This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# A Multi-Network Ensemble Strategy for the BraTS 2026 Brain Metastasis Segmentation Challenge

Nicolo Pecco[1,2][0000−0002−6367−7903] and Lorenzo Veronese[3][0000−0001−9204−3363]

[1] Vita-Salute San Raffaele University, Milan, 20132, Italy
pecco.nicolo@hsr.it
[2] Department of Neuroradiology and CERMAC IRCCS San Raffaele Hospital, Milan, 20132, Italy
[3] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, 20133, Italy
lorenzo.veronese@polimi.it

**Abstract.** Brain metastasis segmentation on magnetic resonance imaging (MRI) poses significant challenges, as lesions are often small, spatially heterogeneous, and exhibit substantial variability between pre- and post-treatment scans. Post-surgical cases introduce additional difficulty due to the presence of resection cavities, which are rare and visually variable. In this work, we presented a multi-network ensemble strategy for Task 1 of the BraTS-METS 2026 challenge. The proposed approach combined three complementary 3D segmentation models, MedNeXt, nnU-Net, and STU-Net, trained on multi-sequence MRI scans including T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and FLAIR images. Model predictions were fused using class-specific weighted probability averaging for standard tumor-related labels and the resection-cavity class, across three stages. A first ensemble stage optimized the segmentation of non-enhancing tumor core, surrounding FLAIR hyperintensity, and enhancing tumor, while a second stage applied class-specific calibration for resection cavities. A final connected-component post-processing step was employed to reduce small false-positive predictions. On the validation leaderboard, the final ensemble improved the mean Dice score from 0.609 to 0.627 compared with the non-post-processed ensemble, with gains across enhancing tumor, tumor core, whole tumor, and resection cavity regions. These results suggest that class-specific ensemble fusion and conservative post-processing provide a practical strategy for robust brain metastasis segmentation in heterogeneous clinical MRI datasets. Our code is available at https://github.com/NicoloPecco/Segmentation_METs_BraTS2026.

**Keywords:** Brain metastasis segmentation · BraTS-METS · Deep learning · Ensembling · MRI

## 1 Introduction

Brain metastases are the most common intracranial tumors in adults and usually originate from systemic cancers of the lung, breast, and skin [14, 22]. A substantial

fraction of patients with solid tumors develop intracranial disease, and as improved systemic therapies extend survival, the incidence of brain metastases and the demand for imaging-based surveillance both continue to rise [14]. Magnetic resonance imaging (MRI) is the standard for detecting and monitoring these lesions, owing to its high soft-tissue contrast and sensitivity to small enhancing tumors and surrounding edema.

Clinically, treatment response is assessed with the Response Assessment in Neuro-Oncology Brain Metastases (RANO-BM) criteria, which use a single (unidimensional) diameter measurement of target lesions [15]. However, a single linear measurement provides only a coarse approximation of tumor burden and may fail to capture the full spatial extent of the disease. In contrast, volumetric quantification of both enhancing lesions and perilesional edema provides a more comprehensive characterization of tumor burden and is increasingly important for stereotactic radiosurgery planning and outcome prediction. Producing such volumes requires delineating every lesion, a task currently performed manually by expert readers. This laborious and time-consuming process suffers from considerable inter- and intra-rater variability [4]. The difficulty is further exacerbated by lesion size: brain metastases are frequently small, and detecting and segmenting sub-centimetre lesions remains particularly challenging, historically yielding low Dice similarity coefficients. Post-treatment imaging adds further complexity, as resection cavities and treatment-related changes can resemble true tumors.

To address these limitations, the field has increasingly adopted deep learning approaches [5], which now enables automatic detection and multi-class segmentation of metastases across multi-sequence MRI, specifically pre-contrast T1-weighted (T1W), contrast-enhanced T1-weighted (T1C), T2-weighted (T2W), and T2 fluid-attenuated inversion recovery (FLAIR) images [6, 27, 3, 2]. The Brain Tumor Segmentation (BraTS) challenge series [18, 1, 26] has played a central role in this effort by providing standardized datasets and evaluation frameworks, and the BraTS-METS challenge extended this framework to brain metastases [19]. Building on this foundation, the BraTS-Lighthouse 2025 metastases task (Task 4) focused on segmentation across the full treatment timeline using a large, multi-institutional dataset acquired with diverse scanners and clinical protocols [17].

