A Detection-Aware SegResNetDS Ensemble and a Measured Null Space for Brain Metastasis Segmentation in BraTS 2026

Mahmoud Ghoraba[1], Mohamed Ghoraba[2], Mohanad Ghonim[3], and Mohamed Ghonim[4]

[1] Undergraduate Student, Software Engineering, German International University, Cairo, Egypt
mahmoud.ghoraba@student.giu-uni.de
[2] Undergraduate Student, Data Science, German International University, Cairo, Egypt
mohamed.ghoraba@student.giu-uni.de
[3] Department of Radiology, University of Arkansas for Medical Sciences, Little Rock, AR, USA
mghonim@uams.edu
[4] Department of Radiology, University of Mississippi Medical Center, Jackson, MS, USA
mghonim@umc.edu

Abstract. Brain metastases are the most common intracranial tumors, and segmenting them by hand is slow work: a single scan can carry a dozen lesions, many only a few millimeters across. We built a pipeline for the BraTS 2026 Metastases challenge using MONAI Auto3DSeg with a SegResNetDS backbone, trained from scratch on the 1,296 provided cases with no pretrained weights: five folds split by patient rather than by case, 245 epochs each on four NVIDIA H200 GPUs, averaged into one probability map and post-processed with per-region thresholds, cavity hysteresis and label-hierarchy enforcement. On the 179 validation cases this reaches lesion-wise instance F1 of 0.72 for tumor core, 0.71 for enhancing tumor, 0.69 for whole tumor, and 0.28 for resection cavity. A case-level decomposition shows the cavity figure to be precision-driven: our final two configurations detect the same 8 cavity instances in the same 23 cavity-bearing cases and differ only in false positives. We also report interventions that did not improve performance under the configurations tested, four of which reflect measured properties of the lesion-wise evaluation setting rather than of the model. Cavity detection remains unresolved: 15 of 23 cavity-bearing cases are missed under every threshold we swept.

Keywords: Brain metastasis segmentation · Lesion-wise detection · Resection cavity · MONAI Auto3DSeg · Metric behavior

1 Introduction

Brain metastases are the most common intracranial tumors in adults and develop in up to 40% of patients with solid cancers [9]. Patients may present with numerous small lesions, making manual segmentation slow and inconsistent across serial examinations, while accurate segmentation supports treatment planning, disease-burden measurement, and response assessment [21]. Small metastases remain frequently missed by automated methods, and vessels, postoperative changes and artifacts produce false positives [4, 22].

The BraTS-METS challenges provide standardized multi-institutional datasets and evaluation. The 2023 edition was won by an Auto3DSeg/MONAI model at a detection-aware lesion-wise Dice of 0.65 ± 0.25, with performance declining sharply for small lesions [14]; the 2025 edition added pre- and post-treatment MRI [12] and was won by an nnU-Net approach [1]. BraTS-METS 2026 separates detection from segmentation quality explicitly, evaluating small lesions primarily by instance-level F1 and larger ones by lesion-wise Dice and normalized surface distance. Task 1 scores enhancing tumor, tumor core, whole tumor and resection cavity, and prohibits external annotated datasets and pretrained weights [2].

This paper contributes a reproducible five-fold SegResNetDS pipeline for Task 1, a case-level decomposition showing our measured resection-cavity gain to be precision rather than recall, and four empirically characterized properties of the lesion-wise evaluation setting that constrain the interpretation of results obtained under it. Code and configuration are available at `https://github.com/MahmoudGhoraba/brain-tumor-segmentation-brats`.

2 Methods

2.1 Data

Task 1 uses the BraTS-METS corpus [14] in its 2025 release [12], contributed by eight institutions. We trained on the 1,296 labeled cases and evaluated on the 179 unlabeled validation cases, with 303 additional cases held for the test phase. The validation leaderboard is unranked, so reported validation results describe our own progress rather than a standing.

Each case contains four co-registered MRI volumes—T1, post-contrast T1, T2, and T2-FLAIR—already skull-stripped and resampled to an isotropic 1 mm^3 grid. Labels cover enhancing tumor, non-enhancing tumor core, surrounding FLAIR-hyperintense tissue, and resection cavity.

Case identifiers encode patient and timepoint. This is important because 646 of the 1,296 training cases belong to a longitudinal UCSD cohort, meaning that the same patient appears across multiple timepoints. We therefore assign folds by patient rather than by case to prevent patient-specific anatomy from appearing in both training and validation. Resection cavity labels are also highly concentrated: 151 training cases (11.6%) contain a cavity, all from the UCSD cohort.

