Epistemic Uncertainty-Guided Supervision Weighting for Brain Metastasis Sub-region Segmentation

Ajesh Saviour Paravila[1] [0009-0008-5787-1866]

Institute of Biomedical Engineering and Nanomedicine
National Health Research Institutes, Miaoli, Taiwan
ajesh.saviour@gmail.com

Abstract. Brain metastasis segmentation presents persistent challenges in accurately delineating tumor subregions with high morphological variability and heterogeneous MRI appearance, particularly for non-enhancing tumor core (NETC) and resection cavity (RC) – structures that exhibit significant inter-case variability and appear in a subset of patients following surgical intervention. Standard segmentation frameworks apply uniform supervision across all classes, an assumption that is ill-suited to these subregions where inter-model disagreement is concentrated at boundaries and model confidence is unreliable[7]. In this work, we propose a reliability-weighted supervision framework that modulates the per-class training signal based on epistemic uncertainty estimated from cross-validation fold ensemble disagreement. Classes with high inter-fold disagreement within a case receive down-weighted supervision, reducing the influence of the most ambiguous subregions on model optimization. Built on top of nnU-Net without architectural modification, our framework is evaluated on the BraTS 2026 metastases segmentation task. Experimental results demonstrate consistent improvement over the vanilla nnU-Net baseline, with the largest gains observed in NETC and RC – precisely the subregions where inter-fold ensemble disagreement is highest. Ablation experiments reveal that pure epistemic weighting outperforms both aleatoric uncertainty estimation and epistemic weighting combined with inverse class frequency correction, suggesting that in the metastasis setting, rare subregion difficulty stems from appearance variability rather than representation deficit – a finding that distinguishes the metastasis from the pediatric failure mode and motivates task-specific reliability characterization.

Keywords: Brain tumor segmentation · nnUNet · Epistemic uncertainty · Class imbalance · Reliability weighting .

1 Introduction

Brain metastasis is the most common intracranial tumor in adults, arising from extracranial malignancies-most frequently lung cancer, breast cancer, and melanoma-with distinct imaging characteristics[8, 11] and outcomes that vary by primary tumor[9]. Automated segmentation of tumor subregions from multiparametric MRI-including enhancing tumor (ET), non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), and resection cavity (RC)-supports treatment planning, recurrence monitoring, and radiomic analysis [3, 4]. The BraTS challenge series [2] has driven substantial progress, with recent editions introducing increasingly nuanced evaluation criteria, including explicit benchmarking against human inter-rater variability in BraTS 2026 [1].

Despite this progress, NETC and RC remain challenging: NETC has poorly defined boundaries with surrounding SNFH, while RC is a highly variable post-surgical structure that is absent in many patients. Standard frameworks such as nnU-Net [5] apply uniform loss weighting across voxels, treating ambiguous NETC boundaries and heterogeneous RC regions as equally informative as well-defined ET. This mismatch between supervision and annotation reliability can undermine performance in clinically important subregions.

We address this mismatch with a reliability-weighted supervision framework that estimates per-class reliability from epistemic uncertainty, measured by inter-fold disagreement of a cross-validation ensemble, and down-weights classes with high disagreement. Unlike aleatoric uncertainty, which measures input-level ambiguity, fold-based epistemic uncertainty identifies classes where independently trained models disagree[6], which we hypothesize reflects annotation uncertainty in regions such as the NETC-SNFH boundary and RC interior. This approach distinguishes metastasis segmentation from settings such as pediatric tumors, where performance is often limited by extreme data scarcity[10, 12]. Rather than relying on standard class-frequency balancing, which may amplify localized annotation noise, our framework uses task-specific reliability to address this ambiguity during optimization.

Our framework requires no architectural changes and operates entirely within the training pipeline, with reliability weights computed once from baseline fold-inference outputs at negligible additional cost. Experiments on the BraTS 2026 brain metastasis dataset demonstrate consistent improvement over vanilla nnU-Net, with the largest gains in NETC and RC.

