Rare-Region Optimization and Ensemble Composition for Pre- and Post-Treatment Brain Metastases Segmentation

Andres Felipe Romero Grajales[1 [0000−0001−7639−5721]]

Center for Research and Formation in Artificial Intelligence (CinfonIA), Universidad de los Andes, Bogotá, Colombia
af.romerog1@uniandes.edu.co

Abstract. Brain metastases segmentation in the pre- and post-treatment setting requires delineating four regions: non-enhancing tumor core, surrounding FLAIR hyperintensity, enhancing tumor, and, in post-treatment cases, the resection cavity (RC). RC is by far the hardest: it appears in only 12.9% of training cases and is frequently abandoned during training. We address BraTS 2026 Task 1 with a residual-encoder nnU-Net and attack RC at three levels. Two independent observations show that exposing a fold to RC at the population rate is not sufficient for RC to be learned: a fold stratified so that RC appears in 12.7–13.1% of its training cases, matching the dataset prevalence, produced zero RC voxels, and a baseline fold learned RC by epoch 27 and lost it at epoch 29, remaining at exactly zero for 471 further epochs with its data unchanged. An architecture-level probe diverged before convergence and is inconclusive. An optimization-level intervention (foreground oversampling with a class-weighted cross-entropy term on RC) raises RC lesion-wise Dice from 0.394 to 0.461, but only when paired one-to-one with a strong RC-competent fold; diluted into a four-model average the gain disappears, and control pairings of two baseline folds do not reproduce it. Instance-level analysis shows the residual error is false-positive dominated: the true-positive count is unchanged across every ensemble built on the specialist. A connected-component filter removing sub-100 mm^3 RC components cuts false positives from 128 to 11, raising RC Dice to 0.518, although with RC scored on only 22 of 179 cases this is not statistically significant. Our final configuration attains lesion-wise Dice of 0.650, 0.681, 0.649, and 0.518 for enhancing tumor, tumor core, whole tumor, and resection cavity. For rare regions, ensemble composition matters more than ensemble size, and the number of scored cases limits what any single comparison can establish.

Keywords: Brain metastases · Image segmentation · nnU-Net · Class imbalance · Model ensembling · Post-treatment imaging.

1 Introduction

Brain metastases (BM) are the most common intracranial tumors in adults, occurring in 10–30% of cancer patients, and their incidence is rising with improved systemic therapies and more sensitive MRI [1,2]. Accurate delineation underpins treatment planning and response assessment, yet manual segmentation is labor-intensive and subject to inter-rater variability [3]. The BraTS challenge has benchmarked brain tumor segmentation since 2012 [4]; the BraTS-Metastases track was introduced in 2023 [3] and later expanded to post-treatment imaging [5]. BraTS 2026 Task 1 addresses pre- and post-treatment studies jointly under a four-label scheme: non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), enhancing tumor (ET), and resection cavity (RC).

Two properties of the task shape an effective solution. First, the evaluation is a statistically-aware, rank-based scheme over many region-level metrics in which missing predictions are not penalized to the worst possible value; it therefore rewards consistency across regions rather than peak performance on any one, and the marginal value of improving a *weak* region exceeds that of further optimizing a strong one [6,7]. Second, and centrally for this work, the resection cavity is severely under-represented, appearing in only 12.9% of training cases, and is the dominant source of error in post-treatment studies. Together these identify the resection cavity as the binding constraint on rank-based performance. Our contributions follow from that diagnosis. Our contributions are the following. We present **a three-level analysis of a rare region**, attacking RC at the data level (RC-stratified cross-validation), the architecture level (a transformer encoder), and the optimization level (targeted oversampling with a class-weighted loss), reporting all three outcomes including the negative ones. We give **evidence that RC failure is not an exposure problem**: a fold stratified at population prevalence learns no RC, and a baseline fold learns RC and then loses it mid-training with its data held fixed. We show **ensemble composition matters more than ensemble size**: paired one-to-one the specialist raises RC Dice from 0.394 to 0.461, while diluted one-in-four its contribution vanishes and control pairings of two baseline folds do not reproduce it. Finally we offer **a diagnosis of the residual error and its statistical limits**: RC error is false-positive-dominated, and a connected-component filter raises RC Dice to 0.518, but not significantly on the 22 cases where RC is scored, which we treat as a finding about rare-region evaluation.