2.2 Evaluation metric

Several findings below are properties of the scorer rather than the model, so the metric [19] is worth stating before the system tuned against it.

Predictions and labels are decomposed into connected components—*instances*—per region. A ground-truth instance counts as detected if any predicted component overlaps it after a small dilation, with no minimum overlap requirement and no penalty for piling several predicted components onto one lesion: extras yield neither an extra true positive nor an extra false positive. A predicted component overlapping nothing is a false positive.

The 27 mm^3 split. Instances are partitioned by volume into *small* ($<$ 27 mm^3) and *large* buckets, scored separately as well as pooled. On this isotropic 1 mm^3 grid that threshold is 27 voxels—a lesion roughly $3 \times 3 \times 3$—and the split exists because the challenge weights detection of sub-resolution disease differently from segmentation quality on lesions large enough to delineate. Every per-bucket figure in Section 3 refers to this partition, and the two buckets use *different case denominators* (Section 3.2).

Two aggregation properties govern the reported figures. Instance F1 is computed per case and then averaged, so a cohort value is a mean of per-case F1 rather than a pooled ratio of instance counts. A case contributes to that mean only when its ground truth or its prediction is non-empty for the region: a case empty in both is excluded, whereas the same case carrying any prediction is included with an F1 of zero. Each bucket applies this rule independently, so the small and large buckets average over different case sets and their rates are not additive (Section 3.2); Section 4.1 quantifies the effect on cavity scoring. Dice, HD95 and normalized surface distance are computed over matched instance pairs only. We reimplemented the metric locally to search configurations offline; two aspects—false-positive apportionment between buckets, and the empty-case convention above—are not confirmable against the released code, which has neither a small/large split nor cohort aggregation, and we established the latter empirically instead.

2.3 Architecture and backbone choice

We used SegResNetDS [15, 6], a 3D residual encoder-decoder with deep supervision [10] at depth four, configured and trained entirely through MONAI Auto3DSeg [3, 16]; we wrote no training-loop code. Configuration: ROI $128 \times 160 \times 128$, `init_filters=32`, `blocks_down=[1,2,2,4,4]`, instance normalization [20], which suits data whose intensities have no absolute units and shift across eight sites.

Why this backbone. We did not benchmark against nnU-Net [7], MedNeXt [17], or SwinUNETR [5], and the choice was deliberate. Auto3DSeg's SegResNetDS won BraTS-METS 2023 under the same detection-aware metric family [14], making it the strongest published prior for this task and scorer, and its resolved template is public and needs no bespoke training code. Our experimental

budget went instead to characterizing the evaluation setting and the cavity failure, which we judged more transferable than an architecture comparison already well covered by the BraTS literature.

The network emits four independent sigmoid channels rather than a softmax, because the regions nest rather than partition—whole tumor (WT) $\supseteq$ tumor core (TC) $\supseteq$ enhancing tumor (ET), plus an independent resection-cavity (RC) channel—so decoding to one label per voxel requires an explicit combination step enforcing that nesting (Section 2.5). Intensity normalization (per-channel z-scoring within a brain mask) and augmentation (random flips, 90° rotation, intensity scaling and shifting, Gaussian noise and smoothing, contrast adjustment) are unmodified from the resolved Auto3DSeg template.

2.4 Training

The 1,296/179/303 partition is fixed by the challenge; all cross-validation here is internal to its 1,296 labeled cases. Within them we used five folds assigned by the *patient* component of the case identifier rather than by case index. Because 646 training cases come from a single longitudinal cohort, an index-based split would place different timepoints of one patient on both sides of a fold boundary, letting patient-specific anatomy be seen in training and scored in validation—acute for a resection cavity, whose shape persists across a patient's own follow-up. Held-out fold sizes were 251, 267, 256, 252, and 270. We trained on four NVIDIA H200 GPUs (143 GB each), folds 0–3 concurrently and fold 4 queued behind fold 0, at batch size nine with two crops per image (Auto3DSeg initially selected five, which exceeded memory at this resolution). Every fold ran the full 245 epochs, checkpointing every two and selecting the best by average held-out Dice over the four regions—best-checkpoint selection over a fixed schedule, not early stopping. Optimization used AdamW [11] at learning rate 2×10^{-4} with warmup-cosine schedule and weight decay 10^{-5}. The loss was `DiceCELoss` [13] with `include_background=True`, `squared_pred=True`, `smooth_nr=0`, `smooth_dr=`10^{-5}, applied to the sigmoid outputs at all four deep-supervision scales. We used no gradient clipping, dropout, or custom initialization, and did not fix random seeds or enforce deterministic kernels—a defect discussed in Section 4.3. We combine the five sigmoid maps by arithmetic mean before post-processing; the environment (Python 3.11.13, PyTorch 2.13.0+cu130, MONAI 1.6.0) is pinned in the container specification.

