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.

Ensemble nnU-Nets for pre- and post-operative Brain Metastases Segmentation in BraTS 2026 Task 1

Alessia Lindemann[1,2], Nadia Milanesi[2], Luca A. Cappellini[1,2], Giovanni Savini[1,2], Letterio S. Politi[1,2], Riccardo Levi[1,2]
1 Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy 2 IRCCS Humanitas Research Hospital, Rozzano, Milan, Italy
alessia.lindemann@st.hunimed.eu

Abstract. Automated brain metastases segmentation must combine sensitivity to small lesions with reliable semantic separation of enhancing tumor (ET), non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), and resection cavity (RC). We present a clinically guided multistage inference pipeline for BraTS 2026 Task 1. Two complementary nnU-Net v2 residual-encoder segmentation sources are fused with an internal OOF-selected recall-dominant weight of 0.7/0.3 and combined with auxiliary small-lesion detection and RC-specific evidence. Probability-aware re-fusion and connected-component post-processing filter low-confidence ET/NETC components, suppress RC predictions on baseline scans identified by the supplied longitudinal suffix, relabel uncertain RC according to local class probabilities and shell context, and consolidate RC at connected-region level. The ensemble weight and RC-consolidation operating point were selected by internal OOF proxy sweeps; semantic RC thresholds were expert-defined from qualitative failure-mode review and were not numerically optimized against validation labels or leaderboard scores. On the 179-case official validation set, the final pipeline achieved lesion-wise instance F1 of 0.7997 for ET, 0.4901 for RC, 0.7906 for TC, and 0.7548 for WT. The method provides an interpretable framework for targeted RC refinement after complementary evidence has been combined.

Keywords: brain metastases segmentation · BraTS 2026 · nnU-Net · ensemble · semantic refinement · resection cavity

1 Introduction

Brain metastases are the most common intracranial tumors in adults, occurring in approximately 20% of patients with systemic cancer and arising most frequently from primary lung, breast, and melanoma tumors [1]. Surgical resection is frequently indicated for large, symptomatic, or diagnostically uncertain lesions, yet resection alone does not guarantee durable local control; postoperative clinical assessment of the resection cavity and adherence to structured follow-up imaging are therefore critical for detecting local recurrence early and guiding adjuvant treatment [2]. Brain metastases are commonly monitored using unidimensional measurements under RANO-BM criteria, yet volumetric estimates of tumor and surrounding signal abnormality can provide complementary information for clinical assessment [3]. Manual delineation is labor-intensive, particularly in patients with multiple lesions, and small metastases remain difficult to detect and segment reliably. Prior deep-learning approaches to automatic brain

metastasis segmentation have reported high accuracy but were largely developed and validated on limited, single-institution cohorts, raising concerns about their generalizability across scanners, centers, and patient populations [4], underscoring the need for large, open, multi-institutional datasets such as BraTS-METS.

BraTS 2026 Task 1 requires automated detection and delineation of non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), enhancing tumor (ET), and resection cavity (RC) in both pre- and post-treatment MRI. The challenge additionally evaluates small-lesion detection, reflecting the clinical importance of lesions below 27 mm^3 [5].

The proposed solution treats final segmentation as the output of a clinically guided inference pipeline rather than a single hard prediction. Its design preserves lesion sensitivity through a recall-oriented system and dedicated detector, stabilizes tumor-core partitioning through a complementary core-aware system, and resolves residual ambiguity using component size, scan timepoint, softmax probabilities, and local context.