2 Methods

2.1 Dataset and Preprocessing

We use the BraTS 2026 Task 1 dataset [5]: multi-parametric MRI (native T1, post-contrast T1, T2, T2-FLAIR) of pre- and post-treatment brain metastases annotated under the four-label BM scheme, with 1,296 training and 179 validation cases. RC is present in only 167 of the training cases (12.9%), making it

by far the rarest region. We adopt the standard nnU-Net preprocessing pipeline: resampling to $1.0 \times 0.90 \times 0.86$ mm^3 (the median spacing after transposition) and per-channel z-score normalization over a foreground mask.

2.2 Network Architecture and Baseline Training

We employ the residual-encoder nnU-Net (ResEnc-L) [8,9]: a 3D U-Net whose encoder has six stages with 32/64/128/256/320/320 channels and 1/3/4/6/6/6 residual blocks, $3 \times 3 \times 3$ kernels, strided-convolution downsampling, instance normalization and leaky ReLU, and a mirrored decoder with one convolution per stage; patch size $160 \times 224 \times 192$, batch size 2, deep supervision enabled. Baselines are trained 500 epochs per fold under the standard scheme [8]: SGD with Nesterov momentum 0.99, initial learning rate 10^{-2} with polynomial decay, Dice plus cross-entropy loss, default augmentation. We train folds f0–f3 of a five-fold split; the final model is not a subset of these but a pairing of one with the variant below.

2.3 RC-Targeted Optimization

Section 3.2 shows that guaranteed RC exposure does not produce RC learning. We therefore introduce a training variant that intervenes on the optimization objective and on patch selection, leaving the architecture, preprocessed data, cross-validation split, and schedule length identical to the baseline. Two modifications are applied. First, **aggressive foreground oversampling**: the fraction of training patches forced to be centered on a foreground voxel is raised from the nnU-Net default of 0.33 to 0.90. RC occupies a small volume within a small minority of studies, so under the default the loss observes RC rarely; 0.90 was chosen as an aggressive setting short of 1.0, retaining some background-centred patches. Second, an **RC-weighted cross-entropy**: the cross-entropy term is class-weighted with $w = [1, 1, 1, 1, 3]$ over (background, NETC, SNFH, ET, RC), so an RC voxel error incurs three times the penalty of any other class; the Dice term is unmodified. The weight of 3 was chosen as a moderate upweighting, large enough to counteract abandonment and small enough that we did not expect it to destabilize the dominant classes. Neither value was tuned; we return to this in Section 4.1.

Sharing plan, data, patch size, and split with the baselines, its outputs lie on the identical voxel grid and can be averaged without resampling. We train it on fold 0 and denote it *rc*; any ensemble containing both *rc* and f0 holds two models trained on identical data, differing only in objective.

2.4 Architecture-Level Probe

We also probed architectural diversity with a Swin UNETR encoder [10] in the same framework (identical preprocessing, patch size, and split; deep supervision disabled, AdamW with cosine annealing). It diverged to a non-finite loss

at epoch 68 of 500 and was abandoned; reported numbers come from the best checkpoint before divergence (epoch 61). A model trained to 12% of its budget supports no conclusion about the architecture, and we report it only for transparency.

2.5 RC False-Positive Suppression

As Section 3.3 will show, the residual RC error is dominated by spurious detections rather than inaccurate delineation. We therefore apply a connected-component (CC) filter. A connected component, a maximal set of RC-labelled voxels joined through shared faces, is precisely what the challenge metric treats as one predicted lesion, so a single stray voxel costs a whole false-positive lesion. We extract components in 3D with 6-connectivity, take each volume as voxel count times the image spacing product, and reassign to background every RC (label 4) component below a threshold T; other labels are untouched, so ET, TC, and WT are invariant.