2 Methods

Our framework augments the standard nnU-Net training pipeline with a reliability-adapted supervision mechanism that scales each class's loss contribution according to epistemic uncertainty estimated from cross-validation fold ensemble disagreement, with the resulting per-class scalar weight applied uniformly to all voxels of that class within a case. The pipeline is illustrated in Figure 1. No modifications are made to the network architecture, preprocessing, or inference procedure. The

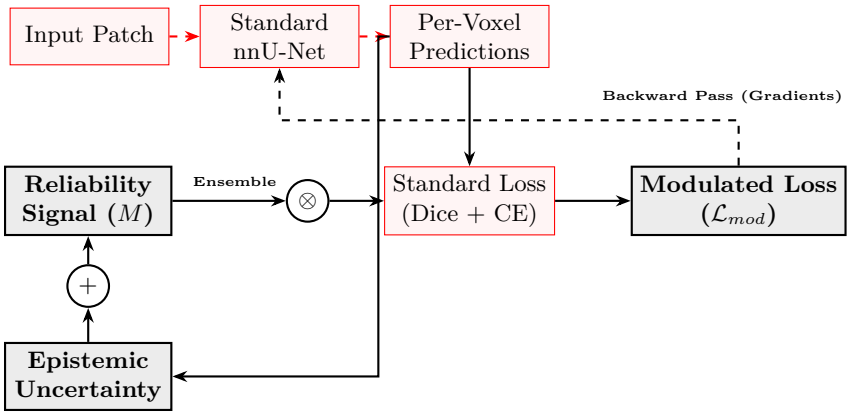

Fig. 1. The proposed training pipeline framework. Components from the standard nnU-Net architecture, preprocessing, and inference are in pink (dashed lines). The core contribution is strictly contained within the active training supervision phase (gray boxes), modulating standard loss with epistemic uncertainty.

reliability weights are computed once prior to training from baseline fold inference outputs and remain fixed throughout.

2.1 Baseline and Notation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the training set of multiparametric MRI volumes and segmentation labels, where $y_i \in \{0, 1, \ldots, C\}$ with $C = 4$ foreground classes (ET, NETC, SNFH, RC) and class 0 denoting background. The standard nnU-Net training objective combines soft Dice and cross-entropy loss with uniform voxel contribution:

$$\mathcal{L}_{\text{uniform}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}$$

Our framework replaces this with a reliability-weighted objective in which each voxel's loss contribution is scaled by its estimated supervision reliability.

2.2 Epistemic Uncertainty Estimation

Following nnU-Net's 5-fold cross-validation, five independent models are trained on non-overlapping 80/20 data splits. Full-dataset inference is performed with each fold model, producing five softmax probability maps per training case. Each case is scored by all five fold models, including the model(s) trained on that case; this yields a denser ensemble signal than restricting to out-of-fold predictions only, at the cost of the resulting variance being a biased, in-sample heuristic for reliability rather than a purely out-of-sample epistemic uncertainty estimate. To assess the impact of this choice on our internal fold-0 ablation, we also computed an out-of-fold-only variance (excluding the model trained on each fold's validation

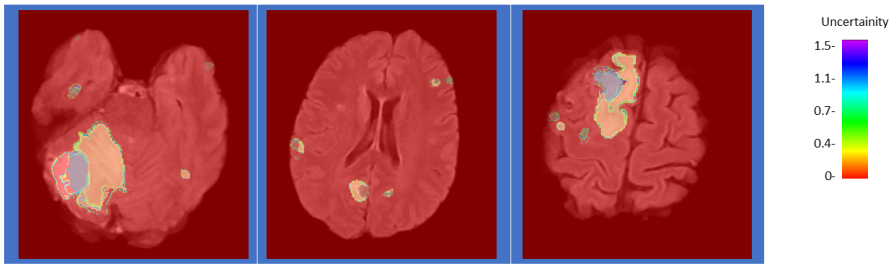

Fig. 2. Spatial correspondence between high inter-fold variance for metastasis cases is boundary-concentrated rather than the blind failure mode.

cases). Results with out-of-fold weights (mean Dice 0.5879) were consistent with the in-sample variant (0.6001), confirming that our conclusions are robust to this modeling choice.

