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.

Resection-Cavity-Aware Segmentation of Brain Metastases for BraTS 2026

Alex Kagozi[1], Moinak Bose[1], Longwei Wang[1], and Lina Chato[1]

Department of Computer Science, University of South Dakota, Vermillion, SD 57069, USA
{alex.kagozi,moinak.bose}@coyotes.usd.edu,
{longwei.wang,lina.chato}@usd.edu

Abstract. Accurate segmentation of brain metastases on pre- and post-treatment multiparametric MRI is challenging due to heterogeneous lesion appearance, treatment-related anatomical change, and severe class imbalance for the resection cavity (RC), present in only 12.9% of cases and exclusively post-treatment. We present our approach to the BraTS 2026 Brain Metastases Segmentation Challenge, benchmarking nnU-Net, Residual Encoder U-Net, and MedNeXt-L, and addressing the RC imbalance with a dedicated specialist trained exclusively on RC-positive cases and fused via a per-fold agreement-threshold channel override, alongside a separately investigated prototype-guided nnU-Net trainer with class-specific prototype memory at the encoder bottleneck. Our submitted pipeline pairs MedNeXt-L (five-fold) with the RC specialist and anatomically-aware post-processing, reaching RC Dice 0.569 (+0.172 over MedNeXt-L alone) and the best validation Dice aggregate (0.6573) among the eight configurations we evaluated. A lesion-wise detection analysis shows this Dice gain trades away RC detection precision, underscoring that rare, treatment-specific labels require explicit label-aware modeling and evaluation beyond a single overlap metric. Our team submission to BraTS MET 2026 is "Prototype-nnunet".

Keywords: Brain metastases · BraTS 2026 · nnU-Net · MedNeXt · Resection cavity · Prototype learning · Class imbalance · Lesion detection · Medical image segmentation

1 Introduction

Brain metastases (BM) are the most common intracranial malignancy, occurring in 20–40% of patients with systemic cancer [1]. Multiparametric MRI (mpMRI) is the modality of choice for BM detection, providing complementary contrasts through native T1-weighted (T1n), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2w), and T2-FLAIR sequences. Automated segmentation enables reproducible assessment of tumor volume and treatment response, critical for radiosurgery planning and longitudinal monitoring.

BraTS has anchored brain tumor segmentation benchmarking since Menze et al. [2], with later iterations refining its datasets and evaluation protocol [3,

4]. The BraTS 2026 Brain Metastases Segmentation Challenge [5] frames this as five-class voxel labeling (background, NETC, SNFH, ET, RC) spanning both pre- and post-treatment imaging, and evaluates a lesion-level F1 detection metric alongside Dice and NSD. RC is exclusive to post-treatment cases, but only those with prior surgical resection, creating severe label imbalance. The 2025 Lighthouse challenge report [6], which first introduced the RC label, notes the same pattern, with RC lagging well behind the other three classes.

The nnU-Net framework [7] is a strong, self-configuring baseline for 3D medical image segmentation, automating preprocessing, architecture selection, and training configuration. Residual Encoder U-Net (ResEnc) [8] improves feature extraction via residual skip connections, while MedNeXt [9] integrates ConvNeXt-inspired large-kernel blocks for superior long-range spatial modeling. We adopt this architecture family rather than proposing a new backbone, focusing on the RC-specific label-imbalance problem, studied separately in the glioma post-operative literature but rarely as one class in a joint multi-class task.

Motivated by the specific challenges of this task, our work makes four contributions: a systematic benchmarking of four 3D CNN architectures (nnU-Net, ResEnc-M, ResEnc-L, MedNeXt-L) trained from scratch for the five-class BraTS 2026 MET task; an RC specialist trained exclusively on the 167 RC-positive post-treatment cases, combined with a per-fold agreement-threshold channel override and anatomically-aware post-processing, yielding the highest RC Dice (0.569) among all evaluated configurations; a prototype-guided nnU-Net trainer with a `3D Prototype Fusion Block` at the encoder bottleneck, achieving the best RC result (Dice 0.456) among general full-dataset models; and a lesion-wise detection (F1) analysis showing the fold-agreement override trades RC detection precision for voxel-level Dice, invisible to Dice/NSD alone.