To choose T we exploited the per-case scoring output, which reports RC true and false positives per case. Cases with zero true positives and at least one false positive contain *only* spurious components; cases with true positives and none false contain *only* correct ones. This partitions the validation set into 48 pure-false-positive and 12 pure-true-positive cases (6 mixed excluded), recovering both size distributions without the ground truth. They are almost disjoint: the 169 spurious components have a median volume of 1 mm^3 (75th percentile 8), against 142 mm^3 (75th percentile 1611, maximum 15818) for the 23 correct ones. We fixed $T = 100$ mm^3, the largest threshold at which no pure-true-positive case loses its lesion. A sweep shows the choice is not sharp below that point: $T = 50$ clears 81% of pure-false-positive cases against 85% at $T = 100$, both at no cost to true positives; degradation begins at $T = 250$, and at $T = 1000$ only 7 of 12 are retained. T is validation-tuned, held fixed for the testing phase, which is its first held-out evaluation.

2.6 Inference, Ensembling, and Implementation

At inference we apply sliding-window prediction with Gaussian patch weighting and mirroring test-time augmentation on all axes. The final segmentation is the arg-max of per-voxel softmax probabilities averaged across models, which preserves each model's confidence and has proven effective in prior BraTS editions [11,12,13]; the CC filter is applied last. All experiments use nnU-Net v2 (v2.7.0) on a single NVIDIA A100. For the testing phase the method is packaged as a self-contained Docker image reading case folders from a read-only input directory and writing one flat segmentation per case, consistent with the containerized, federated evaluation used by the challenge [14]. Source code, the custom trainer, and the container are public at https://github.com/afromerogr/brats2026-met-uniandes.

2.7 Statistical Analysis

The challenge scores four composite regions: enhancing tumor (ET), tumor core (TC = ET + NETC), whole tumor (WT = ET + NETC + SNFH), and resection cavity (RC). We report lesion-wise Dice similarity coefficient (DSC), normalised surface Dice (NSD), 95th-percentile Hausdorff distance (HD95), and small-instance detection F1. All values are means over the cases where the metric is defined, which differs by region: ET and TC are scored on 156 cases, WT on 161, and **RC on only 22 of 179**. We report mean ± SD with 95% confidence intervals from 10,000 bootstrap resamples, and compare configurations with the paired Wilcoxon signed-rank test on per-case scores. The RC sample size materially limits the power of every RC comparison.

3 Results

3.1 Training-Set Cross-Validation and the Dynamics of Collapse

Table 1 reports per-class pseudo-Dice on each fold's held-out partition at the end of training, with the longest interval in which each of the two rarest classes was predicted entirely absent. Pseudo-Dice is computed on downsampled patches during training and is *not* comparable in scale to the lesion-wise validation metric; it characterizes training behavior, not validation performance.

Two patterns are visible. First, transient collapse is the norm rather than the exception: NETC was predicted as absent for 169, 227, and 265 consecutive epochs in folds f0, f1, and f2, recovering fully in every case, and every fold shows an early RC collapse within the first 30 epochs. Second, f2 is the only run in which a collapse proved permanent. Its RC pseudo-Dice first exceeded 0.1 at epoch 23, peaked at 0.242 at epoch 27, and fell to exactly zero at epoch 29, where it remained for 471 of the remaining epochs. NETC collapsed in the same epoch and recovered at epoch 294; SNFH (0.864) and ET (0.756) were unaffected throughout. The training data was fixed: a region demonstrably learned by epoch 27 and extinguished by epoch 29, while two regions continued improving and a third recovered 265 epochs later, cannot be explained by how often RC appeared in the training sample.