2.5 Post-processing

Inference uses Auto3DSeg's `SlidingWindowInfererAdapt` at the training crop size, adjacent windows overlapping by 62.5% with Gaussian blending. The frozen chain runs in this order, and the configuration is version-controlled and verified by whole-file MD5 against the Synapse-hosted artifact.
(1) Region thresholding on the five-fold mean sigmoid: WT 0.5, TC 0.45, ET 0.25, ET lower because missing a small metastasis costs more than over-segmenting one already found. **(2)** RC hysteresis: a cavity component exists only if some

voxel exceeds a 0.97 seed, after which a 0.5 flood sets its extent. **(3)** Site-proxy gate (Section 2.6). **(4)** WT-without-ET filter: drop WT components with fewer than 20 ET voxels unless they overlap a predicted cavity. **(5)** RC rim test: drop cavity components enclosed by a filled ET envelope, read as residual necrotic core. **(6)** A 15 mm^3 volume floor on every region. **(7)** Containment, `tc |= et` and `wt |= tc`. **(8)** Conflict resolution: a cavity keeps a contested voxel only if its raw probability is higher. **(9)** The 15 mm^3 floor again, since conflict resolution can shrink a component below a floor it already cleared. **(10)** Brain masking to nonzero post-contrast T1.

On the 179 validation cases the second volume pass removed 44–63 residual fragments per tumor region plus one 0.26 mm^3 cavity fragment without changing the detection status of any case, for the reason given in Section 4.1.

The evaluation container reuses this inference script unmodified, with all dependencies pinned and no network access, and executed to completion on the hidden test set. Configuration and code are therefore exactly reproducible; numerical output is not, since cuDNN benchmark-mode selection under mixed precision perturbs raw probabilities between runs on identical hardware. Across three complete runs, voxel-level output differed in roughly half the cases while the instance-level TP, FP, FN and F1 returned by the official scorer were identical in every case, region and bucket. We therefore report run-to-run agreement in the official instance metrics, rather than byte identity, as the reproducibility criterion appropriate to a mixed-precision pipeline. The chain contains six hand-set parameters, which raises a fair concern about challenge-specific overfitting; Sections 3.1 and 3.3 bound it—the tumor regions move 0.0170 in total across all threshold tuning, and the cavity parameters buy precision without changing the detection pattern at any value we swept.

2.6 A site proxy for treatment status

A resection cavity presupposes prior surgery, so treatment status constrains this channel in a way that has no analogue for the tumor regions, and that status is not supplied with the challenge data. Because cavity-bearing cases here are drawn almost entirely from one contributing institution (Section 2.1), we approximate it with a dataset-informed prior: the frozen configuration suppresses the cavity channel for cases whose institutional origin, inferred from the patient component of the case identifier, makes prior surgery unlikely. The measured effect is one-sided. Removing the gate reduces cohort cavity instance F1 by 0.0853 on held-out data, and none of the 151 cavity-bearing training cases is suppressed by it, so within this corpus its entire contribution is false-positive reduction at no measured cost in recall (Section 3.3).

The prior is corpus-derived rather than image-derived, and its identifier-matching implementation cannot be validated against test-phase identifiers in advance; incorrect transfer would be silent in either direction. The container therefore verifies the flagged-case count against the 213 of 303 test cases documented as originating from that institution, and outside a pre-registered band of

170–250 substitutes an image-only classifier (0.8689 AUC, no identifier dependency) selecting the highest-ranked 213 cases. We report it as a prior fitted to this dataset's composition rather than a generally transferable component.

3 Results

3.1 Leaderboard progression

Table 1 lists our four submissions to the unranked validation leaderboard. The composite column is the unweighted mean of all-instance F1 across the four regions; it is used for readability and is not an official ranking statistic.