2 Dataset

The BraTS 2026 MET dataset [5, 10, 6] comprises 1,296 training cases from 810 patients and 179 validation cases, each with four co-registered mpMRI sequences (T1n, T1ce, T2w, T2f). Training cases include expert voxel-wise segmentation masks (background 0, NETC 1, SNFH 2, ET 3, RC 4); validation ground truth is blind and not released to participants. Following challenge conventions, we report lesion-wise DSC and NSD for the composite regions ET (label 3), Tumor Core (TC = NETC∪ET), Whole Tumor (WT = NETC∪SNFH∪ET), and RC (label 4). Dataset composition is summarized in Table 1. All cases were processed through the nnU-Net planning and pre-processing pipeline.

3 Preprocessing and Data Augmentation

All models used the nnU-Net automatic preprocessing pipeline [7]: spatial resampling to the dataset-median voxel spacing, per-modality foreground z-score normalization, non-zero bounding-box cropping, and conversion to nnU-Net's internal format. Fingerprinting determined patch size (128^3 for most folds), target

Table 1: BraTS 2026 MET dataset composition.

Training set (1,296 cases, 810 patients)	**Cases/Samples**	**Fraction**
Pre-treatment (session-000)	685	52.9%
Post-treatment (session-001+)	611	47.1%
NETC present	698	53.9%
SNFH present	1148	88.6%
ET present	1250	96.5%
RC present (77 of 810 patients)	167	12.9%
Post-treatment, RC absent	444	34.3%
Validation set (blind, no local ground truth)	**179**	—

spacing, network depth, and batch size; the RC specialist's pre-processing used only the 167 RC-positive cases (Section 4.4). Augmentation included random rotations, scaling, elastic deformations, Gaussian noise/blur, brightness/contrast, low-resolution simulation, and gamma, applied on the fly; no external data or pretrained weights were used. Inference retained nnU-Net's default eight-way mirroring TTA ensemble for all configurations.

4 Methods

4.1 Baseline nnU-Net

The standard 3D full-resolution nnU-Net [7] served as the primary baseline, automatically configuring a plain convolutional U-Net [11]. Training used 1,000 epochs, polynomial learning-rate decay ($10^{-2} \to 10^{-5}$), 5-fold cross-validation on all 1,296 cases, and a combined Dice + cross-entropy loss, $\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}$, each weighted equally and applied at all deep-supervision scales.

4.2 Residual Encoder U-Net (ResEnc-M and ResEnc-L)

Residual Encoder U-Net variants [8] replace nnU-Net's plain convolutional encoder blocks with residual blocks, improving gradient propagation and feature reuse while preserving the self-configuring pipeline. We evaluated ResEnc-M ($\sim$31M parameters) and ResEnc-L ($\sim$62M), both trained with the same Dice+CE objective and 5-fold cross-validation as the baseline.

4.3 MedNeXt-L

MedNeXt-L [9] is a ConvNeXt-inspired architecture using large-kernel depthwise separable convolutions for a larger receptive field than standard 3^3 convolutions. We trained it with a custom `MedNextTrainerL` (expansion ratios $[3, 4, 8, 8, 8, 8, 8, 4, 3]$, kernel size k=3, deep supervision disabled due to decoder-interface incompatibility with nnU-Net's multi-scale losses). Unlike the other architectures, which use nnU-Net's default SGD schedule (Section 5), MedNeXt-L

used AdamW (lr $= 10^{-4}$, weight decay 10^{-5}) with polynomial decay (power 0.9), following the original recipe, with 5-fold cross-validation and the same Dice+CE loss.

4.4 Post-Treatment RC Specialist and Fold-Agreement Override

Because RC is absent from all pre-treatment cases and present in only 12.9% of all cases, general-purpose models learn to suppress RC predictions. We therefore trained a dedicated RC specialist using *only* the 167 RC-positive training cases (77 of the 810 total patients). This maximizes RC-specific gradient signal but leaves the specialist with no notion of when *not* to predict RC; false positives are controlled below and in Section 4.6.