3.2 Does RC-Stratified Cross-Validation Help?

The per-fold variance in RC suggests an intuitive remedy: stratify the folds on RC presence so each is guaranteed a representative share of the 167 RC cases. We generated a five-fold stratified split in which every validation partition contains 33–34 RC cases (12.7–13.1%), matching the population rate, and trained a fold with the same architecture and schedule. Contrary to expectation it still failed to learn RC, producing zero RC voxels on the validation set (RC Dice 0.000; ET 0.645, TC 0.667, WT 0.603) while performing comparably to other single folds elsewhere. With the f2 trajectory of Section 3.1, this gives two independent observations pointing the same way: exposure at the population prevalence is

Table 1. Training-set cross-validation: final per-class pseudo-Dice on each fold's held-out partition, and the longest consecutive interval in which the class was predicted absent. *rc* is the RC-targeted variant, trained on the same fold-0 split as f0. Pseudo-Dice is a training-time patch metric, not on the scale of the lesion-wise validation metric.

Run	NETC	SNFH	ET	RC	dead interval (ep.)	
					NETC	RC
f0	0.803	0.892	0.857	0.691	169	8
f1	0.854	0.895	0.821	0.607	227	27
f2	0.898	0.900	0.838	**0.000**	265	**471**
f3	0.831	0.907	0.836	0.733	3	25
rc (fold 0)	0.714	0.877	0.730	0.711	2	7

not sufficient, and the failure mode is abandonment during optimization rather than absence from the sample. Each is a single run, and neither establishes the mechanism by which abandonment occurs.

3.3 Interventions and the Role of Ensemble Composition

Table 2 collects the baseline fold ensembles, the three intervention levels, and the composition experiments, all against the f0+f1+f3 baseline (RC 0.394).

Among baseline ensembles, adding the RC-dead fold f2 reduces RC Dice from 0.381 to 0.327: a fold that lost RC degrades the softmax average rather than being harmlessly outvoted. The data-level intervention fails outright; the architecture probe reaches RC 0.138 with degraded performance everywhere, characterizing an unstable, under-trained model rather than the architecture. The optimization-level fold alone reaches RC 0.389, indistinguishable from the baseline.

Its value appears only in combination, and only in the right one. Averaged one-to-one with the RC-competent fold f3, RC rises to 0.461 while ET, TC, and WT stay within 0.005 of the baseline. Diluted one-in-four into f0+f1+f3 the same model yields RC 0.385, below the baseline it was added to. Two controls test whether pairing with f3 suffices: vanilla pairings f0+f3 and f1+f3 reach 0.366 and 0.401, straddling the 0.394 baseline and both well short of 0.461. Of the paired comparisons only rc+f3 versus f0+f3 reaches significance (+0.096, 95% CI [+0.018, +0.187], $p = 0.024$); rc+f3 versus f1+f3 does not (+0.060, $p = 0.42$), nor does pairing versus dilution (+0.076, $p = 0.42$).

The mechanism is visible in the instance-level counts of Table 3. Across every configuration built on the RC-oversampled fold, the RC true-positive count is held at exactly 20 lesions: neither dilution nor pairing changes what the specialist detects. What varies is the false-positive count, falling from 171 under dilution to 128 for the one-to-one pairing, with RC Dice rising as it falls. The control pairings sit outside this family, with different true-positive counts (19 and 21) and far higher false-positive counts (250 and 195). Under lesion-wise Dice every spurious

Table 2. Interventions and ensemble composition, validation lesion-wise DSC, relative to the f0+f1+f3 baseline. *rc* denotes the RC-oversampled fold-0 model of Section 2.3. Only *rc* paired one-to-one with f3 improves RC; dilution and vanilla control pairings do not. The final row adds the connected-component (CC) filter. Column maxima in bold; the ET, TC, and WT maxima fall on configurations other than the final model.