Table 1. Validation-leaderboard progression. Between the final two submissions, ET, TC and WT are identical to four decimal places and RC moves alone.

Submission	Configuration	ET F1	TC F1	WT F1	RC F1	Composite
9773954	4-fold, flat 0.5 baseline	0.6913	0.7072	0.6499	0.1235	0.5430
9774077	4-fold, tuned, RC seed 0.99	0.7075	0.7152	0.6933	0.1667	0.5707
9774162	5-fold, max-seed 0.99, 250 mm^3 floor	0.7083	0.7160	0.6919	0.2593	0.5939
9774164	**5-fold, mean-seed 0.97 (shipped)**	**0.7083**	**0.7160**	**0.6919**	**0.2756**	**0.5980**

What changed at each step. Submission 9773954 is the four-fold ensemble at a flat 0.5 threshold with no filtering. Submission 9774077 adds the tuned chain of Section 2.5 at cavity seed 0.99, where all three tumor regions make essentially their entire gain. Submission 9774162 adds fold 4, raises the cavity floor to 250 mm^3 and seeds on the *maximum* across models; the shipped 9774164 reverts to a mean-based seed lowered to 0.97 (Section 3.3) and restores the 15 mm^3 floor. The composite rose 0.0550 across the four, each region contributing one quarter of its own change: cavity +0.0380 (69%, on ΔF1 +0.1521), whole tumor +0.0105 (19%), enhancing tumor +0.0043 (8%) and tumor core +0.0022 (4%). One confound deserves naming: the first two submissions use four folds and the last two five, so the step from 9774077 to 9774162 changes both the cavity configuration and the ensemble size—across that transition ET, TC and WT shift by at most 0.0014, an order of magnitude below the 0.0170 total tumor-region movement attributed to threshold tuning across the whole progression (Section 2.5), so we do not read the fifth fold as materially responsible for the accompanying 0.0926 cavity gain (0.1667 to 0.2593). By contrast, the tumor regions are unchanged to *four decimal places* across the final transition (9774162 to 9774164), where the only configuration change was the cavity seed; since nothing else moved, the 0.0163 cavity gain there (0.2593 to 0.2756) is attributable to the seed by construction.

3.2 Shipped-submission metrics

Table 2 gives the official metric set for the shipped submission. Small-instance F1 runs between 0.14 and 0.21 across the three tumor regions, well below the all-instance figures.

Table 2. Per-region metrics, shipped submission. Large-instance F1 is 0.7143, 0.7161, 0.6870 and 0.0391 for ET, TC, WT and RC. DSC, NSD and HD95 are lesion-wise, mean ± SD over matched pairs. **Rates in the two buckets are averaged over different case sets and are not additive**: all-instance rates use all 179 cases, small-instance rates only the cases containing at least one small ground-truth instance in that region (75, 74, 76 and 6 for ET, TC, WT and RC).

Region	Bucket	TP/case	FP/case	FN/case	F1	DSC	NSD	HD95
ET	all	3.4078	0.3631	3.4581	0.7083	0.63 ± 0.18	0.67 ± 0.19	91 ± 75
ET	small	0.8400	0.4533	5.8800	0.2061	—	—	—
TC	all	3.4190	0.3520	3.3966	0.7160	0.64 ± 0.18	0.67 ± 0.19	90 ± 75
TC	small	0.8514	0.4459	5.8378	0.2089	—	—	—
WT	all	3.1117	0.3799	3.4190	0.6919	0.51 ± 0.17	0.48 ± 0.18	107 ± 79
WT	small	0.5789	0.4868	5.4474	0.1435	—	—	—
RC*	all	0.0447	0.0168	0.1285	0.2756	0.23 ± 0.00	0.15 ± 0.00	239 ± 0

*The cavity dispersion statistics return exactly 0.0 while the means stay reproducible (DSC recomputes to 0.2319 over 22 non-null cases, naive SD 0.346); submission 9774162 returns 0.0135, 8.379 and 0.00624 for the same three. A property of the scoring pipeline, not of any algorithm. The RC small bucket is all-zero (1.1667 FN/case) and omitted.

Two entries need comment. RC large-instance F1 (0.0391) falls well below its all-instance value (0.2756) although the buckets share the same three false positives and differ by a single true positive; the difference follows from the denominator, since the large bucket admits more cavity-free cases carrying spurious large components and a handful of them dominates a mean over 23 cavity-bearing cases. Lesion-wise HD95 is computed over matched pairs on a corpus of predominantly small lesions, where one poorly localized instance contributes a distance bounded only by head size; the 90–107 mm tumor values and 239 mm cavity value reflect that aggregation rather than boundary accuracy on well-detected lesions.