2 Methods

2.1 Dataset Description

Our local audit of the BraTS-METS Task 1 training distribution [11] rooted at MICCAI-LH-BraTS2025-MET-Challenge-Training identified 650 official-facing top-level training studies from 549 patients and an additional UCSD - Training subdirectory containing 646 studies, yielding 1,296 recursively discovered training studies from 810 patients for the top-1296 model preparation. Each real training study used for model preparation came from this root or one of its subdirectories. Each study included four MRI sequences: T1-weighted pre-contrast (T1W), T1-weighted post-contrast (T1C), T2-FLAIR (FLAIR), and T2-weighted (T2W). Manual annotation masks provided the following labels: background (label 0), NETC (label 1), SNFH (label 2), ET (label 3), and RC (label 4) [3]. All released scans underwent standardized preprocessing by the challenge organizers prior to distribution, including conversion from DICOM to NIfTI, rigid co-registration to the SRI24 anatomical template, resampling to 1 mm^3 isotropic resolution, and deep-learning-based skull stripping to remove non-brain tissue and reduce re-identification risk [6]. Voxel-wise intensity normalization was not applied by the organizers; per-channel z-score normalization was instead performed downstream as part of the standard nnU-Net preprocessing pipeline [7].

2.2 Deep Learning Segmentation Network

The proposed segmentation procedure used two complementary deep learning systems, defined as recall-oriented (RO-UNet) and core-aware (CA-UNet). Each system was implemented with nnU-Net v2 in the 3D full-resolution residual-encoder U-Net configuration, operated on the four MRI sequences (T1W, T1C, FLAIR, and T2W), and produced the four BraTS-METS foreground labels (NETC, SNFH, ET, and RC). The submitted container expects all four modalities in either challenge-style names or nnU-Net suffixes 0000-0003; missing-modality inference was not evaluated and no imputation strategy was included in the submitted container.

The 650-case top-level cohort and the 1,296-study recursive preparation served different purposes. The 650-case cohort was used for the formal internal patient-wise five-fold OOF metric report, with exactly 130 cases per fold and no patient split across folds. The 1,296-study preparation was also split patient-wise for OOF prediction generation: the audited fold file contained 1,296 studies from 810 patients, with 258-260 studies per fold and no patient assigned to more than one fold. The final U-Net + Step 4 system itself did not train a new network; it is a post-hoc inference build that combines the CA segmentation source, the RO checkpoint family, softmax fusion, RC-gating/consolidation rules, and the final semantic RC relabeling stage. The cumulative validation configurations denoted U-Net + Step 1 through U-Net + Step 4 use these sources and deterministic stages in increasing order. The ensemble weight was selected on matched OOF predictions from the CA and RO segmentation sources using a stride-3 sweep of 432 cases from the 1,296-study preparation. The official validation leaderboard was used to evaluate successive challenge configurations, but neither validation annotations nor validation leaderboard scores were used to select the ensemble weight or post-processing thresholds; those were selected from OOF training predictions or predefined expert rules. No external image or label source was used for final model preparation; the recall-oriented RC branch metadata count of 1,496 reflects 1,296 real training studies plus 200 training-only copy-paste augmented RC cases generated from cavities and target brains within the same training root and added only to training folds.

The RO checkpoint family was prepared on the top-1296 training set from the challenge training root and included three independently trained nnU-Net checkpoints for four-class segmentation, small-lesion detection, and auxiliary RC evidence. The segmentation checkpoint was trained for 500 epochs with a loss defined as a weighted combination of multiclass Tversky and cross-entropy terms [8]. To avoid ambiguity with alternative Tversky notations, our trainer implemented the foreground Tversky index as (TP + s)/(TP + alpha FN + beta FP + s); therefore alpha weights false negatives in this code. This alpha/beta convention differs from formulations in which alpha weights false positives. The configured alpha = 0.7 and beta = 0.3 penalized false negatives more than false positives, favoring recall of under-segmented structures while preserving voxel-wise class discrimination through cross-entropy. The detection checkpoint used the same architecture and recall-oriented trainer but was trained on derived binary masks retaining only small NETC, ET, and RC lesion components up to 27 mm^3, with locally attached SNFH context. The auxiliary RC checkpoint was trained as a separate RC-focused nnU-Net on masks containing only label 4 for resection cavity; its

training set included the 200 training-only RC copy-paste augmentations described above.