The specialist is a standard 3D nnU-Net with 4-fold cross-validation. At inference, each fold predicts an RC mask, fused with a 2 out of 4 fold-agreement threshold to obtain $\hat{m}_{RC}^{\text{spec}}$ (a tie, not a strict majority). We apply a *union* rule rather than fully replacing the general model's decision: a voxel is RC if the specialist threshold is met *or* MedNeXt-L already predicted RC there, else the ensemble's arg-max class:

$$\hat{y}(x) = \begin{cases} \text{RC}, & \hat{m}_{RC}^{\text{spec}}(x){=}1 \ \vee \ \hat{y}_{\text{MedNeXt}}(x){=}\text{RC}, \\ \hat{y}_{\text{MedNeXt}}(x), & \text{otherwise.} \end{cases} \tag{1}$$

The specialist can therefore only *add* RC voxels, never removing one the ensemble already predicted, decoupling the RC decision from the general backbone at the cost of being unable to correct the general model's own RC false positives, partially mitigated by post-processing (Section 4.6).

4.5 Prototype-Guided nnU-Net Trainer

To investigate whether class-specific internal representations help rare-label segmentation, we modified a ResEnc-M backbone with a `3D Prototype Fusion Block` at the encoder bottleneck. The block maintains $K{=}8$ learnable prototypes $\mathbf{p}_{c,k} \in \mathbb{R}^{320}$ per class $c \in \{0, \ldots, 4\}$ (40 total). ℓ_2-normalized bottleneck features attend to the prototype bank via cosine similarity and four-head cross-attention; a learned sigmoid gate blends attended and original features, $\hat{\mathbf{F}} = g \odot \mathbf{F}' + (1 - g) \odot \mathbf{F}$, so the network can fall back to unmodified features when prototype evidence is uninformative. An auxiliary prototype cross-entropy loss (weight $\lambda_{\text{proto}}{=}0.1$) against downsampled ground truth encourages class specialization. Training used 5-fold cross-validation and the nnU-Net SGD schedule (Section 5). The design combines compact prototypical representations [12] with interpretable reasoning [13], hypothesizing that class-specific prototypes reduce ambiguity between visually similar classes (e.g., ET and SNFH).

4.6 Post-Processing

Pre-treatment RC zeroing. Since RC is anatomically impossible in pre-treatment cases, RC voxels were zeroed for any pre-treatment case (session-000), identified

via the session identifier. This enforces a hard clinical constraint and eliminates a common false-positive source.

Connected-component filtering. Predicted RC regions smaller than 50 voxels were relabeled as background, a threshold chosen empirically to suppress isolated noise while preserving genuine small RC regions.

5 Experimental Setup

All training and evaluation used four NVIDIA A100 GPUs (40 GB each) on the Nautilus Kubernetes cluster of the National Research Platform [14]. Training used nnU-Net's default 250 steps/epoch for 1,000 epochs per configuration. The nnU-Net baseline, ResEnc-M/L, the RC specialist, and the prototype trainer used nnU-Net's default SGD (Nesterov $\mu = 0.99$, decay $10^{-2} \to 10^{-5}$); MedNeXt-L used AdamW (Section 4.3). Inference used sliding-window step size 0.25 (75% overlap) and eight-way mirroring TTA for all configurations. The RC specialist used 4-fold cross-validation (matching the submitted configuration); all others used 5-fold. Table 2 also reports two ensembles: a 3-model softmax average (nnU-Net + ResEnc-M + MedNeXt-L) and that ensemble with the RC fold-agreement override ("Ensemble + RC v2"). Our code and implementation details are available in this GitHub repository: `https://github.com/kagozi/brats26-mets`.

6 Results

6.1 Quantitative Validation Results

Table 2 reports lesion-wise DSC and NSD on the BraTS 2026 validation set (179 cases) for all completed configurations. Our final submitted pipeline (MedNeXt-L + RC override) achieved the highest validation score of **0.6573** among the eight candidates evaluated: the unweighted mean of the four class-wise lesion-wise DSC values (ET, RC, TC, WT) in Table 2. The underlying per-case scores come from the official Synapse/MedPerf evaluation platform [15], downloaded after we submitted each candidate configuration's validation-set predictions for scoring, but this aggregate itself is our own Dice-only selection metric rather than an official leaderboard ranking output, and we used it to choose the final Docker submission. It therefore excludes NSD and the detection F1 in Table 3, discussed next. MedNeXt-L achieves the highest DSC and NSD for ET, TC, and WT among single models. Among non-specialist, non-prototype models, ResEnc-M has the highest RC DSC (0.434), while Prototype Guided nnUNet reaches 0.456, the strongest among full-dataset models. The three-model ensemble did not outperform MedNeXt-L and slightly reduced RC DSC to 0.394, suggesting softmax averaging dilutes architecture-specific strengths rather than uniformly improving all labels.

