Precision Over Recall: Brain Metastasis Segmentation with a Learned False-Positive Rejector

Hari Girish[1], Michael Kong[1], Sherman Chan[1], Nihal Patil[1], Jany Chan[1,2], Arnab Chakravarti[1,2], and Simeng Zhu[1,2]

[1] The Ohio State University, Columbus, OH, USA
[2] The Ohio State University Comprehensive Cancer Center – Arthur G. James Cancer Hospital and Richard J. Solove Research Institute, Columbus, OH, USA
simengzhu@gmail.com

Abstract. Brain metastases are typically small, often multifocal, and evaluated by lesion-wise metrics that penalize every spurious predicted component, so that the balance between sensitivity and precision determines much of the achievable score. We present our entry to the BraTS-METS 2026 challenge (Task 1), a two-stage pipeline built on nnU-Net. The first stage is an ensemble of five cross-validation fold models from a single nnU-Net configuration, trained with an extended 5000-epoch schedule and a Dice-plus-TopK cross-entropy loss. The second stage is a learned connected-component false-positive rejector applied exclusively to the enhancing-tumor prediction: each predicted component is described by confidence, shape, and intensity features and scored by a gradient-boosted classifier trained solely on out-of-fold predictions, with a decision threshold derived from a lesion-wise Dice cost–benefit analysis and fixed before submission. In the course of development we evaluated four separate interventions intended to recover additional small enhancing lesions; each detected more true lesions, each also introduced a comparable or larger number of false ones, and each reduced lesion-wise enhancing-tumor Dice. Reversing this direction and removing components instead produced our largest single improvement, raising lesion-wise enhancing-tumor Dice from 0.6964 to 0.7187 and normalized surface distance from 0.7608 to 0.7839 on the challenge validation set, with corresponding gains in tumor core and whole tumor through subregion nesting. The same operation reduces the small-instance F1 score, a trade-off we report and analyze explicitly because both metric families contribute to the challenge ranking. For our system, precision rather than sensitivity was the binding constraint on lesion-wise performance.

Keywords: Brain metastases · Segmentation · nnU-Net · False-positive reduction · Lesion-wise evaluation · BraTS

1 Introduction

Brain metastases are the most common intracranial tumor in adults, and their number, size, and treatment response are tracked across serial MRI. Because the

enhancing-tumor (ET) region captures contrast-enhancing abnormal tissue, its accurate delineation is central to imaging-based response assessment, although enhancement alone is not specific for viable tumor: it primarily reflects blood–brain-barrier disruption, and radiation necrosis, pseudoprogression, and inflammation may all enhance after treatment. The non-enhancing tumor core, FLAIR hyperintensity, and any resection cavity must also be delineated to characterize disease burden. Manual delineation is labor-intensive and subject to substantial rater variability. Brain metastases also pose challenges distinct from primary gliomas: lesions are frequently subcentimeter, multiple lesions may occur in one patient, and the tissue compartments can be hard to distinguish across the multi-institutional acquisitions in BraTS-METS [4].

The nnU-Net framework [1] derives its preprocessing, network architecture, training schedule, and postprocessing automatically from the properties of a dataset, and outperformed most specialized solutions across 23 public biomedical segmentation datasets. It has become a strong and widely used baseline for brain-tumor and brain-metastasis pipelines.

Small-lesion ET segmentation nevertheless remains disproportionately difficult under the BraTS-METS evaluation protocol. BraTS-METS 2026 scores larger lesions with lesion-wise Dice and normalized surface distance (NSD) and smaller instances with an instance-level F1 score, separated by a volume threshold. The evaluator relates predicted structures to reference lesions using connected-component processing, lesion grouping, and region-specific logic; within lesion-wise Dice, unmatched false-positive and false-negative components are strongly penalized, including zero-Dice contributions. A few spurious ET components, often small enhancing structures or other metastasis mimics such as demyelinating lesions or subacute ischemic changes, can therefore depress a case's score far more than a voxel-wise measure would suggest.

We evaluated several interventions intended to increase small-lesion sensitivity, including foreground oversampling, inverse-lesion-size loss reweighting, and cascaded detectors contributing additional candidates. Raising sensitivity is expected to admit additional false positives; what matters under a lesion-wise metric is whether the recovered lesions outweigh them. In our study they did not: every sensitivity gain came with a comparable or larger increase in false-positive components, and lesion-wise ET Dice fell in each case, indicating that precision rather than recall was the principal limitation for our baseline system.