The CA segmentation source provided a cavity-aware/core-aware four-class softmax used as conservative segmentation evidence. It was used as a single segmentation branch producing both voxel-wise masks and four-class softmax probabilities for subsequent ensemble fusion. We therefore describe the submitted U-Net + Step 4 system as an inference-time ensemble of already trained sources rather than as a newly trained architecture.

At inference time, the softmax outputs of these two base segmentation sources were combined through an internal OOF grid search over candidate RO weights (0, 0.3, 0.5, 0.7, and 1.0), optimizing proxy lesion-level F1 for ET, TC, and WT on matched CA/RO OOF predictions. The sweep evaluated 432 stride-sampled OOF cases from the 1,296-study preparation and selected a recall-dominant 0.7 weight for RO-UNet and 0.3 for CA-UNet. The resulting ensembled segmentation was then further refined by branch fusion and post-processing steps designed to improve false-positive control and RC-aware labeling.

2.3 Semantic Refinement and Optimization Pipeline

Following the prediction obtained by CA and RO-UNet, an optimization post processing pipeline was designed to further improve detection and segmentation of brain lesions. The complete pipeline comprises six stages:

1. four-class softmax weighted ensemble of the RO and CA UNet segmentations;
2. re-fusion with small-lesion detection and RC predictions;
3. probability-based component filtering for ET and NETC;
4. baseline-timepoint RC gating;
5. semantic RC relabeling using component probabilities and local shell context;
6. RC-aware connected-region consolidation.

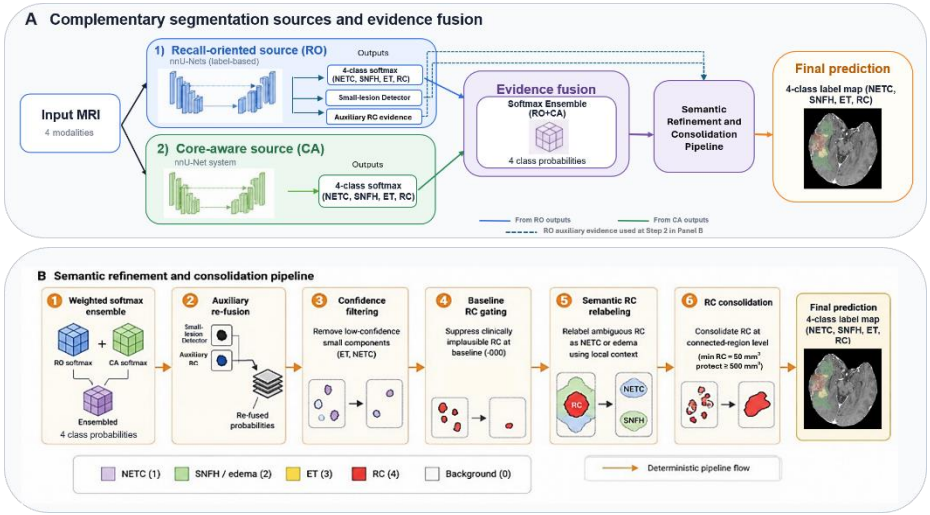

Fig. 1. Overview of the optimized inference pipeline. The RO stream supplies a four-class segmentation softmax, an auxiliary small-lesion detector, and auxiliary RC evidence; the CA stream supplies a complementary four-class segmentation softmax. The six deterministic stages are shown in the same order as described in the text.

1) Softmax Ensemble of Complementary Segmenters

The four-class softmax outputs of RO-UNet and CA-UNet are combined voxel-wise with weights of 0.7 and 0.3, respectively. These weights were selected by the internal OOF grid search described above, rather than fitted on validation labels. In the local sweep, wA = 0.7 achieved the highest mean proxy lesion-level F1 across ET, TC, and WT (0.7133), compared with 0.7095 for wA = 1.0 and 0.6878 for wA = 0.0.

The ensembled softmax is converted to a multiclass label map based on the highest probability class, while the probability map is retained for later confidence filtering and correct attribution of RC class.