Table 2: Lesion-wise validation DSC and NSD for all evaluated configurations. Best values per metric are shown in **bold**. The submitted pipeline is indicated by †.

Configuration	DSC ↑				NSD ↑			
	ET	**RC**	**TC**	**WT**	**ET**	**RC**	**TC**	**WT**
nnU-Net baseline	0.6854	0.3686	0.7152	0.6663	0.7464	0.2735	0.7629	0.6890
ResEnc-M	0.6849	0.4344	0.7107	0.6728	0.7466	0.3543	0.7615	0.6954
ResEnc-L	0.6923	0.4220	0.7100	0.6742	0.7547	0.3521	0.7601	0.6975
MedNeXt-L	**0.7031**	0.3969	**0.7279**	**0.6853**	**0.7673**	0.3232	**0.7783**	**0.7081**
Ensemble (3-model)	0.6995	0.3938	0.7246	0.6802	0.7604	0.3022	0.7726	0.7029
Ensemble + RC v2	0.6980	0.5047	0.7188	0.6728	0.7592	0.4457	0.7668	0.6951
MedNeXt + RC override†	0.6888	**0.5687**	0.6974	0.6741	0.7527	**0.5043**	0.7541	0.6992
Prototype Guided nnUNet	0.6917	0.4558	0.7213	0.6833	0.7536	0.3894	0.7723	0.7060

The RC specialist provides a decisive improvement. Compared with MedNeXt-L alone (RC DSC 0.397), MedNeXt + RC override reaches RC DSC 0.569 (+0.172, +43%) and RC NSD 0.504 (+0.181), at a modest TC DSC cost (−0.031) from occasional over-segmentation of residual ET as RC. Prototype Guided nnUNet reaches RC DSC 0.456, +0.021 over its ResEnc-M backbone, without sacrificing ET, TC, or WT.

6.2 Lesion-wise Detection Performance (F1)

BraTS 2026 additionally scores a lesion-level detection F1 (true positive only if a prediction sufficiently overlaps a ground-truth instance), complementary to DSC/NSD since a model can overlap well on detected lesions while still being penalized for missed or spurious ones. Table 3 reports all-instance detection F1 for ET/TC/WT/RC, plus RC precision, recall, and false positives per case, for all eight configurations.

The submitted override has the highest RC DSC (0.569) but the *lowest* RC F1 (0.283), driven by a recall/precision trade. Relative to MedNeXt-L alone, it raises RC recall from 0.452 to 0.675 while precision falls from 0.820 to 0.270. ET, TC, and WT detection F1 track their DSC rankings closely, so this dissociation is specific to RC and the override mechanism.

Splitting predictions by ground-truth RC status sharpens this further. Since the official validation-set labels are never released to participants, ground-truth RC status per case was inferred from the per-case lesion-instance counts returned by the official evaluation platform: a case is counted as RC-negative if the platform reports zero ground-truth RC instances (true positives plus false negatives equal zero) and RC-positive otherwise. On the 156 validation cases with no RC in the ground truth, the submitted pipeline predicts at least one spurious RC lesion in 33 cases (21.2%, 50 false-positive lesions in total), while on the 23 RC-positive cases it reaches lesion-wise RC DSC 0.569 (consistent with the aggregate RC DSC of 0.5687 in Table 2) and recall 67.7%. With no RC-negative training

Table 3: Lesion-wise detection F1 (ET/TC/WT/RC) and RC-specific precision, recall, and false positives per case (179 validation cases). Best value per column in **bold**. The submitted pipeline is indicated by †.