We therefore present a two-stage entry. The first stage averages the softmax probabilities of the five cross-validation fold models of a single nnU-Net configuration, trained with an extended 5000-epoch schedule and a Dice-plus-TopK cross-entropy loss. The second is a learned connected-component false-positive rejector applied only to the ET prediction, scoring each component on confidence, shape, and intensity features; its threshold was chosen by a lesion-wise Dice cost–benefit analysis on internal cross-validation predictions and fixed before submission. This postprocessing stage, not any change to the network, raised lesion-wise ET Dice from 0.6964 to 0.7187 while *reducing* small-instance F1 (ET

$0.4886 \rightarrow 0.4717$); since Dice, NSD, and F1 all enter the ranking, we report this trade-off rather than the favorable metric alone.

2 Data

2.1 Dataset

We use the BraTS-METS (Task 1) 2026 data release [3], comprising co-registered, skull-stripped multiparametric MRI with four sequences per study (post-contrast T1-weighted, T1c; pre-contrast T1-weighted, T1n; T2-weighted, T2w; and T2 FLAIR, T2f) and voxel-wise labels for the four regions of interest: non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), enhancing tumor (ET), and resection cavity (RC). All models were trained on the 1296 released training studies (comprising both pre- and post-treatment examinations); all model-selection decisions reported here were made on the 179-case online challenge validation set, for which reference annotations are withheld and evaluated through the challenge's MedPerf-based infrastructure [2]. Within a study the four sequences and the reference annotation share a common voxel grid and are skull-stripped, but the release is *not* resampled to a common grid *across* studies: the training set contains 62 distinct voxel spacings and 126 distinct volume dimensions, with per-voxel volumes spanning 0.065–2.42 mm^3 (median 1.00). Raw voxel counts are therefore not comparable between studies, and every lesion volume we report is converted to mm^3 using that study's own voxel volume.

Label and lesion-size distribution. Figure 1 characterizes the reference annotations of the training set by subregion, and two properties of the cohort directly shape the design of our method.

First, the subregions differ greatly in how often they occur. ET is annotated in 96% of studies and is markedly multifocal (7.8 lesions per study on average, 10 118 lesions in total), and SNFH in 89%; NETC appears in 54%. RC, by contrast, is present in only 13% of studies (253 lesions in total) and its lesions are the largest of any subregion (median 524 mm^3). RC is thus both rare and large, which is precisely why it contributes no small instances to the evaluation and reports a small-instance F1 of zero for every model we submitted (Section 4).

Second, and more consequentially for our method, the ET lesion-volume distribution is strongly right-skewed: the median ET lesion is only 40 mm^3, 41% of ET lesions fall below the 27 mm^3 boundary that places an instance in the detection-scored regime, and 91% are below 1000 mm^3. A model's ET score is therefore governed by how it handles a large number of very small components rather than a few large ones. This is the premise underlying the false-positive rejector of Section 3.4: when most lesions are small and numerous, each spurious small component carries substantial cost under a lesion-wise metric.

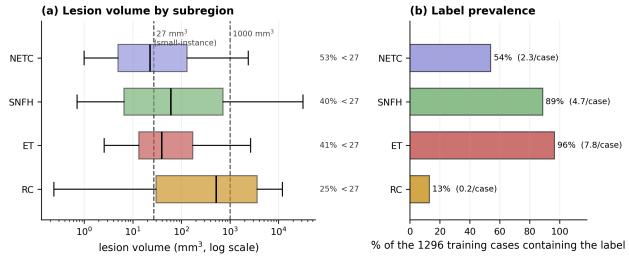

Fig. 1. Reference annotations of the BraTS-METS 2026 training set (1296 studies) by subregion: enhancing tumor (ET), surrounding non-enhancing FLAIR hyperintensity (SNFH), non-enhancing tumor core (NETC), and resection cavity (RC). **(a)** Distribution of individual lesion volumes on a logarithmic axis (boxes span the interquartile range, whiskers the 5th–95th percentiles); the dashed lines mark the 27 mm^3 boundary below which instances are scored by an instance-level F1 score rather than by lesion-wise overlap, and the 1000 mm^3 cut we use internally to designate "small" lesions. The percentage beside each row is the fraction of that subregion's lesions falling below 27 mm^3. **(b)** Fraction of studies in which each subregion is annotated at all, with the mean number of lesions per study in parentheses. Volumes are computed using each study's own voxel volume, since the release is not resampled to a common grid, and lesions are plain 26-connectivity components, so these statistics are descriptive of the cohort rather than evaluator-exact.