2) Detection and RC Re-fusion

After softmax ensembling, the ensembled label map is fused with two auxiliary outputs from RO-UNet. The detection output is a binary small-lesion support map trained from small NETC, ET, and RC components with locally attached SNFH context; the auxiliary RC output is cavity-specific evidence used during fusion and is distinct from the RC class already present in the four-class segmentation softmax.

During re-fusion, detection-supported candidate components are promoted only when they are sufficiently small and not already covered by the segmentation core; this logic uses the detector as a small-lesion rescue signal rather than as a separate semantic class. The configured fusion retains detection components up to 250 mm^3 and ignores components whose overlap with the segmentation core exceeds 0.8. These re-fusion guardrails (250 mm^3 maximum detection-component size, 0.8 maximum core-overlap

fraction, and the 125 mm³ auxiliary-RC minimum described below) were expert-fixed implementation constants chosen to restrict auxiliary branches to small or anatomically plausible component-level corrections; they were not included in the numerical OOF sweeps used for the ensemble weight, confidence filter, or RC-consolidation operating point.

Detection and RC re-fusion is performed at connected-component level to keep the branch roles explicit. Let TC = ET + NETC denote the tumor-core prediction coming from the fused segmentation stream. The RC branch is then treated as complementary evidence that can modify only the TC interior, not the outer lesion support. Operationally, each connected RC component is considered only if its size is at least 125 mm³ and if it lies outside enhancing tumor. In set terms, the update acts on TC intersect RC and resolves the local competition between ET, NETC, and RC, whereas TC regions without RC support remain unchanged. This means that RC can relabel plausible cavity-like portions of the core, but it is not allowed to overwrite ET voxels or to create disconnected lesion support on its own. The resulting rule can be summarized as follows: TC and RC are merged only when RC provides anatomically plausible, size-qualified evidence inside an already supported lesion core; otherwise the original TC labeling is preserved.

3) Confidence-Based Component Filtering

Probability-based false-positive control is applied to ET and NETC. A connected component is considered for filtering only when it is both small and low-confidence. The configured thresholds are 50 mm³ for component size and 0.70 for mean softmax confidence for both ET and NETC. They were retained from an earlier OOF confidence-filter sweep over 457 stride-sampled training cases, using internal proxy lesion-level F1 at Dice overlap 0.1 as the objective and evaluating confidence thresholds 0, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8 crossed with size thresholds 0, 50, 100, 200, 400, and 800 mm³. In that sweep, the shared 0.70/50 mm³ operating point improved proxy ET F1 from 0.7064 to 0.7307 and proxy NETC F1 from 0.4731 to 0.5037, and it was kept fixed for the submitted U-Net + Step 4 build. Detection-supported components are protected.

When contextual relabeling is enabled, an uncertain small component is not necessarily removed to background. It can instead be reassigned to the dominant neighboring foreground class within a two-voxel dilation shell, treating the component as a possible local label conflict within a real lesion.

4) Baseline RC Gating

In the supplied longitudinal naming scheme, scans ending in “-000” are treated as baseline timepoints. Because RC is primarily a post-surgical finding, the pipeline suppresses RC predictions on these scans before RC-specific semantic refinement, thereby restricting subsequent cavity-versus-necrosis arbitration to non-baseline cases. In the audited 179-case validation set, 107 scans carried the “-000” suffix and 23 of those baseline predictions contained RC voxels that were removed by this gate. If a baseline scan lacks this suffix, the container will not suppress RC because it has no independent timepoint

metadata; non-baseline suffixes are processed without suppression. This challenge-specific proxy limits transferability, whereas routine clinical records ordinarily provide surgical status and treatment history.

5) Semantic RC Relabeling

Semantic RC relabeling operates on connected RC components in non-baseline cases. Each component is evaluated using its size in mm^3, mean softmax support for NETC, SNFH, ET, and RC, and the class composition of a two-voxel shell surrounding the component.

RC components of at least 500 mm^3 are preserved, as are components with mean RC probability of at least 0.45. Components touching ET context are preserved when mean RC support is no more than 0.03 below mean NETC support. A remaining weak RC component is relabeled as SNFH when at least 45% of its two-voxel shell is SNFH and mean SNFH support is within 0.03 of both NETC and RC support. It is relabeled as NETC when the shell majority is NETC and mean NETC support is within 0.03 of RC support, or when there is no ET contact and NETC support is within the same margin.