The U-Net [23] and its self-configuring successor nnU-Net [11] remain the strongest baselines, alongside encoder-decoder variants such as SegResNet [20], attention-guided designs like Attention U-Net [21], architecture-search models such as DiNTS [9], and transformer-based networks including UNETR [8] and Swin UNETR [7, 25]. These architectures encode different inductive biases: convolutional models are well-suited to capturing local detail, whereas transformer-based encoders model long-range contextual dependencies. Consequently, no single network is optimal across all subregions, especially given the severe class imbalance from small lesions and the distinct resection-cavity class [16]. For the BraTS 2026 challenge (Task 1), we therefore developed a class-specific ensemble for brain-metastasis segmentation, evaluated within the challenge's MedPerf-based framework [13]. The ensemble combines models with

complementary inductive biases and applies class-specific weights and calibration to the hardest subregions.

Our main contributions can be summarized as follows:

- addressing brain-metastasis segmentation across the full treatment timeline, covering both pre- and post-treatment scans and the resection-cavity subregion;
- designing an ensemble of STU-Net [10], MedNeXt [24], and nnU-Net [11], leveraging their complementary inductive biases for robust multi-class lesion segmentation;
- introducing a two-stage, class-specific weighting scheme for ensemble fusion, allowing each model to contribute differently to individual tumor subregions and improving segmentation performance.

## 2 Methods

### 2.1 Dataset

We conducted our experiments on the dataset released within the BraTS-Lighthouse brain metastases segmentation challenge (BraTS 2026, Task 1) [17], including pre- and post-treatment multi-centric MRI studies (single-timepoint cases and longitudinal dataset) acquired under standard clinical conditions using a variety of scanners and acquisition protocols. The data were organized into three subsets: i) a training set of 1296 patients with expert ground-truth (GT) masks (the only subset with public GT, used for all model development); ii) an official validation leaderboard set of 179 scans for which the GT was not provided and whose predictions were scored by the online evaluation server; and iii) a held-out, unseen test set used for the final performance assessment and ranking. For each case, four co-registered MRI sequences were furnished: native pre-contrast T1-weighted (T1W), contrast-enhanced T1-weighted (T1C), T2-weighted (T2W), and T2 fluid-attenuated inversion recovery (FLAIR). All volumes were preprocessed by the challenge organizers, including conversion to NIfTI format, co-registration, resampling to $1\text{mm}^3$ isotropic resolution, and skull stripping. Each voxel was assigned to one of five classes: background, non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH, i.e., the peritumoral edema), enhancing tumor (ET), and resection cavity (RC), the latter present only in post-treatment (post-surgical) cases.

### 2.2 Data Partitioning

Because the GT was available only for the 1296 labeled training scans, all model training and selection were performed on this set. The training set was partitioned into an internal training set of 1011 cases and an internal validation set of 285 cases, corresponding to an approximate 80%/20% split stratified by the four foreground labels (NETC, SNFH, ET, and RC) to preserve their relative frequencies in both subsets (see Table 1 for training/validation set class distributions). For the UCSD dataset, which includes a longitudinal subset of patients, the partition was performed at the patient level to avoid data leakage between subsets.

**Table 1.** Foreground classes and their approximate prevalence across the labeled training and validation cases.

| Value | Class | Prevalence |
|---|---|---|
| 1 | NETC – non-enhancing tumor core | ∼60% |
| 2 | SNFH – surrounding FLAIR hyperintensity (edema) | ∼85% |
| 3 | ET – enhancing tumor | ∼95% |
| 4 | RC – resection cavity | 12.9% |

### 2.3 Evaluation Metrics

Segmentation quality was quantified using the Dice Similarity Coefficient (DSC) and the Normalized Surface Distance (NSD). During model development, DSC was computed for the four native segmentation classes: non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), enhancing tumor (ET), and resection cavity (RC). For the official validation leaderboard and unseen test set, performance was reported using both DSC and NSD for ET, RC, tumor core (TC = ET + NETC), and whole tumor (WT = TC + SNFH). DSC measured volumetric overlap, whereas NSD assessed boundary agreement by quantifying the proportion of surface points lying within an acceptable distance of the corresponding reference surface. Both metrics were computed lesion-wise, thereby penalizing missed lesions (false negatives) and spurious detections (false positives).

### 2.4 Experimental Setup

We combined three complementary families of 3D segmentation models.