3.3 The resection cavity, decomposed

Submission 9774164 has 8 true positives, 3 false positives and 23 false negatives, so 31 ground-truth cavity instances; 23 of the 179 cases carry at least one cavity, against a pre-scoring estimate of 21.3 ± 4.0 from the validation set's UCSD case count and within-institution training prevalence. Of those 23 cavity-bearing cases, 6 are fully detected, 2 partially detected and 15 missed entirely, and this pattern is *identical* across submissions 9774162 and 9774164 even though their leaderboard cavity F1 differs by 0.016 (0.2593 against 0.2756). The difference is attributable entirely to false positives: two fewer cavity-bearing cases carry a spurious additional component (2 against 0) and one fewer cavity-free case receives any spurious detection (4 against 3). Under the empty-case exclusion this moves the included-case count from 27 to 26 and the summed per-case F1 from 7.00 to 7.17. No additional cavity was detected.

The 15 missed cases persisted under every configuration tested—seven seed values, two seed parameterizations and five volume floors—none of which changed the count by a single case. No voxel in these cases exceeds the seed threshold at

any setting we evaluated, which locates the limitation in the model's probability output rather than in the decision rule applied to it.

Gating. On the out-of-fold set (n=1294)*, cohort RC instance F1 is 0.4841 with the shipped gate, 0.3988 without one, and 0.600 under an oracle gate using ground-truth cavity status, which bounds what any case-level gate can achieve. The dataset-informed prior thus recovers roughly 40% of the available headroom (0.0853 of the 0.2012 gap between the ungated and oracle figures) at no measured cost in recall. These are out-of-fold values on a cohort whose composition differs from the validation set, so the relation between the three figures rather than their magnitudes is the comparison of interest.

Why the seed had to move. We originally tuned the cavity seed to 0.99 on per-fold out-of-fold probabilities, but every submission runs inference on the five-model *mean*—a different distribution with different peak statistics. Lowering the seed to 0.97 roughly doubled detected cavity instances, from 4 to 8. We measured why on cached five-model probabilities. The peak of the models' mean divided by the mean of their individual peaks cannot exceed 1, and equals 1 exactly when all five peak at the same voxel; its median is 0.7145 on unseen validation cases against 1.0000 on memorized cross-fold data, and the median pairwise distance between the models' cavity argmax voxels is 103 mm—about the width of a brain. On unseen cases the models disagree about *where* the cavity is, not merely how sure they are, and averaging spatially disagreeing predictors flattens the peak of the mean. An absolute threshold calibrated on a per-fold distribution therefore does not transfer to the ensembled mean it runs against.

4 Discussion

4.1 Findings that generalize beyond this challenge

The four observations below concern the metric, the matcher, the loss function and the evaluation substrate rather than our particular model, and each constrains how results obtained under the same evaluation should be interpreted.

An empty-case convention creates a step penalty no threshold sweep can see. A case with empty ground truth and a non-empty prediction enters the cohort mean with F1= 0, while the same case predicting nothing is dropped entirely, so the first spurious prediction in a lesion-free case costs about $M/(N+1)$—roughly six times a typical marginal candidate's contribution at $M\approx0.69$, N=179. Being a step rather than a gradient it is invisible to any continuous threshold sweep, and it dominates in sparse regions like the cavity

* One case (`BraTS-MET-00695-000`) is excluded from every out-of-fold aggregation in this work: a `Spacingd` rounding artifact leaves its predicted probability array one voxel short of its ground-truth grid in two dimensions, a reproducible geometry mismatch rather than a stochastic failure. This accounts for 1,295 of the 1,296 training cases. The identity of a second excluded case, needed to reach the n=1294 reported here, is not recoverable from our records; we report the figure as logged rather than silently adjusting it to 1,295.

