Anatomy-Preserving Rare-Subtype Adaptation for Pediatric Brain Tumor Segmentation

Hari Mohan Rai[1] and Dauren Omarbekov[1]

Nazarbayev University, Astana, Kazakhstan
hari.rai@nu.edu.kz, dauren.omarbekov@nu.edu.kz

Abstract. Pediatric brain tumor subregion segmentation is strongly imbalanced: whole-tumor anatomy is comparatively stable, whereas enhancing tumor (ET), cystic component (CC), and edema (ED) may be absent or small. We present an anatomy-preserving system for BraTS-PEDs 2026 Task 2. Two independent five-fold residual-encoder nnU-Net ensembles supply tumor anatomy, one defining the submitted segmentation and one whose tumor core localizes an enhancing-tumor specialist, and the specialist reaches the output only through a constrained transplant: it may relabel non-enhancing tumor core (NETC) as ET and nothing else, the whole case is rolled back once a proposal exceeds 800 voxels, and a terminal tumor-core guard closes the assembly, so tumor core, whole tumor and the cystic mask are invariant by construction. On the 91-case challenge validation set the frozen submission achieved Dice 0.588, 0.917, 0.204, 0.000, 0.944 and 0.944 for ET, NETC, CC, ED, tumor core and whole tumor. We also report a negative result: four learned rare-subtype specialists, each with a large ground-truth oracle ceiling, transferred almost nothing under strict held-out evaluation, and importing a challenge winner degraded CC. Pulled back from the registry by digest, the submitted image reprocessed all 91 cases and differs from the frozen reference in ten of them by twelve voxels in total, leaving presence, connected components and instance topology unchanged. Code, container recipe and the verification record are released publicly.

Keywords: Pediatric brain tumor · MRI segmentation · nnU-Net · rare subtypes · anatomy-preserving refinement · reproducible containers

1 Introduction

The BraTS series has benchmarked brain tumor segmentation since its adult glioma editions [12, 1] and now extends to pediatric populations, where tumor biology, subregion prevalence, and cohort size all differ [8, 7]. The strongest entries of the first pediatric edition were ensembles around nnU-Net [9, 3], itself an encoder–decoder U-Net [13] whose residual-encoder presets remain highly effective for gross tumor anatomy [5], and that is our starting point.

Pediatric subregions are hard because the classes differ greatly in prevalence, volume, appearance, and clinical meaning, and a flat multiclass objective is dom-

inated by background and large tumor. In our labeled cohort, at a 10-voxel presence threshold, NETC was present in 291 of 293 cases and ET in 196, whereas CC and ED were present in only 94 and 64. A model can therefore score well on tumor core (TC) and whole tumor (WT) while suppressing genuine rare tissue or hallucinating it where there is none. Our error analysis showed that most ET and CC errors were not failures to find tumor at all: rare-subtype voxels were absorbed into NETC. That motivated a narrower task, preserving the reliable anatomy and deciding only whether selected NETC voxels should be promoted.

We first tried the obvious solution, training a corrector on the errors of the frozen anatomical models, and it did not work: four rare-subtype specialists with large ground-truth oracle ceilings failed to transfer on held-out folds, as did a winning model of the 2025 pediatric edition (Sec. 3.1). The submitted system follows from that. Two independent residual-encoder ensembles are trained, one defining the submitted anatomy and one localizing the specialist, so a specialist error cannot move the anatomical decision; the transplant is restricted to the single transition NETC→ET and bounded by an inclusive whole-case budget, leaving TC, WT, and the cystic mask invariant by construction; and the submission carries an executable verification chain of hash-locked weights and operators, synthetic threshold tests, and a full replay of the pulled image, so the artifact that is scored is provably the artifact that was validated. Sec. S1 of the supplementary material summarizes it as a schematic and as pseudocode. Code, container recipe, and the verification record are at `https://github.com/1nqi/hmnunet-brats2026-task2`.