Level / role	Configuration	ET	TC	WT	RC
Baseline folds	f0+f1	**0.661**	0.691	0.645	0.381
	f0+f1+f2	0.654	0.684	0.652	0.327
	f0+f1+f3	0.655	0.679	0.647	0.394
Data	RC-stratified split	0.645	0.667	0.603	0.000
Architecture	Transformer probe*	0.391	0.405	0.426	0.138
Optimization	*rc* alone	0.650	0.673	0.636	0.389
Dilution (1:4)	*rc* + f0+f1+f3	**0.661**	0.689	0.646	0.385
Control pair	f0 + f3	0.660	**0.692**	**0.660**	0.366
Control pair	f1 + f3	0.648	0.676	0.645	0.401
Pairing (1:1)	*rc* + f3	0.650	0.681	0.649	0.461
Final	*rc* + f3 + CC	0.650	0.681	0.649	**0.518**

* under-trained (epoch 61/500, diverged); not an architectural conclusion.

component contributes a zero to the per-lesion average, so suppressing them raises the score without detecting anything new. Equal-weight softmax averaging explains dilution symmetrically: three generalists outvote one specialist, pulling borderline RC probabilities below threshold, whereas a single peer leaves the specialist enough weight to prevail where confident.

3.4 Suppressing RC False Positives, and Its Statistical Limits

The CC filter removed 184 RC components and reduced the cases carrying any RC prediction from 66 to 24, close to the 23 expected at the 12.9% prevalence observed in training. RC false positives fall from 128 to 11 at a cost of two true positives, RC Dice rises from 0.461 to 0.518, detection F1 from 0.217 to 0.516, and HD95 falls from 126.9 to 110.0 mm. ET, TC, and WT are unchanged to four decimal places, as only label 4 is modified.

The paired analysis qualifies this substantially. The filter alters the score in only 10 of the 22 cases where RC is scored, improving 6 and worsening 4, giving a mean paired difference of +0.057 RC Dice with 95% CI $[-0.000, +0.128]$ and $p = 0.28$. **The improvement is not statistically significant.** We report the leaderboard mean because it is the challenge's ranking quantity, but a reader should not treat the 0.461 to 0.518 change as a demonstrated effect. The mechanism, that RC score is governed by false-positive count, is supported by the instance counts of Table 3; the magnitude is not resolvable at $n = 22$.

Figure 1 illustrates the intended behavior and its cost: a 9745 mm^3 cavity retained unchanged, and a 93 mm^3 true-positive component removed because it falls 7 mm^3 below the threshold. Of the two true positives lost across the

Table 3. Resection cavity in detail on the validation set. RC is scored on 22 of 179 cases; DSC is mean ± SD with a 95% bootstrap CI. Instance counts are totals across all cases. Within the *rc* family the true-positive count is constant and RC Dice tracks the false-positive count; the control pairings have different true-positive counts and are not part of that series.

Configuration	RC DSC	95% CI	NSD	HD95	TP	FP	F1
f1 + f3 (control)	0.401 ± 0.340	[0.261, 0.542]	0.324	162.3	19	250	0.185
f0 + f3 (control)	0.366 ± 0.332	[0.232, 0.501]	0.310	151.6	21	195	0.206
rc alone	0.389 ± 0.329	[0.257, 0.526]	0.351	148.2	20	174	0.216
rc + f0+f1+f3	0.385 ± 0.329	[0.253, 0.522]	0.310	164.2	20	171	0.202
rc + f3	0.461 ± 0.347	[0.320, 0.607]	0.372	126.9	20	128	0.217
rc + f3 + CC	$\mathbf{0.518} \pm 0.344$	[0.377, 0.656]	**0.386**	**110.0**	18	**11**	**0.516**

validation set, one was matched at an RC Dice of 0.001, a technically matched lesion with negligible overlap; the other reduced a case from two matched cavities to one.