The module also permits conservative promotion from NETC to RC in non-baseline cases. A NETC component is eligible only when its volume is 80–6000 mm^3, mean RC probability is at least 0.30, RC support exceeds NETC support by at least 0.03, SNFH occupies no more than 35% of the local shell, and foreground occupies at least 60% of that shell. This bidirectional rule addresses both unsupported RC labels and plausible cavities absorbed into NETC-heavy mixed cores.

6) RC Consolidation

The final stage consolidates RC at connected-region level. The configured operating point uses tau = 0.20 for the local RC arbitration score, a minimum RC size of 50 mm^3, and RC protection for regions of at least 500 mm^3. This operating point was selected from a pool of 1,296 matched OOF cases; for computational efficiency the sweep evaluated all 167 RC-positive cases and a 400-case sample of RC-negative cases. Within this 567-case sweep set, tau = 0.20 with 500 mm^3 protection improved internal proxy RC F1 from 0.5791 for the prior tau = 0.30/no-protection baseline to 0.5808. The historical RC copy-paste augmentation was not explicitly donor/recipient-restricted within each OOF fold. Accordingly, this RC-specific OOF sweep is treated as internal development evidence rather than as an independent validation result. Volumetric thresholds are computed from the NIfTI voxel spacing, whereas dilation shells are voxel-based and therefore tied to the processed image grid.

2.4 Source Code Availability

The complete inference and post-processing source code is publicly available at https://github.com/allindemann/brats-2026-mets-task1-v18. The submitted Docker image is referenced in the Synapse submission record. The repository includes nnU-Net configurations and custom trainers, dataset-preparation scripts, softmax ensembling, detection/RC re-fusion, confidence filtering, semantic RC relabeling, connected-region

consolidation, the Dockerfile, and execution instructions required to reproduce the submitted predictions. The source-code snapshot used for revision corresponds to commit 8814f64.

3 Results

3.1 Qualitative Examples

Figure 2 shows representative validation outputs from the final pipeline. The examples illustrate the range of label configurations encountered in post-treatment and baseline cases, including a large RC-containing lesion, a small central core component, and an SNFH-dominant region. Because validation ground truth is not publicly released, the overlays are presented as qualitative illustrations of the final predictions rather than as evidence of correction accuracy [3].

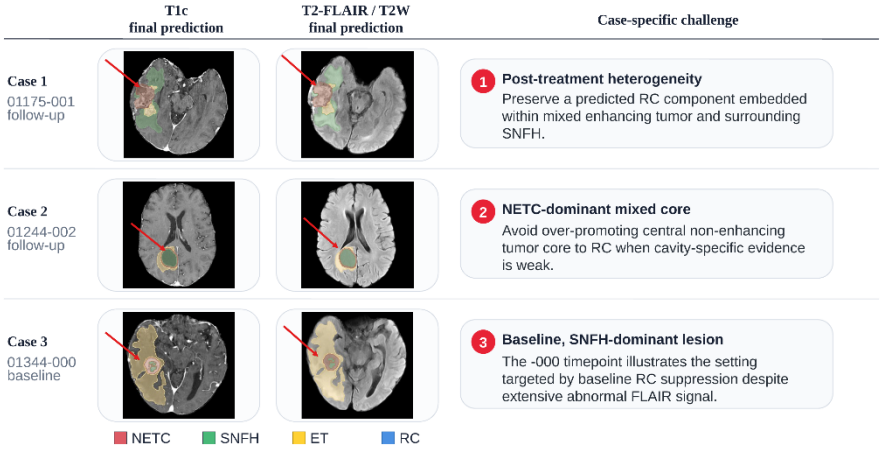

Fig. 2. Representative final validation predictions overlaid on T1c and T2-FLAIR. Rows show difficult cases; colors denote NETC, SNFH, ET, and RC.