2 Materials and Methods

2.1 Data and Label Contract

We used 293 labeled training cases and 91 unlabeled challenge validation cases, under one fixed five-fold split with 234 to 235 training and 58 to 59 held-out cases per fold, so every case was held out exactly once. Each case contained T1-weighted native (T1N), contrast-enhanced T1-weighted (T1C), T2-weighted (T2W), and T2-FLAIR (T2F) MRI [8, 7]. In accordance with the challenge rules, we also cite the MedPerf federated benchmarking framework [6]. No challenge validation labels were used for training.

The official integer labels were ET=1, NETC=2, CC=3 and ED=4, with composite masks TC = $\{1, 2, 3\}$ and WT = $\{1, 2, 3, 4\}$. This contract was enforced by executable tests before training and packaging.

2.2 Frozen Anatomical Ensembles

Both ensembles are five-fold 3D full-resolution nnU-Nets with residual encoders, trained with region-based supervision on the four sigmoid channels WT, TC, ET and CC, the formulation that has proved effective in earlier BraTS editions [4]. The *primary* ensemble P combines an extra-large residual encoder at 500

epochs, an extra-large one at 250 epochs with foreground oversampling raised to 0.5, and a large one at 250 epochs, each contributing all five folds with mirroring test-time augmentation [16]; the three probability volumes are averaged in single precision, decoded with a strict threshold and post-processed once (Sec. 2.3). The *hybrid* ensemble H replaces the 500-epoch model by its 250-epoch counterpart and is otherwise identical. It is never submitted: its tumor core defines the specialist's region of interest, and it is the fallback of the terminal guard. Using two ensembles keeps the anatomical decision and the specialist trigger separate, since a proposal is admitted only where both predict NETC.

Both ensembles use the 3D full-resolution configuration at $1 \times 1 \times 1$ mm with per-channel z-scoring. The extra-large encoder trains on $160 \times 256 \times 224$ patches at batch size 3, the large one on $128 \times 224 \times 224$ at batch size 2; both have six stages, 32 to 320 features and 1, 3, 4, 6, 6, 6 residual blocks per stage. Optimization is SGD, Nesterov momentum 0.99, weight decay 3×10^{-5}, initial rate 10^{-2}, polynomial decay $\eta_t = \eta_0(1 - t/T)^{0.9}$ over $T = 250$ or 500 epochs of 250 iterations. The spatial pipeline is rotation up to 30° per axis and scaling in $[0.7, 1.4]$, each at $p = 0.2$, then mirroring on all three axes, with elastic deformation disabled; intensity augmentation adds Gaussian noise ($p = 0.1$), blur ($p = 0.2$), brightness and contrast in $[0.75, 1.25]$ ($p = 0.15$ each), simulated low resolution ($p = 0.25$), and gamma in $[0.7, 1.5]$ with ($p = 0.1$) and without ($p = 0.3$) inversion. This is the nnU-Net 2.8.1 schedule, which our trainers change only in epochs and foreground oversampling.

Class imbalance is handled before the loss rather than inside it. Region-based supervision gives each of WT, TC, ET, and CC its own sigmoid channel with a Dice and binary cross-entropy objective under deep supervision, so a rare subregion is its own binary target rather than a minority class in a softmax. Patch sampling then centers a fixed fraction of every batch on a foreground voxel: 0.33 for the 250- and 500-epoch configurations, 0.5 for the oversampled member of each ensemble and for the specialist, which also sees only tumor-core crops. We did not reweight the objective; Sec. 3.1 reports what happened when we attacked the imbalance with learned correctors instead.

