). https://doi.org/10.5281/zenodo.19714728
5. Bancerek, M., Rudzki, P., Nalepa, J.: Segmentation of pre and post-treatment brain metastases using nnU-Nets. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. LNCS, vol. 16376, pp. 161–172. Springer (2026). https://doi.org/10.1007/978-3-032-16365-3_15
6. Capellán-Martín, D., Parida, A., Jiang, Z., et al.: Adaptable segmentation pipeline for diverse brain tumors with radiomic-guided subtyping and lesion-wise model ensemble. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. LNCS, vol. 16376, pp. 456–467. Springer (2026). https://doi.org/10.1007/978-3-032-16365-3_41

7. Cho, S.J., Sunwoo, L., Baik, S.H., Bae, Y.J., Choi, B.S., Kim, J.H.: Brain metastasis detection using machine learning: a systematic review and meta-analysis. Neuro-Oncology **23**(2), 214–225 (2021). https://doi.org/10.1093/neuonc/noaa232
8. 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**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
9. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jaeger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI 2024). LNCS, vol. 15009, pp. 488–498. Springer (2024). https://doi.org/10.1007/978-3-031-72114-4_47
10. Karargyris, A., Umeton, R., Sheller, M.J., Aristizabal, A., George, J., Wuest, A., Pati, S., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence **5**, 799–810 (2023). https://doi.org/10.1038/s42256-023-00652-2
11. 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. LNCS, vol. 16376, pp. 173–181. Springer (2026). https://doi.org/10.1007/978-3-032-16365-3_16
12. 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
13. Lin, N.U., Lee, E.Q., Aoyama, H., Barani, I.J., Barboriak, D.P., Baumert, B.G., et al.: Response assessment criteria for brain metastases: proposal from the RANO group. The Lancet Oncology **16**(6), e270–e278 (2015). https://doi.org/10.1016/S1470-2045(15)70057-4
14. Machura, B., Kucharski, D., Bozek, O., Eksner, B., Kokoszka, B., Pekala, T., et al.: Deep learning ensembles for detecting brain metastases in longitudinal multi-modal MRI studies. Computerized Medical Imaging and Graphics **116**, 102401 (2024). https://doi.org/10.1016/j.compmedimag.2024.102401
15. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (2024). https://doi.org/10.1038/s41592-023-02151-9
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
17. Moawad, A.W., Janas, A., Baid, U., Ramakrishnan, D., Saluja, R., Ashraf, N., Maleki, N., Jekel, L., Yordanov, N., Fehringer, P., et al.: The brain tumor segmentation (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI. arXiv preprint arXiv:2306.00838 (2024). https://doi.org/10.48550/arXiv.2306.00838
18. Reinke, A., Tizabi, M.D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A.E., et al.: Understanding metric-related pitfalls in image analysis validation. Nature Methods **21**(2), 182–194 (2024). https://doi.org/10.1038/s41592-023-02150-0
19. Umeh, F., Yordanov, N.Y., Maleki, N., Amiruddin, R., Moawad, A., Pytlarz, M., Chukwurah, C., Aboian, M.: Taking advantage of MONAI and DiNTS frameworks to develop a state-of-the-art algorithm for automatic segmentation of brain

metastases. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. LNCS, vol. 16376, pp. 182–189. Springer (2026). https://doi.org/10.1007/978-3-032-16365-3_17