3.2 Training and Validation Results

Training-set performance was estimated by patient-wise OOF evaluation across the top-level 650-case cohort. It is reported separately from official validation because the OOF summary uses voxel-wise regional metrics and a lesion-detection threshold sweep, whereas validation follows the challenge lesion-wise format. OOF lesion detection used Hungarian one-to-one matching of TC connected components; a pair was considered matched when its Dice met the selected threshold. At Dice threshold 0.5, the mean case-wise lesion-level F1 was 0.5083, with mean case-wise precision 0.5593 and recall 0.9533. These quantities were averaged separately across cases and therefore should not be recombined as a single pooled harmonic mean. The same OOF summary yielded WT Dice 0.8069, TC Dice 0.9197, ET Dice 0.8766, and NETC Dice 0.6812. At a lower Dice threshold of 0.1, used as a small-lesion detection operating point, the OOF mean case-wise F1 was 0.8245 with precision 0.8896 and recall 0.8211 [9, 10].

On the 179-case official validation submission, Table 1 reports cumulative pipeline upgrades in functional order. Step 1 is the cavity-aware hybrid baseline; step 2 adds the recall-oriented operating point and confidence-based false-positive control; step 3 adds the post-hoc RO-CA ensemble with RC timepoint gating and connected-region RC consolidation; and step 4 adds semantic RC relabeling. Because step 3 combines multiple operations, it should not be interpreted as an isolated ablation of each mechanism inside that group.

Mean lesion-wise instance F1 increased from 0.6012 at step 1 to 0.7088 at step 4. The final semantic RC stage produced a targeted change, increasing RC instance F1 from 0.4717 to 0.4901 without degrading ET, TC, or WT at the reported precision. In the final submission, ET and TC remained the strongest regions, with lesion-wise instance F1 values of 0.7997 and 0.7906, respectively, whereas RC remained the most challenging target.

Table 1. Validation metrics by pipeline upgrade and target region.

Pipeline step	Region	Lesion-wise instance F1	Mean lesion-wise Dice	Mean lesion-wise NSD
U-Net + Step 1	ET	0.7757	0.6957	0.7569
	RC	0.1163	0.4012	0.3161
	TC	0.7814	0.6973	0.7503
	WT	0.7314	0.6131	0.6327
U-Net + Step 2	ET	0.7971	0.6977	0.7552
	RC	0.4310	0.4878	0.3634
	TC	0.7748	0.6964	0.7417
	WT	0.7446	0.6461	0.6630
U-Net + Step 3	ET	0.7997	0.7211	0.7807
	RC	0.4717	0.4617	0.3481
	TC	0.7906	0.7343	0.7806
	WT	0.7548	0.6760	0.6954
U-Net + Step 4	ET	0.7997	0.7211	0.7807
	RC	0.4901	0.4617	0.3481
	TC	0.7906	0.7344	0.7807
	WT	0.7548	0.6760	0.6954

4 Discussion

Neuroradiological review identified four recurrent RC failure patterns: undercalled cavities within mixed NETC/RC lesions, RC assigned to central necrosis, RC assigned to SNFH-dominant regions, and RC predictions on baseline scans. These observations motivated separate lesion recovery, baseline gating, semantic relabeling, and connected-region consolidation.

The semantic RC stage targets uncertain cavity semantics while protecting plausible cavities and ET. In the cumulative validation table, its measurable effect was concentrated in RC instance F1, increasing RC from 0.4717 to 0.4901, while ET, TC, and WT remained unchanged at the reported precision and RC Dice/NSD did not improve. Without validation ground truth or a fully component-wise ablation for every sub-rule, we do not claim that each semantic rule has been independently proven; rather, we

interpret the stage as a targeted, non-regressing correction for the most ambiguous label.

4.1 Design Rationale and Limitations

The central design choice is to separate spatial false positives from semantic label conflicts. Small low-confidence ET/NETC components are filtered or reassigned from neighboring foreground context, whereas uncertain RC components are evaluated directly against NETC and SNFH support. Conversely, a non-baseline NETC component with stronger RC evidence can be promoted under the conservative criteria above.