1. STU-Net [10], a fully convolutional encoder-decoder architecture analogous to U-Net, but scaling to very large capacities through depth-wise separable convolutions with large kernels. We used the *Large* variant (≈440 M parameters), trained within the nnU-Net-v2 framework [11] via the `STUNetTrainerLarge` trainer in the `3d_fullres` configuration with a patch size of $160 \times 224 \times 192$, a batch size of 1, and a target spacing of $1.0 \times 0.898 \times 0.859$ mm.
2. MedNeXt [24], a fully-convolutional architecture inspired by ConvNeXt that replaces standard residual blocks with large-kernel depthwise-separable ConvNeXt blocks, enabling long-range context without self-attention. We used the *Large* variant with kernel size 5, trained within the nnU-Net-v1 framework in the `3d_fullres` configuration with a patch size of $112 \times 160 \times 128$, a batch size of 2, and an isotropic target spacing of $1.0 \times 1.0 \times 1.0$ mm.
3. NnU-Net [11], a self-configuring fully-convolutional framework that automatically adapts preprocessing, network topology, and training to the dataset. We trained it in the `3d_fullres` configuration with a patch size of $112 \times 160 \times 128$, a batch size of 2, and a target spacing of $1.0 \times 0.898 \times 0.859$ mm.

All models were trained for 1000 epochs on a single train-validation split on an NVIDIA A10 GPU (24GB VRAM, with Python 3.11) and best checkpoint

was used for final evaluation. Standard data augmentation was used for the training set. Probability maps (softmax outputs) were saved for each model and subsequently combined via weighted probability averaging (cfr. Section 2.5). Our team identifier on the Synapse platform was `TheRealBoys`. The full pipeline including evaluation scripts and model weights are open sourced at: https://github.com/NicoloPecco/Segmentation_METs_BraTS2026.

### 2.5 Ensembling and weighted probability

Given the softmax outputs $p_m(c, \mathbf{x})$ of model $m \in \{\text{MedNeXt, nnU-Net, STU-Net}\}$ for class $c$ at voxel $\mathbf{x}$, the final segmentation was $\hat{y}(\mathbf{x}) = \arg\max_c \hat{p}(c, \mathbf{x})$. The weights were allowed to differ between the RC channel and the non-RC channels (NETC, SNFH, ET). Weight selection was performed in two stages on the internal validation fold: first for the non-RC labels (Stage 1) and then separately for RC (Stage 2). A third stage was subsequently applied as post-processing, in which small connected components were removed from the final segmentation (Stage 3).

*Stage 1: Ensemble weights for NETC, SNFH, and ET* After model training, we optimized the ensemble weights for the three labels NETC, SNFH, and ET. Specifically, each weight combination was selected from the set $0, 0.25, 0.5, 0.75, 1.0$, resulting in 124 possible configurations. For each configuration, the same set of weights was applied to all probability channels. The optimal configuration was selected as the one achieving the highest mean DSC across NETC, SNFH, and ET on the internal validation set and used for Stage 2.

*Stage 2: RC-specific weighting and calibration* In the second stage, we kept the optimal ensemble weights identified in Stage 1 fixed for the non-RC labels and performed a separate optimization for the RC channel. A separate RC probability map was computed using an independently optimized set of model weights and voxels where $\hat{p}(\text{RC}, \cdot) > \tau$ were then assigned the RC label, overwriting the Stage 1 prediction. We evaluated the same 124 model-weight combinations of Stage 1 for the RC channel and additionally swept the threshold $\tau \in 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$, yielding 744 configurations in total.

*Stage 3: Post-processing and thresholding.* Because the evaluation was performed on a lesion-wise basis, model performance was particularly sensitive to false positives (FPs) and false negatives (FNs). While FNs are difficult to recover through post-processing, FPs can be partially reduced by removing small isolated predictions that are unlikely to represent true lesions. Therefore, we applied a connected-component size filter, retaining only predicted lesions with at least 50 voxels. However, the choice of threshold requires a careful trade-off. Post-treatment cases may contain small residual or recurrent lesions, whereas pre-treatment cases often include larger and more clearly visible tumor components. Unlike previous challenge pipelines, where relatively large class-specific thresholds were applied to pre-treatment cases (e.g., 250 voxels for WT, 150 voxels for TC, and 100 voxels