where most cases are empty. **Redundant detections carry no cost under any-overlap dilation matching.** A ground-truth lesion is credited as found if *any* predicted component's dilation touches it, so two components on one lesion yield one true positive and no additional false positive. Component splitting therefore cannot recover credit a merged prediction has already earned, and the volume-floor ordering fix removed dozens of fragments per region without changing any region's F1—we infer, from the unchanged score, that those fragments overlapped already-matched lesions. **Tversky's β has a regime-dependent sign.** The usual gradient-dilution argument for recall-favoring $\beta{>}\alpha$ [18] runs: with Tversky index = TP/(TP+αFP+βFN), near TP$\approx$ 0 the derivative $\partial T/\partial p \approx 1/(\beta \cdot \mathrm{FN})$, so raising β weakens the correction where it is most needed. That holds only for a missed lesion sitting *alone* in its patch. Add a well-detected large lesion to the same patch—the ordinary multi-lesion case—and the denominator becomes TP-dominated, the derivative becomes $\approx \beta/\mathrm{TP}$, and raising β now *strengthens* the correction. Numerically, moving (α, β) from $(0.5, 0.5)$ to $(0.1, 0.9)$ lowers the gradient at an isolated missed lesion (9.98 to 5.55×10^{-2}) while raising it when a large detected lesion shares the patch (2.75 to 4.55×10^{-5}): the same parameter change, opposite in sign. The result is analytic and verified numerically rather than established by a controlled training comparison, and it indicates that the direction of a recall-reweighting adjustment depends on patch composition. **Cross-fold ensembling is contaminated on longitudinal cohorts.** A resection cavity persists across a patient's own timepoints, so any substrate letting one fold's model see another timepoint of the same patient will flatter detection of patient-specific structures: cavity cohort F1 on such a substrate reached 0.7333 against 0.2756 on the real leaderboard. The configurations are not matched, so the comparison indicates magnitude rather than an exact contrast, but the separation is large enough that we do not treat this substrate as informative in absolute terms. Diagnostics run on it inherit the problem: our first test of the per-fold to ensembled-mean shift used cross-fold predictions on seen patients, where max(mean) = mean(max) holds by construction, so it could only have returned agreement.

4.2 A measured null space

Several plausible interventions did not improve performance under the configurations tested. We distinguish those measured with adequate sensitivity from those whose measurement design could not have detected the effect they targeted; the second group is underpowered rather than negative, and we do not read it as evidence against the underlying approach.

Measured without improvement. A joint threshold and volume sweep located an interior optimum, and five subsequent refinements produced no further gain. Watershed splitting of merged whole-tumor components recovered no additional true positives across the out-of-fold set, consistent with the matcher property above, since no unmatched ground-truth lesion remained for splitting to recover. Fifty-seven fusion configurations improved no region's small-lesion recall without a disproportionate increase in false positives. A max-across-models

cavity seed, intended to exploit cross-model disagreement, detected no additional cavities and doubled false positives, consistent with a maximum-based gate responding to a single model's isolated peak rather than to agreement among models. Cavity volume floors from 15 to 2000 mm^3 did not separate true from false positives on the shipped configuration, whose remaining cavity false positives are themselves cavity-sized. **Underpowered rather than negative.** Four further arms—test-time augmentation, connected-lesion-aware sampling, a lesion-equal-weight loss terminated early on an undiagnosed flat loss, and a Tversky arm in which MONAI's `TverskyLoss` lacks `squared_pred` and so altered the false-positive penalty alongside the intended reweighting—each rested on a single seed without a matched control. These should be repeated with adequate controls before being interpreted as evidence about the underlying methods. Bootstrap standard errors over the out-of-fold set (0.02 for ET/TC/WT, 0.097 for the cavity) characterize how well a 179-case sample represents the population, and do not apply to paired comparisons on identical cases.

4.3 Limitations

The principal limitation is the 15 cavity-bearing cases that were missed under every decision rule tested. Since no voxel in these cases exceeded the seed threshold, addressing this failure likely requires an architectural change, such as a dedicated detection or centroid-heatmap head [23]. We did not characterize these cases by lesion size, interval since surgery, contributing site, or proximity to the ventricular system, which would be a natural direction for future work.

Small-instance F1 also remains low despite evidence that sub-threshold probability information is informative: max-probability AUC against size- and depth-matched sham locations is 0.7211 within the sub-0.1 band, above the 0.59 ceiling expected from pure noise. This suggests a calibration rather than purely perceptual limitation, motivating future work on calibrated component-level scores rather than raw probability thresholds.

Finally, random seeds were not fixed and deterministic kernels were not enforced, so individual training-side experiments lack variance estimates. The cavity gate also carries a generalization risk because its benefit was measured on a corpus with a highly concentrated institutional distribution. Moreover, the validation configuration was tuned and evaluated on the same 179-case unranked leaderboard.