Learned lesion recovery and semantic arbitration are therefore separated: recall-oriented segmentation and detection prioritize sensitivity, the core-aware system supplies a complementary tumor-core prior, and component-level rules act only after the probability maps have been fused.

The deterministic stages can correct only errors represented in the available probability maps and local context. The ensemble weight, confidence-filter thresholds, and RC-consolidation parameters were selected from internal OOF training predictions, whereas the Detection/RC re-fusion guardrails and semantic RC thresholds were expert-fixed constants defined from implementation constraints and qualitative neuroradiological failure-mode review; they therefore remain hand-specified constants in the final container rather than parameters optimized jointly end-to-end. The RC copy-paste augmentation used for the auxiliary RC checkpoint was training-only and used no external images or labels. Because the historical augmentation was not explicitly donor/recipient-restricted within each fold, the RC-specific OOF results and consolidation sweep are treated as internal development evidence rather than as independent validation results. The "-000" convention, absence of independent clinical timepoint metadata, and voxel-based shell operations may limit transferability. The submitted container also expects four MRI channels; no modality was missing in the audited inputs, but behavior with an absent T2W sequence was not evaluated. Future work should quantify the isolated contribution and robustness of each stage on fully held-out data with explicit timepoint metadata and missing-modality stress tests.

5 Conclusion

We present a clinically guided ensemble and semantic-refinement pipeline combining complementary segmentation, small-lesion detection, and RC-specific post-processing. By separating lesion recovery from RC arbitration and delaying hard decisions until probability fusion, the system achieved its strongest performance for ET and TC, while RC remained the most difficult target. Because the semantic stage was non-regressing rather than globally improving at the available precision, it should be interpreted as a targeted and transparent mechanism for ambiguous RC predictions.

Acknowledgements. BraTS-METS 2026: Synapse (syn74274097); challenge and dataset papers are cited [6,11]. Supported by ANTHEM (HMN076-1). **AI disclosure.** AI tools were used solely for language editing; all scientific content was reviewed and approved by the authors. **Ethics and consent**. This secondary analysis used only de-identified challenge data; no new data or identifiable patient information were collected. Source institutions obtained the required IRB, data-transfer approvals before data sharing [11]. **Competing interests.** The authors declare no competing interests.

References

1. Achrol, A.S., Rennert, R.C., Anders, C., et al.: Brain metastases. Nature Reviews Disease Primers 5, 5 (2019).
2. Mahajan, A., Ahmed, S., McAleer, M.F., et al.: Post-operative stereotactic radiosurgery versus observation for completely resected brain metastases: a single-centre, randomised, controlled, phase 3 trial. Lancet Oncology 18(8), 1040–1048 (2017).
3. Lin, N.U., Lee, E.Q., Aoyama, H., et al.: Response assessment criteria for brain metastases: proposal from the RANO group. Lancet Oncology 16(6), e270–e278 (2015).
4. Grøvik, E., Yi, D., Iv, M., Tong, E., Rubin, D., Zaharchuk, G.: Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. Journal of Magnetic Resonance Imaging 51(1), 175–182 (2020).
5. BraTS 2026 Challenge: Task 1—Brain Metastases. Synapse ID syn74274097 (accessed 26 July 2026).
6. Moawad, A.W., Janas, A., Baid, U., et al.: The Brain Tumor Segmentation–Metastases (BraTS-METS) Challenge 2023: brain metastasis segmentation on pre-treatment MRI. Machine Learning for Biomedical Imaging 3, 204–247 (2025).
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. Nature Methods 18, 203–211 (2021).
8. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: MLMI 2017, LNCS 10541, 379–387 (2017).
9. Reinke, A., Tizabi, M.D., Baumgartner, M., et al.: Understanding metric-related pitfalls in image analysis validation. Nature Methods 21, 182–194 (2024).
10. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods 21, 195–212 (2024).
11. 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).