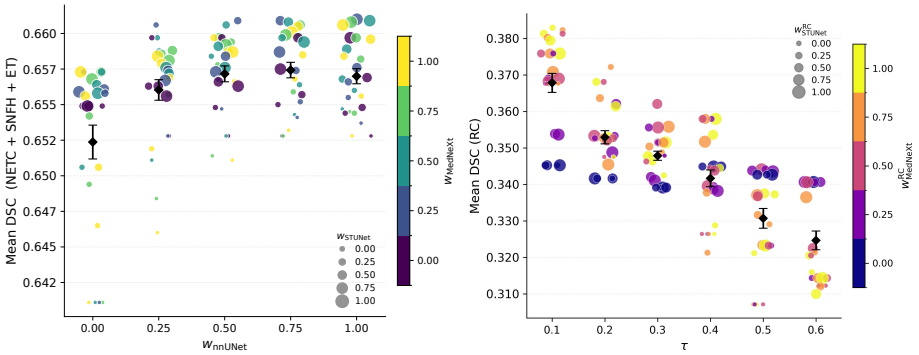


**Fig. 1.** Strip plot of Stage 1 ensemble weight combinations (left) and Stage 2 RC threshold search (right), both evaluated on the internal validation set. Left: each point represents one combination; the $y$-axis reports the mean DSC averaged over NETC, SNFH, and ET. Points are grouped along the $x$-axis by weights of nnUNet, coloured by MedNeXt, and scaled in size proportionally to STUNet. Right: nnUNet RC weight is fixed to 0 (optimal); each point represents one combination; the $y$-axis reports mean DSC on the RC class. Points are grouped by threshold $\tau$, coloured by MedNeXt, and scaled by STUNet. In both panels, black diamonds indicate the group mean $\pm$ one standard error.

for ET), more conservative thresholds have been adopted for post-treatment lesions, including a 50-voxel threshold for ET [12]. Based on this rationale, we selected a minimum connected-component size of 50 voxels to remove very small isolated predictions likely to represent segmentation artifacts, while minimizing the risk of discarding small residual or recurrent lesions.

## 3 Results

### 3.1 Stage 1: Ensemble weights for NETC, SNFH, and ET

The ensemble weight search identified the configuration MedNeXt/nnU-Net/STU-Net = 1.00/0.5/0.25 as the best-performing setting when optimizing the mean Dice score across NETC, SNFH, and ET. The weight-search landscape showed only minimal differences among the top-ranked configurations, indicating a broad performance plateau rather than a single sharply optimal solution. Overall, the trend suggested that nnU-Net and STU-Net were the main contributors to the ensemble performance, while MedNeXt provided a smaller incremental contribution (Fig. 1, left). This configuration achieved Dice scores of 0.535 for NETC, 0.738 for SNFH, and 0.728 for ET, corresponding to a mean non-RC Dice of 0.667 on the internal validation set (Table 2) while 0.702, 0.666, 0.680, and 0.386 on TC, WT, ET, and RC, respectively, in the validation leaderboard set.

**Table 2.** Single-model, Stage 1 and Stage 2 ensemble performance on the internal validation set. Values are reported as mean DSC ± standard deviation. The best results for each metric are in bold.

| Model | NETC | SNFH | ET | RC | Mean |
|---|---|---|---|---|---|
| MedNeXt | $0.523 \pm 0.381$ | $0.720 \pm 0.283$ | $0.717 \pm 0.243$ | $0.380 \pm 0.395$ | 0.585 |
| nnU-Net | $0.519 \pm 0.385$ | $0.725 \pm 0.280$ | $0.714 \pm 0.253$ | $0.115 \pm 0.264$ | 0.518 |
| STU-Net | $0.531 \pm 0.378$ | $0.717 \pm 0.281$ | $0.717 \pm 0.250$ | $0.343 \pm 0.356$ | 0.577 |
| Stage 1 | $\mathbf{0.535 \pm 0.383}$ | $\mathbf{0.738 \pm 0.272}$ | $\mathbf{0.728 \pm 0.244}$ | $0.294 \pm 0.378$ | 0.574 |
| Stage 2 | $\mathbf{0.535 \pm 0.383}$ | $0.735 \pm 0.275$ | $0.720 \pm 0.254$ | $\mathbf{0.416 \pm 0.395}$ | **0.602** |

**Table 3.** Stage 2 and Stage 3 ensemble performance for `TheRealBoys` team on the validation leaderboard (Val. LB). Values are reported as DSC and NSD. The best results for each metric are in bold.