Configuration	Detection F1				RC Detection		
	ET	TC	WT	RC	Prec.	Recall	FP/case
nnU-Net baseline	0.7983	**0.8125**	**0.7964**	0.4554	0.5987	0.4833	0.057
ResEnc-M	0.7707	0.7831	0.7715	0.4083	0.4851	0.5472	0.102
ResEnc-L	0.7993	0.8052	0.7940	0.3799	0.3951	0.5472	0.147
MedNeXt-L	0.7936	0.8032	0.7885	**0.5052**	**0.8204**	0.4515	**0.017**
Ensemble (3-model)	**0.8009**	0.8109	0.7935	0.4845	0.5982	0.4833	0.057
Ensemble + RC v2	0.7996	0.8098	0.7916	0.3293	0.3226	0.6111	0.224
MedNeXt + RC override†	0.7860	0.7904	0.7828	0.2825	0.2701	**0.6752**	0.319
Prototype Guided nnUNet	0.7806	0.7954	0.7838	0.4535	0.4992	0.5790	0.102

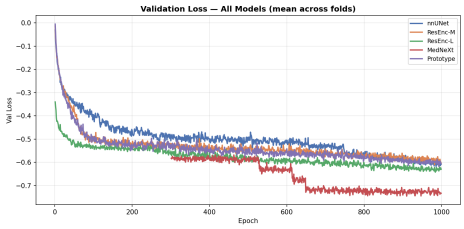


Fig. 1: Final-epoch validation loss comparison across all five full-dataset trainers (5-fold mean), showing a ranking broadly consistent with the DSC results in Table 2.

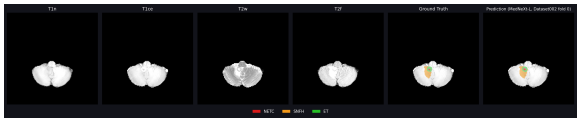
exposure, this false-positive rate is expected, and only partially controlled by the fold-agreement threshold and post-processing.

6.3 Training Dynamics

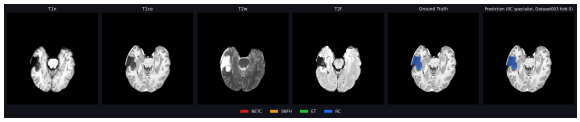
Figure 1 compares 5-fold averaged validation loss (nnU-Net's compound Dice+CE loss) across all five full-dataset trainers. ResEnc-L converges fastest early on but this does not translate into a correspondingly large final DSC gap over ResEnc-M (Table 2), suggesting diminishing returns from added capacity once loss converges. The prototype-guided trainer tracks its ResEnc-M backbone closely, indicating the added parameters do not destabilize training.

7 Qualitative Examples

Since the official validation set is blind, true prediction-vs-ground-truth comparison is only possible on training-set cases; Figure 2 shows held-out fold-0 predictions. Fig. 2a (MedNeXt-L) closely matches NETC/SNFH/ET with no



(a) Pre-treatment, predicted by MedNeXt-L.



(b) Post-treatment with RC, predicted by the RC specialist.

Fig. 2: Held-out fold-0 cross-validation examples, BraTS-MET-00004-000 and BraTS-MET-01121-002 (training-set cases, not the blind official validation set): T1n, T1ce, T2w, T2f, ground truth, and prediction (NETC red, SNFH orange, ET green, RC blue).

RC in either ground truth or prediction. Fig. 2b (RC specialist) shows a large, well-localized cavity with close agreement; RC is not always this prominent, often much smaller and confined to a narrow slice range, part of why the average RC DSC (0.569) reflects many harder cases and motivates both the specialist and the sub-50-voxel filter (Section 4.6).

8 Discussion

Our results reveal a consistent pattern. Architectural improvements to the general-purpose backbone (nnU-Net → ResEnc → MedNeXt-L) yield steady ET/TC/WT DSC gains but not RC. MedNeXt-L has the best general DSC but the second-lowest RC DSC among single models (0.397), showing that stronger global feature extraction does not help a label absent from over 87% of training cases.