For case i, let $\{p_i^{(k)}\}_{k=1}^{5}$ denote the per-fold softmax outputs, $p_i^{(k)} \in \mathbb{R}^{(C+1)\times D\times H\times W}$. The per-class epistemic uncertainty map is the inter-fold variance:

$$u_i^c = \mathrm{Var}_k\left[p_i^{(k),c}\right], \quad c \in \{0,\ldots,C\}$$

A brain mask derived from the union of foreground predictions suppresses spurious variance outside the brain:

$$\mathcal{M}_i = \left\{ v : \max_k \sum_{c>0} p_i^{(k),c}(v) > 0.05 \right\}$$

Figure 2 illustrates the spatial correspondence between high inter-fold variance and the clinically ambiguous subregions NETC and RC in representative metastasis cases. In contrast to the pediatric setting where CC and ED exhibit near-zero variance despite near-zero Dice - a blind failure mode driven by underrepresentation - metastasis subregion variance is instead concentrated along the NETC-SNFH interface and within the heterogeneous interior of RC, consistent with (though not a direct measurement of) genuine boundary annotation ambiguity in these regions.

2.3 Reliability Weight Computation

Absent class detection RC and other subregions are absent in a substantial proportion of metastasis cases. Rather than inferring absence from variance, we use the available ground truth labels directly: for each case i and class c, if the case contains no voxels of that class ($\sum_v y_i^c(v) = 0$), the class is flagged as absent. All present classes receive computed reliability weights; absent classes are assigned a minimum weight $\epsilon = 0.05$ to prevent gradient contributions from spurious predictions.

Per-case normalization Mean variance per present class is normalized to $[0, 1]$ across present classes:

$$\tilde{u}_i^c = \frac{\bar{u}_i^c - \min_{c'} \bar{u}_i^{c'}}{\max_{c'} \bar{u}_i^{c'} - \min_{c'} \bar{u}_i^{c'}}$$

where $\bar{u}_i^c$ is the mean variance for class c within $\mathcal{M}_i$.

Epistemic weight The reliability weight inverts normalized uncertainty:

$$w_i^c = 1 - \tilde{u}_i^c$$

All weights are clipped to $[\epsilon, 1.0]$ with $\epsilon = 0.05$.

Note that no frequency correction term is applied here. Ablation experiments show that the specific inverse class frequency correction we tested does not improve, and marginally reduces, performance in the metastasis setting (Section 3.3) - a result we interpret as consistent with metastasis subregion difficulty stemming more from appearance variability and annotation ambiguity than from class underrepresentation, though we did not test alternative frequency-correction formulations. Based on these ablations, we use epistemic weighting alone for the metastasis configuration.

2.4 Aleatoric Uncertainty Estimation

We evaluate several reliability formulations. **Epistemic** weights use inter-fold variance as described in Section 2.3. **Aleatoric** weights instead use per-voxel predictive entropy:

$$H_i(v) = -\sum_{c=0}^{C} \bar{p}_i^c(v) \log \bar{p}_i^c(v)$$

averaged per-class within the brain mask and normalized identically to epistemic weights. **Frequency correction** scales weights by inverse class frequency N/N_c. **Spatial** variants apply per-voxel weights directly without class-level averaging.