| Submission ID | Set | Model | Metric | ET | RC | TC | WT | Mean |
|---|---|---|---|---|---|---|---|---|
| 9768076 | Val. LB | Stage 2 | DSC | 0.680 | 0.386 | 0.702 | 0.666 | 0.609 |
| 9768076 | Val. LB | Stage 2 | NSD | 0.741 | 0.266 | **0.751** | **0.688** | 0.612 |
| 9768120 | Val. LB | Stage 3 | DSC | **0.697** | **0.419** | **0.713** | **0.679** | **0.627** |
| 9768120 | Val. LB | Stage 3 | NSD | **0.748** | **0.277** | 0.748 | 0.684 | **0.614** |

### 3.2 Stage 2: RC-specific weighting and calibration

In the second stage, the RC label was assigned independently from the argmax-based prediction of the other classes. Specifically, a normalized RC probability map was obtained from the weighted combination of MedNeXt and STU-Net probabilities, using MedNeXt/nnU-Net/STU-Net = 1.00/0.00/0.25 followed by thresholding with $\tau = 0.1$. This design allowed RC to be detected without directly competing with NETC, SNFH, and ET within the same argmax operation. This configuration achieved DSC scores of 0.535 for NETC, 0.735 for SNFH, 0.720 for ET, and 0.416 for RC on the internal validation set (Table 2), while achieving DSC scores of 0.680, 0.386, 0.702, and 0.666 for ET, RC, TC, and WT, respectively, on the validation leaderboard set (Table 3).

### 3.3 Stage 3: Post-processing and thresholding

The connected-component post-processing step improved the validation leaderboard performance compared to the Stage 2 ensemble (Table 3, Figure 2). Mean DSC increased from 0.609 to 0.627 after applying Stage 3. Improvements were observed for all reported regions: ET increased from 0.680 to 0.697, RC from 0.386 to 0.419, TC from 0.702 to 0.713, and WT from 0.666 to 0.679. The effect on NSD was smaller. Mean NSD increased slightly from 0.612 to 0.614, with improvements for ET and RC, which increased from 0.741 to 0.748 and from 0.266 to 0.277, respectively. Conversely, TC and WT NSD showed minor decreases, from 0.751 to 0.748 and from 0.688 to 0.684. Overall, Stage 3 mainly improved lesion-wise volumetric overlap, particularly for RC, while producing only marginal

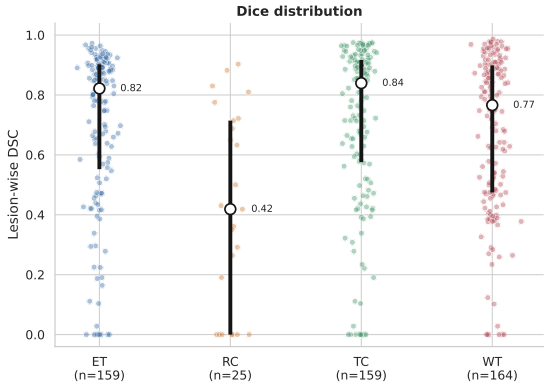


**Fig. 2.** Validation leaderboard dice score distribution for each case (Stage 3). White circles indicate the median Dice score, while vertical black bars denote the interquartile range. The number of valid cases is reported below each region label, highlighting the smaller sample size for RC compared with ET, TC, and WT.

changes in surface-distance performance. Average inference time per case was approximately 45s for MedNeXt, 28s for STU-Net, and 7.5s for nnU-Net, for a combined ensemble inference time of approximately 81s per case.

## 4 Discussion

In this work, we investigated a multi-network ensemble strategy for automated segmentation of pre- and post-treatment brain metastases in the BraTS 2026 challenge setting. The proposed approach combined three complementary 3D segmentation models, namely MedNeXt, nnU-Net, and STU-Net, and fused their softmax outputs using a two-stage weighting strategy (Stage 1 + Stage 2) followed by a final post-processing stage (Stage 3). The first stage optimized the ensemble for the non-RC labels, whereas the second stage treated the RC as an independent class-specific prediction problem. This design was motivated by the substantial class imbalance of RC and by the observation that the optimal models for the main tumor components were not necessarily optimal for the resection-cavity label.