Our specialist trains exclusively on the 167 RC-positive cases, a more extreme correction than restricting to the 611 post-treatment cases, so a union override rather than a learned probability is used; with no RC-negative exposure, predictions must be gated by fold-agreement. Soft blending underperformed this hard override: replacing the ensemble's RC channel with the specialist's raw softmax, without the discrete fold vote (Table 2, "Ensemble + RC v2"), gave RC DSC 0.505 and F1 0.329 — worse Dice but better detection than the submitted override (0.569 / 0.283), showing the fold-agreement vote itself drives the trade-off. A calibrated multi-model blend (general: 0.45 MedNeXt-L + 0.35 prototype + 0.20 ResEnc-M; RC: 0.60 specialist + 0.40 prototype) gave RC DSC 0.541 and aggregate 0.6525, both below the submission. Intersection and learned-gating fusion, and a fold-agreement threshold sweep, remain future work.

Prototype-guided representation. The prototype module improves RC by +0.021 over its non-prototype backbone without any training-data restriction, suggest-

ing class-specific feature attractors help separate RC signal from visually overlapping classes like ET and SNFH, while offering an interpretable representation for clinical quality assurance.

Limitations. Several limitations should be acknowledged. First, we have not yet combined the prototype trainer with the RC specialist, and the RC-zeroing rule depends on reliable treatment-session metadata. In addition, performance on the hidden test set may differ from validation performance because of distribution shift. Cross-validation was performed using session-level rather than patient-level partitioning. Of the 810 training patients, 287 (35%) contributed more than one session, and 23 contributed both an RC-positive and an RC-negative session. However, this overlap does not affect the official results reported for the disjoint 179-case validation set in Tables 2–3. The RC specialist was also trained only on RC-positive cases rather than on all 611 post-treatment cases using balanced sampling. Furthermore, we did not evaluate alternative fold-agreement thresholds ($k \in \{1, 2, 3, 4\}$ of 4) beyond the submitted setting of $k = 2$. These extensions remain important directions for future work because generating revised validation results would require resubmission, which was not available within the revision window. Finally, the observed RC dissociation is specific to the all-instance detection metric. Under the challenge's small-lesion breakdown, all configurations produced zero true positives for the 10 small RC lesions. Therefore, the observed trade-off appears to be confined to larger, well-formed cavities.

9 Conclusion

We presented a training-from-scratch brain metastases segmentation pipeline for BraTS 2026 combining MedNeXt-L, an RC specialist trained exclusively on RC-positive post-treatment cases with a union-based fold-agreement override, and anatomically-aware post-processing, achieving RC DSC 0.569 (a 43% relative improvement over MedNeXt-L alone) and the best validation Dice aggregate (0.6573) among the eight configurations evaluated. Our prototype-guided trainer further shows that class-specific bottleneck representations improve rare-label segmentation without data subsetting. Voxel-overlap and lesion-detection metrics can disagree sharply for rare, spatially small labels such as RC, which benefits most from explicit clinical-constraint modeling and label-specific training rather than architectural scaling alone.

Acknowledgements

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

Disclosure of Interests

The authors declare no competing interests relevant to this article.

References

1. Nayak, L., Lee, E.Q., Wen, P.Y.: Epidemiology of brain metastases. Current Oncology Reports **14**(1), 48–54 (2012)
2. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015)
3. Bakas, S., Reyes, M., Jakab, A., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)
4. Baid, U., Ghodasara, S., Mohan, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
5. BraTS 2026 Challenge: Brain Tumor Segmentation 2026. Synapse Challenge Portal (2026). `https://www.synapse.org/Synapse:syn74274097`
6. 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 preprint arXiv:2504.12527 (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. Isensee, F., Wald, T., Ulrich, C., et al.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. arXiv preprint arXiv:2404.09556 (2024)
9. Roy, S., Koehler, G., Ulrich, C., et al.: MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation. In: MICCAI, pp. 745–755. Springer (2023)
10. 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. arXiv preprint arXiv:2306.00838v3 (2024)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241. Springer (2015)
12. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4077–4087 (2017)
13. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: Deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 8930–8941 (2019)
14. Weitzel, D., Graves, A., Albin, S., et al.: The National Research Platform: Stretched, Multi-Tenant, Scientific Kubernetes Cluster. In: Practice and Experience in Advanced Research Computing 2025 (PEARC '25), Article 69, 1–5. ACM, New York, NY, USA (2025)
15. Karargyris, A., Umeton, R., Sheller, M.J., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence **5**, 799–810 (2023)