2.5 Reliability-Weighted Loss

The scalar weight w_i^c is broadcast over all voxels of class c within each training patch. The weighted loss is:

$$\mathcal{L}_{\text{rel}} = \frac{1}{|\mathcal{P}|} \sum_{v \in \mathcal{P}} w_i^{c(v)} \cdot \mathcal{L}_{\text{CE}}^{(v)} + \sum_{c=1}^{C} \left(1 - \frac{2 \sum_v w_i^c \cdot p^c(v) \cdot y^c(v) + \epsilon}{\sum_v w_i^c \cdot (p^c(v) + y^c(v)) + \epsilon}\right) \cdot \frac{1}{C}$$

where $\mathcal{P}$ is the current training patch, $c(v)$ is the ground truth class of voxel v, $p^c(v)$ is the predicted probability for class c, and $y^c(v)$ is the one-hot ground

truth. The loss operates within nnU-Net's deep supervision framework with weights broadcast identically across all resolution scales. We note that because w_i^c is constant across all voxels of class c within a case, it factors out of both the numerator and denominator sums of the weighted Dice term, leaving w_i^c's influence on the Dice ratio confined to interaction with the additive smoothing constant ϵ. Since ϵ is small relative to the intersection and cardinality sums over a typical training patch, the weighted Dice term's sensitivity to w_i^c is correspondingly small. The cross-entropy term, in which $w_i^{c(v)}$ multiplies each voxel's loss directly without a comparable cancellation, is therefore likely the primary channel through which reliability weighting affects training. An ablation isolating the two terms confirms this: training with weighted CE alone recovers 86.39% of the full E2a improvement, while weighted Dice alone recovers 12.58%.

Tie behavior: If $\max_{c'} \bar{u}_i^{c'} = \min_{c'} \bar{u}_i^{c'}$, all present classes have equal variance and the normalization denominator is zero. We set $\tilde{u}_i^c = 0$ and $w_i^c = 1$ for all present classes, assigning equal reliability when variance provides no basis for differentiation.

Normalization limitation: Per-case normalization discards absolute uncertainty magnitude, so uniformly low- and high-variance cases can receive identical relative weights. We use this scheme because variance magnitudes are not directly comparable across cases with different lesion sizes and burdens; globally calibrated normalization is left to future work.

Class-absence detection: For computing training weights, we use ground truth labels directly: if $\sum_v y_i^c(v) = 0$, the class is assigned $\epsilon = 0.05$.

For test-time generalization, where ground truth is unavailable, we employ the variance threshold $\tau = 10^{-6}$ to infer absence. This avoids train/test mismatch and ensures the same procedure applies to unlabeled cases. On the training set, the threshold achieved 93.46% agreement with ground-truth presence, with 4.33% false-positive and 2.21% false-negative flags.

2.6 Implementation Details

All experiments use nnU-Net v2 [5] with the 3d_fullres configuration and default hyperparameters. Training runs for 1000 epochs with a polynomial learning rate schedule and batch size 2. Reliability weights are computed once from 5-fold baseline inference outputs prior to training and remain fixed. The custom trainer subclasses nnUNetTrainer with only the loss function and training step overridden. Experiments were conducted on NVIDIA A100 GPUs. The training scripts are open-sourced at the following link: https://github.com/zephyr-9598/BraTS-2026-metastasis.

3 Results

3.1 Dataset

The BraTS 2026 metastases segmentation dataset comprises multiparametric MRI scans including T1, T1-contrast enhanced, T2, and FLAIR sequences. Each case is annotated with four tumor subregions: enhancing tumor (ET), non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), and resection cavity (RC), corresponding to label indices 1 through 4 respectively. All experiments use the official training split with 5-fold cross-validation. Official challenge evaluation follows lesion-wise metrics, including Lesion-wise Dice (LW Dice), Lesion-wise Normalized Surface Distance at 1.0mm tolerance (LW NSD@1.0), and instance-level F1 score, aggregated across relevant subregion combinations.