2.2 Preprocessing

All preprocessing is handled automatically by the nnU-Net planner using the dataset fingerprint. Images are resampled to the dataset median voxel spacing of $1.0 \times 0.859 \times 0.859$ mm (third-order spline interpolation for images, nearest-neighbor for labels) and intensity-normalized per channel using a Z-score computed over foreground voxels only. The planner selects a patch size of $112 \times 160 \times 128$ voxels and a batch size of 2 for the `3d_fullres` configuration. We adopt nnU-Net's default on-the-fly data augmentation (random rotations, scaling, Gaussian noise and blur, brightness and contrast, low-resolution simulation, gamma correction, and mirroring).

3 Methods

Our final entry, denoted M_{FPR}, is a two-stage pipeline: an ensemble of five cross-validation fold models from a single nnU-Net [1] configuration (the base model, M_{base}), followed by a learned connected-component false-positive rejector that operates exclusively on the enhancing-tumor (ET) predictions. The base model provides high sensitivity across all four subregions; the rejector supplies the precision that the segmentation network alone did not, and it improved ET Dice where every sensitivity-oriented intervention we tried had reduced it. Figure 2 gives an overview of the complete pipeline, including the out-of-fold procedure used to fit the rejector at training time.

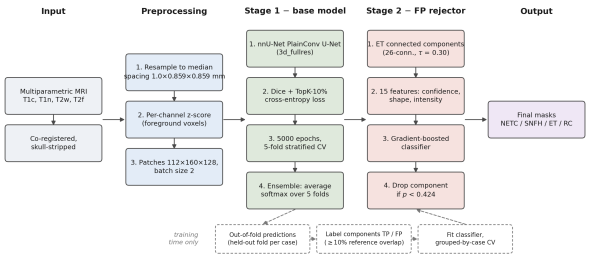

Fig. 2. Overview of the two-stage pipeline. Solid boxes show the inference path: the four sequences are preprocessed, segmented by an ensemble of five fold models (M_{base}, Stage 1), and the ET prediction is then filtered component-by-component by the learned rejector (M_{FPR}, Stage 2). The dashed path is *training time only*: components from out-of-fold predictions are labeled against the reference and used to fit the classifier under grouped-by-case cross-validation.

3.1 Base Segmentation Model

The base model is an nnU-Net with the default `PlainConvUNet` architecture (an encoder–decoder UNet with plain convolutional blocks and instance normalization), trained in the `3d_fullres` configuration. We depart from the nnU-Net defaults in two respects, both of which measurably affected ET performance:

- **Loss function.** We replace the default Dice + cross-entropy loss with a *Dice + TopK cross-entropy* loss (TopK-10%), in which the cross-entropy term is averaged over only the 10% highest-loss voxels per batch. This concentrates the gradient signal on ambiguous boundary and small-lesion voxels, which are disproportionately important for metastasis segmentation.
- **Training length.** We train for 5000 epochs (250 mini-batches each), five times the nnU-Net default of 1000. In our experiments, extending the training schedule was the single change that most increased small-lesion sensitivity: the mean number of correctly detected small-ET lesions per case rose from ∼2.0 (1000 epochs) to 2.48 (5000 epochs), and lesion-wise ET Dice from 0.674 to 0.696, indicating that for this configuration the training schedule, not model capacity, was the binding constraint.

The configuration is trained with five-fold cross-validation over a stratified split (folds balanced by lesion count so that high-metastasis cases are represented in every fold), using SGD with Nesterov momentum (0.99) and nnU-Net's polynomial learning-rate decay from an initial rate of 10^{-2}. At inference, the five fold models are ensembled by averaging softmax probabilities.

3.2 GPU Specification

Training and inference were performed on the Ohio Supercomputer Center cluster. Each of the five folds was trained on a single NVIDIA A100 (40 GB) with

Python 3.10, CUDA 12.4.1, and PyTorch 2.12.0; the 40 GB ceiling determined the $112 \times 160 \times 128$ patch size at batch size 2. Five-fold inference over the 179 validation cases took roughly eight minutes per fold, and the rejector trains on CPU in under a minute.