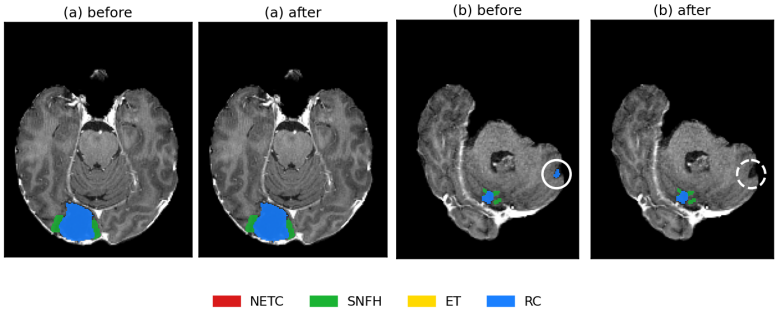

Fig. 1. Connected-component filtering, before and after. (a) A 9745 mm^3 cavity is retained unchanged. (b) A 93 mm^3 true-positive component (circled) is removed because it falls 7 mm^3 below $T = 100$ mm^3.

3.5 Is the Residual Error Delineation or Detection?

Constant true-positive counts do not by themselves establish that the matched lesions are well delineated. We therefore restricted the analysis to cases with at least one RC true positive and no RC false positives, so the score reflects delineation alone. Across the ensemble configurations matched-lesion quality is broadly flat: DSC 0.739 ($n = 9$) under dilution, 0.742 ($n = 12$) for the one-to-one pairing, and 0.727 ($n = 15$) after filtering, with NSD 0.60, 0.63, and 0.55 respectively. This supports the claim that composition moves the aggregate score

through false positives rather than boundary accuracy. Two qualifications follow: the specialist *alone* delineates measurably worse (0.603 DSC, 20.6 mm HD95, $n = 11$), so pairing improves delineation and not only false-positive count; and the filtered configuration's HD95 rises to 15.7 mm because the filter cleared false positives from three additional cases whose boundary accuracy is poor, which enter the clean subset only after filtering.

3.6 Small-Lesion Detection

Enhancing-tumor small-instance F1 is 0.382 for the final configuration and 0.404 for *rc* alone, with tumor core comparable (0.391, 0.412) and whole tumor lower (0.288, 0.292). Small-instance RC detection remained at exactly zero for *every* configuration evaluated. The validation set contains seven small RC lesions over six cases; none was detected by any model here, and the CC filter forecloses them by construction. The filter removes a capability we never demonstrated, but removes it permanently.

4 Discussion

Our results support a consistency-first design under the BraTS 2026 evaluation: because the ranking aggregates many region-level metrics and does not penalize missing predictions to the worst value, improving a weak region is worth more at the margin than further optimizing a strong one [6,7]. Two independent observations locate the RC difficulty away from data exposure: a stratified fold at population prevalence learned no RC, and a baseline fold lost RC within two epochs while its data was unchanged and two other regions kept improving. Acting on the optimization recovers the region, and the training curves suggest it changes optimization dynamics rather than only final scores: RC pseudo-Dice at epochs 100, 250, and 400 is 0.464, 0.577, 0.708 for the RC-targeted model against 0.308, 0.085, 0.616 for the baseline on the identical split, and the long NETC dead interval shrinks from 169 epochs to 2. This is one run per condition and does not separate oversampling from class weighting.

The composition finding is the most transferable result here. Standard practice is to average all available folds; for a rare region this is harmful in two ways, since a fold that lost the region degrades the average and a fold that *specializes* in it is neutralized by dilution. Both follow from equal-weight softmax averaging treating a minority opinion as noise, and both argue for deliberate, metric-informed model selection over uniform aggregation.

The instance-level decomposition sharpens this into a second claim, with an important caveat. Across every ensemble built on the specialist the true-positive count is constant and matched-lesion quality is flat, so the aggregate RC score is governed by how many spurious components accompany the detections. Removing them by a volume filter produces the largest RC gain in this study, but exploits a mismatch between what the network optimizes (voxel-wise overlap, under which a single-voxel error is negligible) and what the challenge measures

(lesion-wise Dice, under which a single-voxel component is a whole false lesion scoring zero). Rare regions are especially exposed, their score being averaged over few instances, and that same scarcity limits what we can conclude: with RC scored on 22 cases, neither the filter's improvement nor the pairing-versus-dilution difference is statistically significant, despite both being visible in the ranking metric. A region scored on a twelfth of the cohort separates methods in the leaderboard ordering long before those differences are resolvable, and its rankings should be read accordingly.