3.2 Baseline Performance

Table 1 reports per-class Dice scores across all ablation configurations evaluated on fold 0 of the 5-fold cross-validation split;we report fold 0 only for the detailed ablation study to keep the table concise, and we confirm the main E1 vs E2a comparison across all 5 folds in Table 2. The improvement pattern is consistent across folds, with E2a outperforming baseline on all classes and all folds. The vanilla nnU-Net baseline (E1) achieves a mean Dice of 0.5569 with notable performance degradation on NETC (0.4974) and RC (0.2668), consistent with standard uniform supervision being insufficient for the most challenging metastasis subregions. Inter-fold variance analysis reveals that NETC and RC exhibit the highest ensemble disagreement among all subregions, with variance concentrated along the NETC-SNFH boundary and within the heterogeneous interior of RC-regions where annotation ambiguity is clinically recognized, though we do not directly measure inter-rater agreement here. In contrast, ET and SNFH achieve comparatively strong performance (0.7372 and 0.7261 respectively), consistent with their more distinct imaging characteristics.

3.3 Ablation Study

Effect of epistemic weighting (E2a) Applying reliability weights derived purely from inter-fold epistemic uncertainty improves mean Dice from 0.5569 to 0.6001. Improvements are observed across all foreground classes, with the most pronounced gains in NETC (+0.0623) and RC (+0.0914). As noted in Section 2.5, the weighted Dice term's sensitivity to w_i^c is structurally limited by cancellation, so this improvement is likely driven predominantly by the weighted cross-entropy term rather than by reliability-weighted overlap optimization. An ablation isolating weighted CE from weighted Dice supports this: CE-only weighting achieves a mean Dice of **0.5942**, recovering 86.39% of the full E2a improvement, while Dice-only weighting achieves **0.5623** (12.58% recovery), consistent with the cross-entropy term driving the majority of the observed

Table 1. Per-class DSC across ablation configurations on fold 0 validation.

	Mean Validation DSC			
Method	**NETC**	**SNFH**	**ET**	**RC**
Vanilla nnUNet (E1)	0.4974	0.7261	0.7372	0.2668
Pure Epistemic (E2a)	**0.5597**	**0.7449**	**0.7374**	**0.3582**
Epistemic + freq. weighing (E2b)	0.5481	0.7185	0.7203	0.1888
Pure Epistemic-Spatial (E3a)	0.5610	0.7214	0.7098	0.2240
Epistemic-Spatial + freq. weighing (E3b)	0.5751	0.7114	0.7202	0.2330
Pure Aleatoric (E4a)	0.4825	0.7001	0.7129	0.2238
Aleatoric + freq. weighing (E4b)	0.5074	0.7224	0.7163	0.2334

Table 2. Cross-fold mean validation DSC ± standard deviation (5 folds).

Method	**NETC**	**SNFH**	**ET**	**RC**
Vanilla nnUNet (E1)	0.493 ± 0.026	0.724 ± 0.018	0.735 ± 0.015	0.261 ± 0.034
Pure Epistemic (E2a)	**0.556 ± 0.023**	0.743 ± 0.019	0.736 ± 0.016	**0.352 ± 0.029**

gain. Notably, the largest improvements occur precisely in NETC and RC-the subregions with highest inter-fold ensemble disagreement-consistent with our hypothesis that epistemic uncertainty identifies clinically relevant ambiguity in these subregions.

Effect of frequency correction (E2b) Adding inverse class frequency correction to the epistemic weights yields a mean Dice of 0.5439, representing a degradation from pure epistemic weighting. Performance drops across all classes, with RC experiencing the most substantial decline (0.1888 vs 0.3582 in E2a). We attribute this difference to the nature of metastasis subregion failure: RC and NETC difficulty stems from appearance variability and annotation ambiguity rather than representation deficit. RC is absent in a substantial proportion of cases, but when present, its heterogeneous appearance generates high epistemic uncertainty-the frequency correction term inadvertently up-weights spurious predictions in cases where RC genuinely should be absent or where the annotation is most unreliable. Pure epistemic weighting selectively identifies the regions where supervision should be trusted, whereas frequency correction introduces a signal that conflicts with the reliability characterization.