3.3 Major Libraries

The pipeline is built on **nnU-Net v2** [1] (PyTorch backend), which handles fingerprinting, preprocessing, network configuration, training, and five-fold ensembling. Postprocessing uses **scikit-learn** [5] for the rejector, **SciPy**/**scikit-image** for connected components and shape features, and **NiBabel**/**NumPy** for I/O and arrays. The submitted container pins nnU-Net 2.7.0, PyTorch 2.12.0 (CUDA 13.0), scikit-learn 1.7.2, SciPy 1.15.3, scikit-image 0.25.2, NiBabel 5.4.2, and NumPy 2.2.6; the rejector must be loaded under the scikit-learn version it was fitted with.

3.4 Postprocessing: Learned ET False-Positive Rejector

The dominant error mode we observed in the base model is false-positive enhancing-tumor components: small, low-confidence ET blobs that the lesion-wise evaluation penalizes severely, because an unmatched predicted component contributes zero Dice for that instance. In our experiments, approaches that raised sensitivity by lowering the effective detection threshold therefore *reduced* lesion-wise ET Dice. We instead allow the base model to operate at high sensitivity and remove false ET components in a dedicated precision stage.

Component extraction and labeling. For every training case we threshold the base model's out-of-fold (OOF) ET probability map at $\tau = 0.30$ and extract connected components under 26-connectivity. A component is labeled a *true positive* if at least 10% of its volume overlaps the reference ET mask, and a *false positive* otherwise. This procedure yielded 8300 components from the training set (7181 true positive, 1119 false positive). This rule defines the rejector's *training labels* only; it is not an evaluation criterion, does not reproduce the official evaluator's lesion matching, and no reported score depends on it.

Avoidance of leakage. Two precautions are used. First, the probability maps are produced under *cross-validation*: each training case is predicted only by the fold model that did not train on it, so the rejector learns from held-out predictions. Second, the rejector is validated with *grouped* five-fold cross-validation in which all components from a given case share a split, preventing components of the same patient from appearing in both training and validation folds. We also excluded two feature families that would have leaked or shifted: features derived from reference tumor-core or whole-tumor overlap (unavailable at inference, since ET is a subset of the tumor core), and cross-fold disagreement (identically zero for OOF predictions, where each case is predicted once, but non-zero at inference: a train/test mismatch).

Features. Each component is described by 15 features spanning three families: (i) *confidence*: the mean, peak, standard deviation, and 25th percentile of the ET probability within the component; (ii) *size and shape*: voxel count, bounding-box extent, solidity, and elongation, the last intended to capture elongated structures such as vessels as distinct from compact metastases; and (iii) *intensity and location*: the mean and local (component-versus-shell) contrast of the T1c signal and of the T1c−T1n subtraction image, together with the normalized component centroid.

Classifier. We use a gradient-boosted decision-tree classifier (scikit-learn's `GradientBoostingClassifier`; 300 estimators, learning rate 0.05, maximum depth 3, subsample 0.8) to predict the true/false label from these features. Under grouped cross-validation the classifier attains an area under the ROC curve of 0.86, indicating that false components are separable from true ones; the two populations differ in both confidence (mean ET probability 0.64 versus 0.83) and size (median 17 versus 109 voxels). The most informative features were the peak and mean ET probability, the T1c−T1n subtraction contrast, component size, and elongation.

Operating point and its selection. We select the decision threshold from a lesion-wise Dice cost–benefit analysis rather than by maximizing classification accuracy. Removing a genuine false positive gains approximately $\overline{\mathrm{DSC}}/(N+F)$ in expectation for that case, whereas erroneously removing a true positive forfeits only that lesion's own (typically low) Dice contribution, since the true positives a classifier discards are predominantly marginal, weakly segmented ones. The break-even drop-precision (the fraction of removed components that must genuinely be false positives for the operation to be neutral) is therefore only 33–41%, far below the near-unity value that removing solely true false positives would imply. We operate at a threshold of 0.424, which on out-of-fold data removes 20% of false-positive components at a drop-precision of 59.4%, clearing break-even by 18–26 percentage points.

Importantly, both the classifier and this threshold were fit and selected *exclusively* on the internal out-of-fold predictions of the 1296 training cases. No reference annotations from the 179-case challenge validation set were used at any point, and the threshold was fixed before the corresponding challenge submission was made; it was not tuned against challenge feedback.

Application. At inference we recompute ET components from the base model's ET probability map, score each with the trained rejector, and relabel every rejected component to background. Because ET $\subseteq$ TC $\subseteq$ WT, removing a false ET component simultaneously removes a false tumor-core and whole-tumor instance, so all three nested regions are affected by a single operation, while the resection-cavity prediction is left untouched by construction.