A separate unresolved failure is small-instance RC detection, exactly zero in every configuration: neither overlap-based training, probability averaging, RC-stratified splitting, nor RC-targeted oversampling recovered any of the seven small cavities, and our filter forecloses them below its threshold. A detection-oriented component rather than a segmentation loss may be required, and a probability-weighted or morphology-aware criterion in place of a fixed volume cutoff is the natural next step: it could suppress low-confidence specks while letting a small high-confidence cavity survive.

4.1 Limitations

Our RC-targeted variant changes foreground oversampling and cross-entropy weighting simultaneously, so their individual contributions cannot be separated here. The four-way ablation requires three further 500-epoch trainings and leaderboard evaluation, and the leaderboard closed on 30 July 2026; our logs show RC pseudo-Dice still fluctuating past epoch 400, so a shortened schedule would not support a reliable comparison. Neither the oversampling fraction nor the class weight was tuned, and nested validation within the training set was not attempted. The threshold T was likewise selected on the validation set and evaluated there. The architecture probe diverged before convergence and supports no conclusion about transformer encoders. All models were trained on the original release; the quality-controlled revision issued 21 July 2026 was not incorporated. Metrics beyond HD95 and voxel-wise overlap require validation ground truth, which is not released. Results are validation-only; testing-set results were unavailable at submission.

Acknowledgments. The author thanks Universidad de los Andes for its institutional and academic support, and is grateful to Prof. Pablo Arbeláez for his educational guidance and for insightful in-person discussions about the challenge. Finally, the author thanks his mother and grandmother for their continued support.

Use of AI tools. An AI assistant (Claude, Anthropic) was used for manuscript drafting and revision and for analysis scripting. All experimental design, execution, and verification of results were performed by the author, who takes full responsibility for the content of this article.

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

References

1. Lamba, N., Wen, P.Y., Aizer, A.A.: Epidemiology of brain metastases and leptomeningeal disease. Neuro-Oncol. **23**(9), 1447–1456 (2021)
2. Achrol, A.S., Rennert, R.C., Anders, C., Soffietti, R., Ahluwalia, M.S., Nayak, L., et al.: Brain metastases. Nat. Rev. Dis. Primers **5**(1), 5 (2019)
3. Moawad, A.W., Janas, A., Baid, U., Ramakrishnan, D., Saluja, R., Ashraf, N., et al.: The Brain Tumor Segmentation – Metastases (BraTS-METS) Challenge 2023: Brain Metastasis Segmentation on Pre-treatment MRI. arXiv:2306.00838 (2024)
4. 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 Trans. Med. Imaging **34**(10), 1993–2024 (2015)
5. Maleki, N., Amiruddin, R., Moawad, A.W., Yordanov, N., Gkampenis, A., Fehringer, P., 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:2504.12527 (2025)
6. 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. Nat. Methods **21**(2), 182–194 (2024)
7. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., et al.: Metrics reloaded: recommendations for image analysis validation. Nat. Methods **21**(2), 195–212 (2024)
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. Nat. Methods **18**(2), 203–211 (2021)
9. 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. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI 2024), pp. 488–498. Springer, Cham (2024)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: 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), pp. 272–284. Springer, Cham (2022)
11. Capellán-Martín, D., Jiang, Z., Parida, A., Liu, X., Lam, V., Nisar, H., et al.: Model ensemble for brain tumor segmentation in magnetic resonance imaging. In: Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation (BraTS 2023, CrossMoDA 2023), pp. 221–232. Springer, Cham (2024)
12. Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2017), pp. 450–462. Springer, Cham (2018)
13. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnU-Net for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2020), pp. 118–132. Springer, Cham (2021)
14. 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. Nat. Mach. Intell. **5**, 799–810 (2023)