Effect of aleatoric uncertainty (E4a, E4b) Replacing epistemic uncertainty with aleatoric uncertainty estimated from mean prediction entropy yields a mean Dice of 0.5298, a decrease from the baseline (0.5569). Adding frequency correction (E4b) partially recovers performance to 0.5449, though this remains below both the baseline and the pure epistemic configuration (E2a, 0.6001). We attribute the underperformance of aleatoric weighting to the broader spatial extent of entropy-based uncertainty, which fires on background voxels near brain boundaries and intensity artifacts in addition to genuine tumor subregion ambiguity-introducing

noise into the supervision weight signal that epistemic fold disagreement avoids by construction. The entropy-based weights are less selective for the specific boundary ambiguity that characterizes NETC and RC, resulting in attenuated supervision across a wider range of voxels including those with reliable annotations.

Scalar vs spatial weighting Spatial reliability weight maps, which apply per-voxel rather than per-class scalar weights within each training patch, underperform their scalar counterparts across all configurations (Table 1). Spatial epistemic weighting achieves 0.5541 versus 0.6001 for scalar epistemic, and spatial epistemic + frequency achieves 0.5599 versus 0.5439 for the scalar equivalent. We hypothesize this may relate to within-patch weight heterogeneity: when a training patch contains voxels with highly variable reliability weights, gradient magnitudes within the patch could become inconsistent, destabilizing optimization relative to the uniform per-class scaling applied by scalar weights; however, we did not directly measure within-patch gradient variance, and this explanation remains a hypothesis rather than a confirmed mechanism. Empirically, under our patch-based 3D training setup, class-level reliability abstraction captures the supervision signal more effectively than the spatially-resolved alternative we tested.

3.4 Final Results

Tables 2 and 3 report results on the official BraTS 2026 Metastases challenge validation set, obtained via blind submission of predictions to the challenge leaderboard (distinct from the internal 5-fold cross-validation split used for the fold-0 ablation in Table 1). The independent test set has not yet been released at the time of writing; test-set evaluation will occur via the challenge's docker submission phase once available. Following the challenge protocol, evaluation on the validation leaderboard is performed using lesion-wise metrics: LW Dice, LW NSD@1.0, and instance-level F1 score. The official aggregated label categories include Enhancing Tumor (ET, label 3), Resection Cavity (RC, label 4), the combined non-enhancing/edema/enhancing core (NETC + SNFH + ET), and the full four-class union (NETC + SNFH + ET + RC).

Table 3. Official BraTS 2026 validation results for nnUNet (lesion-wise metrics); F1 (all instance) is reported as a supplementary measure.

Label	LW Dice	LW NSD@1.0	F1 (small instance)	F1 (all instance)
ET (3)	0.6947	0.7557	0.4472	0.8004
RC (4)	0.5225	0.4334	0.0000	0.4566
NETC + SNFH + ET	0.7234	0.7730	0.4574	0.8144
NETC + SNFH + ET + RC	0.6911	0.7148	0.3625	0.8046

Compared to the vanilla nnU-Net baseline, our custom model (pure epistemic weighting, E2a) achieves consistent improvements across all aggregated categories

Table 4. Official BraTS 2026 validation results for our custom model (lesion-wise metrics); F1 (all instance) is reported as a supplementary measure.