Table 1. Effect of the rejector on the challenge validation set (179 cases), for the three metric families used in the ranking. M_{FPR} is M_{base} after ET false-positive rejection. Lesion-wise Dice and NSD improve for the three nested regions while small-instance F1 decreases; RC is unaffected by construction (its lesions rarely fall below the small-instance threshold).

Metric	Model	ET	TC	WT	RC
Lesion-wise DSC	M_{base}	0.6964	0.7236	0.6827	0.5073
	M_{FPR}	**0.7187**	**0.7429**	**0.6950**	0.5073
	Δ	+0.0224	+0.0193	+0.0123	±0.0000
Lesion-wise NSD	M_{base}	0.7608	0.7757	0.7128	0.4419
	M_{FPR}	**0.7839**	**0.7942**	**0.7235**	0.4419
	Δ	+0.0232	+0.0185	+0.0107	±0.0000
Small-instance F1	M_{base}	**0.4886**	**0.4933**	**0.3895**	0.0000
	M_{FPR}	0.4717	0.4708	0.3726	0.0000
	Δ	−0.0170	−0.0225	−0.0169	±0.0000

4 Results

All results are reported on the 179-case BraTS-METS challenge validation set, whose reference annotations are withheld; scores were obtained from the official challenge evaluation. Our final submission (M_{FPR}) attains lesion-wise Dice of 0.7187 (ET), 0.7429 (TC), 0.6950 (WT), and 0.5073 (RC).

4.1 Effect of the False-Positive Rejector

Applied to the base model's predictions, the rejector removed 110 of 1039 ET components (10.6%). This reduced small-ET false positives per case by 49% ($1.79 \rightarrow 0.91$) at a cost of 0.21 true positives per case, an effective drop-precision of ∼81%, consistent with, and somewhat better than, the 59.4% estimated from out-of-fold data at the selected threshold. Figure 3 illustrates this mechanism qualitatively on two cross-validation cases.

Table 1 reports the effect on all three metric families used in the BraTS-METS 2026 ranking: lesion-wise Dice and NSD improve across the three nested regions (ET +0.0224 and +0.0232) and RC is unchanged, while small-instance F1 *decreases* (ET −0.0170). Averaged over the four regions, lesion-wise Dice rises from 0.6525 to 0.6660 and NSD from 0.6728 to 0.6859, while small-instance F1 falls from 0.3429 to 0.3288. The gains and the loss are therefore of comparable magnitude, and because the official ranking is computed across all participating teams and was not returned to participants, we cannot claim a demonstrated improvement in final rank. Since all three families enter the ranking, we report both directions rather than only the metric that favors our method, and analyze the mechanism in Section 5.

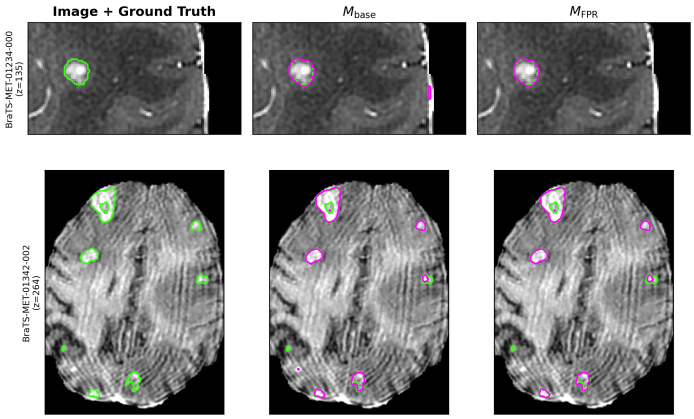

Fig. 3. Qualitative effect of the rejector on two cross-validation cases (challenge validation annotations are withheld). Each row is one axial T1c slice; columns show the reference annotation, M_{base}, and M_{FPR}. Green outlines the reference ET, magenta the predicted ET. **Top row** (BraTS-MET-01234-000, $z = 135$): a solitary metastasis; M_{base} delineates it but adds an elongated component at the cortical margin with no green counterpart, which M_{FPR} removes, leaving the metastasis unchanged. **Bottom row** (BraTS-MET-01342-002, $z = 264$): a multifocal case; M_{base} delineates the larger lesions and adds a small isolated component with no reference counterpart, which M_{FPR} removes while the matched lesions are retained.