The internal validation experiments showed that the three individual models achieved comparable performance for the non-RC regions, with STU-Net obtaining the highest overall single-model mean DSC and nnU-Net performing particularly well for SNFH. However, RC remained the most challenging label across all models, confirming the difficulty of segmenting post-surgical regions in a strongly imbalanced setting. The Stage 1 ensemble improved the non-RC labels, achieving DSC values of 0.535 for NETC, 0.738 for SNFH, and 0.728 for ET, corresponding to a mean non-RC DSC of 0.667. The relatively small differences among the best-performing weight combinations suggest that the ensemble did

not depend on a single unstable configuration, but rather on a broad performance plateau dominated by the contribution of nnU-Net and STU-Net, with MedNeXt providing a smaller but useful complementary contribution.

The independent RC optimization in Stage 2 improved the overall balance of the segmentation output. In particular, assigning RC through a separately thresholded probability map allowed this label to be detected without directly competing with NETC, SNFH, and ET in the same argmax operation. This was beneficial because RC is rare, visually heterogeneous, and present only in post-treatment cases. On the internal validation set, Stage 2 increased the RC DSC to 0.416 while preserving the performance of the non-RC labels. This finding supports the use of class-specific calibration when one label has a substantially different prevalence and appearance compared with the remaining tumor subregions.

The final post-processing Stage 3 further improved the validation leaderboard performance. Removing connected components smaller than 50 voxels increased the mean leaderboard DSC from 0.609 to 0.627, with improvements across all reported regions: ET increased from 0.680 to 0.697, RC from 0.386 to 0.419, TC from 0.702 to 0.713, and WT from 0.666 to 0.679. This indicates that a conservative connected-component filter was effective in reducing false-positive detections without excessively removing true lesions. This trade-off is particularly important in brain metastasis segmentation, where lesion-wise metrics penalize both missed lesions and spurious detections. In contrast to more aggressive post-processing strategies, the selected threshold was intentionally kept low, since very small metastases and small post-treatment residual components may still be clinically relevant.

Overall, the results suggested that the proposed ensemble strategy improved robustness compared with individual models and that class-specific fusion was particularly useful for handling the RC label. The strongest gains were obtained by separating the optimization of the common tumor components from the rare post-surgical class, rather than by applying a single global weighting scheme to all labels. This was consistent with the clinical and imaging heterogeneity of the task: ET, NETC, and SNFH are primarily tumor-related structures, whereas RC represents a treatment-related cavity whose appearance depends on surgical status and post-treatment changes.

Several limitations should be acknowledged. First, model selection and weight optimization were performed on a single internal validation split, and therefore the stability of the selected weights was not assessed across multiple cross-validation folds. Second, the connected-component threshold was selected based on previous challenges BraTS and may not be optimal for all lesion sizes or treatment stages. Finally, although the ensemble improved the overall performance, it increased computational and storage requirements because probability maps from three separate models had to be generated and retained. This overhead has practical implications for real-time clinical deployment: running three networks increases inference time, memory footprint, and weight storage relative to a single model, and the weighted-averaging step requires holding multiple probability maps in memory simultaneously. Lighter deployable configurations are nonetheless feasible,

for example a simplified ensemble using a subset of the models or knowledge distillation into a single network. Finally, the 50-voxel post-processing threshold, although effective in reducing small spurious predictions, was not systematically evaluated across different lesion sizes, and further investigation is needed to identify more appropriate threshold values for different lesion characteristics.

Future work should include training the final models on the full labeled training set, evaluating the stability of the ensemble weights across cross-validation folds, and performing lesion-size-stratified analyses to better understand the trade-off between false-positive removal and sensitivity to small metastases. Additional improvements may also be obtained by incorporating RC-specific sampling strategies, post-treatment case classifiers, or adaptive post-processing rules based on lesion size, anatomical location, and class-specific probability calibration. Furthermore, although the model was trained on a multi-centric dataset, future work should evaluate its robustness across different scanners and acquisition protocols by applying it to external datasets. Despite these limitations, the proposed two-stage ensemble provides a practical and effective strategy for multi-class brain metastasis segmentation in a heterogeneous challenge setting.