Label	LW Dice	LW NSD@1.0	F1 (small instance)	F1 (all instance)
ET (3)	0.7247	0.7857	0.4598	0.8134
RC (4)	0.5825	0.4874	0.0825	0.5366
NETC + SNFH + ET	0.7454	0.7810	0.5024	0.8335
NETC + SNFH + ET + RC	0.7021	0.7348	0.4115	0.8236

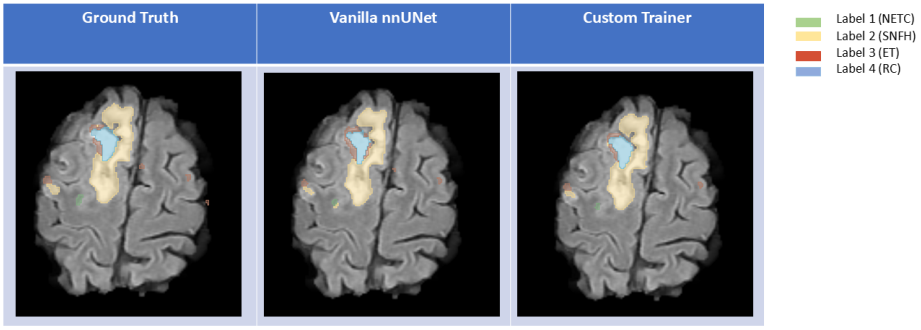

Fig. 3. Qualitative comparison: vanilla nnU-Net vs E2a on representative metastasis case, axial view, showing overall segmentation improvement.

on LW Dice and LW NSD, the primary official ranking metrics. RC, the most challenging structure with the highest inter-fold variance, shows a substantial gain in LW Dice from 0.5225 to 0.5825 (+11.5%) and in NSD (+12.5%). However, on the official ranking F1 (small instance) metric, RC shows very marginal improvement; on the supplementary all-instance F1 metric, RC improves by +17.5%. ET also improves from 0.6947 to 0.7247 (+4.3%) in LW Dice. The aggregated core region (NETC+SNFH+ET) improves from 0.7234 to 0.7454 (+3.0%), while the complete four-class union improves from 0.6911 to 0.7021 (+1.6%).

These gains are corroborated by the NSD and small-instance F1 metrics, which reflect improved boundary delineation and instance-wise detection under the official ranking protocol, respectively. The disproportionate improvement in RC-the subregion with the highest annotation ambiguity and post-surgical appearance variability-is consistent with epistemic down-weighting addressing the heterogeneity challenge unique to brain metastasis cases, without compromising performance on the better-performing subregions.

Figure 3 shows a representative case illustrating the qualitative improvement in NETC and RC segmentation under our framework. The vanilla nnU-Net prediction exhibits fragmented NETC boundaries and incomplete RC delineation-consistent with the high inter-fold variance identified in these subregions. The reliability-weighted model recovers the non-enhancing tumor core with substantially improved boundary continuity and the resection cavity with more complete

coverage, demonstrating that calibrated supervision weighting translates to meaningful segmentation improvement in the most clinically challenging subregions.

4 Discussion

This study shows that, for brain metastasis segmentation, **epistemic reliability weighting alone** is sufficient to improve nnU-Net supervision without architectural changes. Our framework achieves a +11.5% improvement in LW Dice for RC and a +4.3% improvement for ET on the official BraTS 2026 Metastases validation set, with consistent gains observed across the aggregated subregion metrics.The gains concentrate in subregions where fold-level disagreement is highest, supporting the hypothesis that inter-fold variance captures clinically meaningful annotation uncertainty.

A key contrast with our pediatric findings is that **frequency correction is not consistently beneficial** in metastasis. In pediatric data, frequency correction complements epistemic weighting by rescuing blind failures in underrepresented classes. In brain metastasis, where class sparsity is less extreme and annotation consistency is generally higher, that same correction can overemphasize class priors that are already adequately learned. The result is that epistemic-only weighting provides the cleanest and most robust optimization signal.

These findings suggest that reliability weighting should be **data-regime aware**: pediatric cohorts benefit from combined uncertainty and representation correction, whereas adult cohorts may prefer a simpler uncertainty-only formulation. Practically, this reduces hyperparameter sensitivity and simplifies deployment, since no additional balancing between epistemic and frequency terms is required.

Limitations remain. Uncertainty maps are computed once from cross-validation ensembles and kept fixed during training. Iterative uncertainty refresh or online reliability estimation may further improve calibration. Future work should also evaluate epistemic-only supervision across external adult cohorts to confirm robustness under scanner, institution, and protocol shift.