5 Conclusion

We presented a five-fold SegResNetDS ensemble trained from scratch for BraTS 2026 Task 1 with a frozen detection-oriented post-processing pipeline, reaching lesion-wise instance F1 of 0.71 for enhancing tumor, 0.72 for tumor core, 0.69 for whole tumor and 0.28 for resection cavity. Case-level analysis establishes that the cavity improvement was precision-driven: our final two configurations detect the same 8 instances across the same 23 cavity-bearing cases and differ only

in spurious components. Alongside the system we characterized four properties of the lesion-wise evaluation setting (Section 4.1), each relevant to interpreting results obtained under the same evaluation. Cavity detection remains the principal unresolved problem: 15 of 23 cavity-bearing cases are missed entirely and no decision-rule change we tested altered that count. Characterizing those cases, rather than searching further over thresholds, is the direction we would pursue next.

Acknowledgements

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. Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097). Per the challenge's data usage agreement, we additionally acknowledge the MedPerf federated benchmarking infrastructure [8] associated with this dataset.

Disclosure of Interests

The authors declare that they have no competing interests.

References

1. 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. Lecture Notes in Computer Science, vol. 16376, pp. 161–172. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16365-3_15, version 3
2. BraTS-METS Organizing Committee: BraTS 2026 challenge, task 1: Brain metastasis segmentation. Sage Bionetworks Synapse (2026), `https://www.synapse.org/Synapse:syn74274097/wiki/639582`, accessed July 31, 2026
3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. Grøvik, E., Yi, D., Iv, M., Tong, E., Rubin, D.L., 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). https://doi.org/10.1002/jmri.26766

5. 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: International MICCAI Brainlesion Workshop. pp. 272–284 (2021), arXiv:2201.01266
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
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**(2), 203–211 (2021)
8. 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
9. 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
10. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial Intelligence and Statistics (AISTATS). pp. 562–570 (2015)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
12. Maleki, N., Amiruddin, R., Moawad, A.W., Yordanov, N., Gkampenis, A., Fehringer, P., Umeh, F., 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 preprint arXiv:2504.12527 (2025). https://doi.org/10.48550/arXiv.2504.12527, `https://arxiv.org/abs/2504.12527`
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV). pp. 565–571 (2016)
14. Moawad, A.W., Janas, A., Baid, U., Ramakrishnan, D., Saluja, R., Ashraf, N., Maleki, N., et al.: The brain tumor segmentation–metastases (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, `https://arxiv.org/abs/2306.00838`
15. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. pp. 311–320 (2018), arXiv:1810.11654
16. Myronenko, A., Yang, D., Buch, V., Xu, D., et al.: Auto3dseg for brain tumor segmentation from 3d mri in brats 2023 challenge. arXiv preprint arXiv:2510.25058 (2025)
17. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jäger, P.F., Maier-Hein, K.H.: MedNeXt: Transformer-driven scaling of convnets for medical image segmentation. In: MICCAI (2023), arXiv:2303.09975
18. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging. pp. 379–387 (2017), arXiv:1706.05721
19. Saluja, R., et al.: BraTS-2024-Metrics: Lesion-wise scoring for the brats challenges. `https://github.com/rachitsaluja/BraTS-2024-Metrics` (2024), accessed 2026-07-28
20. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)

21. Vogelbaum, M.A., Brown, P.D., Messersmith, H., Brastianos, P.K., Burri, S., Cahill, D., Dunn, I.F., Gaspar, L.E., Gatson, N.T.N., Gondi, V., Jordan, J.T., Lassman, A.B., Maues, J., Mohile, N., Redjal, N., Stevens, G., Sulman, E., van den Bent, M., Wallace, H.J., Weinberg, J.S., Zadeh, G., Schiff, D.: Treatment for brain metastases: ASCO-SNO-ASTRO guideline. Journal of Clinical Oncology **40**(5), 492–516 (2022). https://doi.org/10.1200/JCO.21.02314
22. Wang, T.W., Hsu, M.S., Lee, W.K., Pan, H.C., Yang, H.C., Lee, C.C., Wu, Y.T.: Brain metastasis tumor segmentation and detection using deep learning algorithms: A systematic review and meta-analysis. Radiotherapy and Oncology **190**, 110007 (2024). https://doi.org/10.1016/j.radonc.2023.110007
23. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)