**Acknowledgments.** Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097).

**Disclosure of Interests.** The authors declare no conflicts of interest.

## References

1. Baid, U., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels for the pre-operative scans of the tcga-gbm collection (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific Data **4**(1) (2017). https://doi.org/10.1038/sdata.2017.117
4. Cerina, V., Rui, C.B., Di Cristofori, A., Ferlito, D., Carrabba, G., Giussani, C., Basso, G., De Bernardi, E.: Implication of tumor morphology and mri characteristics on the accuracy of automated versus human segmentation of gbm areas. Scientific Reports **15**(1), 2160 (2025)
5. Cho, S.J., Sunwoo, L., Baik, S.H., Bae, Y.J., Choi, B.S., Kim, J.H.: Brain metastasis detection using machine learning: a systematic review and meta-analysis. Neuro-Oncology **23**(2), 214–225 (2021). https://doi.org/10.1093/neuonc/noaa232
6. 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). https://doi.org/10.1002/jmri.26766

7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop (BrainLes). Springer (2022)
8. Hatamizadeh, A., et al.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 574–584 (2022)
9. He, Y., Yang, D., Roth, H., Zhao, C., Xu, D.: Dints: Differentiable neural network topology search for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
10. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., Qiao, Y.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023)
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
12. Jain, I., Willems, S., Latre, S., De Schepper, T.: On-the-fly data augmentation for brain tumor segmentation. In: Bakas, S., Dennis, E., Astaraki, M., Baid, U., Conte, G.M., Foltyn-Dumitru, M., Jiang, Z., Labella, D., Metz, M.C., Anazodo, U., de Verdier, M.C., Kofler, F., Li, H.B., Linguraru, M.G., Maleki, N. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 48–62. Springer Nature Switzerland, Cham (2026)
13. 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
14. 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
15. Lin, N.U., Lee, E.Q., Aoyama, H., et al.: Response assessment criteria for brain metastases: proposal from the RANO group. The Lancet Oncology **16**(6), e270–e278 (2015). https://doi.org/10.1016/S1470-2045(15)70057-4
16. Machura, B., et al.: Deep learning ensembles for detecting brain metastases in longitudinal multi-modal mri studies. Computerized Medical Imaging and Graphics **116**, 102401 (2024). https://doi.org/10.1016/j.compmedimag.2024.102401
17. Maleki, N., 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
18. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
19. Moawad, A.W., Janas, A., Baid, U., Ramakrishnan, D., et al.: The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri. arXiv preprint arXiv:2306.00838 (2023). https://doi.org/10.48550/arXiv.2306.00838
20. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. arXiv preprint arXiv:1810.11654 (2018)
21. Oktay, O., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

22. Ostrom, Q.T., Wright, C.H., Barnholtz-Sloan, J.S.: Brain metastases: epidemiology. In: Handbook of Clinical Neurology, vol. 149, pp. 27–42. Elsevier (2018). https://doi.org/10.1016/B978-0-12-811161-1.00002-5
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2015. LNCS, vol. 9351, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.: Mednext: Transformer-driven scaling of convnets for medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2023)
25. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
26. de Verdier, M.C., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Tahon, N.H., Foltyn-Dumitru, M., Zhang, J., Alafif, M., Baig, S., Chang, K., D'Anna, G., Deptula, L., Gupta, D., Haider, M.A., Hussain, A., Iv, M., Kontzialis, M., Manning, P., Moodi, F., Nunes, T., Simon, A., Sollmann, N., Vu, D., Adewole, M., Albrecht, J., Anazodo, U., Chai, R., Chung, V., Faghani, S., Farahani, K., Kazerooni, A.F., Iglesias, E., Kofler, F., Li, H., Linguraru, M.G., Menze, B., Moawad, A.W., Velichko, Y., Wiestler, B., Altes, T., Basavasagar, P., Bendszus, M., Brugnara, G., Cho, J., Dhemesh, Y., Fields, B.K.K., Garrett, F., Gass, J., Hadjiiski, L., Hattangadi-Gluth, J., Hess, C., Houk, J.L., Isufi, E., Layfield, L.J., Mastorakos, G., Mongan, J., Nedelec, P., Nguyen, U., Oliva, S., Pease, M.W., Rastogi, A., Sinclair, J., Smith, R.X., Sugrue, L.P., Thacker, J., Vidic, I., Villanueva-Meyer, J., White, N.S., Aboian, M., Conte, G.M., Dale, A., Sabuncu, M.R., Seibert, T.M., Weinberg, B., Abayazeed, A., Huang, R., Turk, S., Rauschecker, A.M., Farid, N., Vollmuth, P., Nada, A., Bakas, S., Calabrese, E., Rudie, J.D.: The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri (2024). https://doi.org/10.48550/ARXIV.2405.18368
27. Wang, T.W., et al.: Brain metastasis tumor segmentation and detection using deep learning algorithms: a systematic review and meta-analysis. Radiotherapy and Oncology **190**, 110007 (2024)