Disclosure of Interests

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

Acknowledgements

We thank God for His continuous guidance and blessings throughout this work. We also gratefully acknowledge the National Health Research Institutes (NHRI), Taiwan, and Dr. Maxim Solovchuk for providing the computational resources and support that made this research possible.

References

1. Bakas, S., Linguraru, M., Kazerooni, A.F., Farahani, K., Astaraki, M., et al.: BraTS 2026 Cluster of Challenges: (Apr 2026). https://doi.org/10.5281/ZENODO.19714728, https://zenodo.org/doi/10.5281/zenodo.19714728
2. Bonato, B., Nanni, L., Bertoldo, A.: Advancing Precision: A Comprehensive Review of MRI Segmentation Datasets from BraTS Challenges (2012–2025). Sensors **25**(6), 1838 (Mar 2025). https://doi.org/10.3390/s25061838, https://www.mdpi.com/1424-8220/25/6/1838
3. Cariola, A., Sibilano, E., Guerriero, A., Bevilacqua, V., Brunetti, A.: Deep learning strategies for semantic segmentation of pediatric brain tumors in multiparametric MRI. Scientific Reports **15**(1), 22595 (Jul 2025). https://doi.org/10.1038/s41598-025-07257-2
4. Huang, Y., Chen, L., Zhou, C.: Multi-Modal Brain Tumor Segmentation via 3D Multi-Scale Self-attention and Cross-attention (Apr 2025). https://doi.org/10.48550/arXiv.2504.09088, http://arxiv.org/abs/2504.09088, arXiv:2504.09088 [eess.IV]
5. 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 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, https://www.nature.com/articles/s41592-020-01008-z
6. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf
7. Kirscher, T., Kahl, K.C., Kovacs, B., Rokuss, M.R., Maier-Hein, K., Coubez, X., Meyer, P., Faisan, S.: Rethinking Post-Hoc Calibration in Semantic Segmentation (Jul 2026). https://doi.org/10.48550/ARXIV.2607.01902, https://arxiv.org/abs/2607.01902, version Number: 1
8. Nayak, L., Lee, E.Q., Wen, P.Y.: Epidemiology of Brain Metastases. Current Oncology Reports **14**(1), 48–54 (Feb 2012). https://doi.org/10.1007/s11912-011-0203-y, http://link.springer.com/10.1007/s11912-011-0203-y
9. Parker, M., Jiang, K., Rincon-Torroella, J., Materi, J., et al.: Epidemiological trends, prognostic factors, and survival outcomes of synchronous brain metastases from 2015 to 2019: a population-based study. Neuro-Oncology Advances **5**(1), vdad015 (2023). https://doi.org/10.1093/noajnl/vdad015
10. Riera-Marín, M., López, J.G., Rodríguez-Comas, J., Ballester, M.A.G., Galdran, A.: Multi-Rater Calibrated Segmentation Models (May 2026). https://doi.org/10.48550/arXiv.2605.02437, http://arxiv.org/abs/2605.02437, arXiv:2605.02437 [cs.CV]
11. Suh, J.H., Kotecha, R., Chao, S.T., Ahluwalia, M.S., Sahgal, A., Chang, E.L.: Current approaches to the management of brain metastases. Nature Reviews Clinical Oncology **17**(5), 279–299 (May 2020). https://doi.org/10.1038/s41571-019-0320-3, https://www.nature.com/articles/s41571-019-0320-3
12. Zarinabad, N., Wilson, M., Gill, S.K., Manias, K.A., Davies, N.P., Peet, A.C.: Multiclass imbalance learning: Improving classification of pediatric brain tumors from magnetic resonance spectroscopy. Magnetic Resonance in Medicine **77**(6), 2114–2124 (Jun 2017). https://doi.org/10.1002/mrm.26318