Table 2. Sensitivity-oriented interventions, each relative to *its own* baseline (they were built on different baselines, so this is not a head-to-head comparison). ΔTP/ΔFP are per-case small-instance ET counts. Every intervention that raised true positives raised false positives at least as much and lost ET Dice; only the rejector, which lowers both, improved it.

Intervention	Baseline	ΔTP	ΔFP	ΔET DSC
Foreground oversampling	PlainConv 1k	−0.04	+0.20	−0.0178
Inverse-lesion-size loss	PlainConv 1k	+0.69	+0.65	−0.0140
Mask-conditioned cascade	ResEnc-M 1k	+0.47	+0.43	−0.0298
Additive ET-detector cascade	4-model ens.	+1.21	+1.32	−0.0700
FP rejector (ours)	M_{base}	**−0.21**	**−0.88**	**+0.0224**

4.2 Sensitivity-Oriented Interventions

Table 2 summarizes the interventions we evaluated that were intended to recover additional small ET lesions. Because these were developed at different points and built on different baseline configurations, each is reported as a change *relative to its own baseline* rather than against a single common reference; the comparison is therefore within-pair and not a head-to-head ranking of the interventions themselves.

The interventions follow a common pattern. Every intervention that increased small-ET true positives also increased small-ET false positives by a comparable or larger amount, and in each case lesion-wise ET Dice decreased. The additive ET-detector cascade illustrates the extreme: it recovered the most additional true lesions of any configuration we trained, yet added even more false positives. Foreground oversampling raised false positives without a corresponding sensitivity gain. By contrast, the rejector moves in the opposite direction on both counts: it removes far more false positives than true positives, and is the only intervention in the table that improved ET Dice.

5 Discussion

Our central finding is that, for the model family and operating regime we explored, lesion-wise ET Dice was limited by precision rather than sensitivity. Four independent attempts to recover additional small enhancing lesions each detected more true lesions and each nevertheless reduced ET Dice, because the detections arrived with a comparable or larger number of spurious components (Table 2). Removing components instead produced the largest improvement we obtained. This conclusion is empirical and specific to our system; it should not be read as a general claim that brain-metastasis segmentation is precision-limited.

The mechanism behind that trade-off is direct. Deleting a component the evaluator would have matched costs one true positive under F1, whereas deleting an unmatched component removes a zero-Dice contribution that depresses lesion-wise Dice. We chose our operating point against lesion-wise Dice; a threshold selected to maximize a combined objective would be expected to be more conservative.

Although the second stage of our pipeline is designed to reject false-positive lesions, it may also discard true lesions, a consequential error in applications such as treatment planning for stereotactic radiosurgery, where an omitted lesion goes untreated. This potential failure mode underscores the need for a human-in-the-loop approach in clinical implementation, which lies beyond the scope of this work. Any prospective deployment would require close collaboration with clinical stakeholders.

Several limitations qualify these results. First, our characterization of the rejected components is quantitative rather than aetiological: they are predominantly small and low-confidence, but we did not audit them radiologically. Second, the rejector was fitted on out-of-fold predictions from a single nnU-Net configuration; applying it to a deliberately higher-sensitivity base did not surpass the configuration reported here, suggesting it benefits from a base that is already well calibrated rather than merely more sensitive. Third, RC is untouched by construction and remains the weakest subregion (0.5073), plausibly data-limited given how rarely it appears in training. Finally, our internal lesion statistics use 26-connectivity components rather than the evaluator's definition; all reported scores come from the official evaluation.

6 Code Availability

The inference pipeline, the trained false-positive rejector, and the training scripts are available at `https://github.com/Hari-Girish/-Precision-Over-Recall-Brain-Metastasis-Segmentation-with-a-Learned-False-Positive-Rejector`. The five nnU-Net fold checkpoints are not tracked in the repository owing to their size; the repository documents how to stage them.

Acknowledgments. We acknowledge the compute resources offered by the Ohio Supercomputer Center.

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

References

1. 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). `https://doi.org/10.1038/s41592-020-01008-z`
2. 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`
3. Maleki, N., Amiruddin, R., Moawad, A.W., Yordanov, N., Gkampenis, A., Fehringer, P., Umeh, F., Chukwurah, C., Memon, F., Petrovic, B., 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`
4. Moawad, A.W., Janas, A., Baid, U., Ramakrishnan, D., Saluja, R., Ashraf, N., Maleki, N., Jekel, L., Yordanov, N., Fehringer, P., 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`
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)