2.3 Region Decoding and Post-processing

Let $p \in [0,1]^{4\times X\times Y\times Z}$ be the averaged region probabilities in channel order $[\mathrm{WT}, \mathrm{TC}, \mathrm{ET}, \mathrm{CC}]$. Labels are assigned by strict thresholding in a fixed paint order, so that a more specific region overwrites a more general one:

$$\hat{y}(x) = \begin{cases} 3, & p_{\mathrm{CC}}(x) > 0.5, \\ 1, & p_{\mathrm{ET}}(x) > 0.5,\ p_{\mathrm{CC}}(x) \le 0.5, \\ 2, & p_{\mathrm{TC}}(x) > 0.5,\ p_{\mathrm{ET}}(x) \le 0.5,\ p_{\mathrm{CC}}(x) \le 0.5, \\ 4, & p_{\mathrm{WT}}(x) > 0.5,\ p_{\mathrm{TC}}(x) \le 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The threshold is strict, so a probability of exactly 0.5 leaves the voxel outside that region, and the paint order is the sequence in which the four region masks are

written onto the label volume, each overwriting whatever the previous one left. A single post-processing pass follows: predicted edema is removed, because ED was not recovered reliably on development data and its presence only degraded whole tumor; ET components below 100 voxels and NETC components below 50 are deleted; and cystic components are retiered by size, removed below 10, relabeled NETC from 10 to 49, and retained as CC at 50 or more. All component analysis uses 26-connectivity. The pass runs exactly once per segmentation and its boundaries are pinned by the synthetic tests of Sec. 3.5.

2.4 Enhancing-Tumor Specialist on the Tumor-Core ROI

Rare-subtype errors are concentrated inside tissue that the anatomical ensembles already detect, so the specialist is confined to that tissue. Let T be the tumor core of the decoded hybrid ensemble before post-processing. The specialist operates on the axis-aligned bounding box of T, dilated by a margin of 16 voxels in each direction and clipped to the image; cases with empty T skip the specialist stage entirely.

Inside this crop we build the six-channel specialist input from the four MRI channels together with two channels derived from the hybrid prediction, and apply a five-fold nnU-Net with a medium residual encoder: five encoder stages with 32 to 320 features, $80 \times 80 \times 80$ patches, and batch size 10, trained with the same 250-epoch schedule, augmentation pipeline, and foreground oversampling rate of 0.5 as the base models. Unlike the ensembles, it is supervised on the three tumor-core labels directly rather than on regions.

The two derived channels are the hard tumor core of the raw hybrid decode and its crop-local signed distance transform $\mathrm{edt}(T) - \mathrm{edt}(\bar{T})$ clipped to ± 20 voxels; neither is z-scored, the four MRI channels enter unmodified apart from the crop, and there are no inter-modality contrast channels. Both are computed inside the container per case after the hybrid decode, so extraction is dynamic rather than a pre-processing pass. Training crops used the same construction on out-of-fold hybrid predictions rather than ground-truth tumor core, so the specialist meets the region-of-interest distribution it was trained on, and its folds inherit the ensembles' out-of-fold assignment, so no case shapes the region of interest of a fold it is scored in. The specialist prediction is pasted back into full image space under identity geometry; the paste is rejected if the recovered array does not match the crop shape and geometry exactly.

The specialist output is combined with the raw hybrid segmentation by a single rule: a voxel the hybrid labels NETC and the specialist labels ET or CC takes the specialist label, every other voxel is unchanged, and the fused volume is post-processed once, yielding F.

2.5 Budgeted Transplant and Terminal Guard

The submitted segmentation is derived from the primary ensemble P; the fused volume F may only propose. The proposal is the set of voxels on which all three

sources agree that a promotion is warranted:

$$\Omega = \big\{x : \; H(x) = \text{NETC} \;\wedge\; F(x) = \text{ET} \;\wedge\; P(x) = \text{NETC}\big\}. \tag{2}$$

The proposal is admitted for the case as a whole, not voxel by voxel:

$$\hat{Y}(x) = \begin{cases} \text{ET}, & x \in \Omega \;\wedge\; |\Omega| \le 800, \\ P(x), & \text{otherwise}, \end{cases} \tag{3}$$

so a case whose proposal exceeds the inclusive budget of 800 voxels is rolled back bit-exactly to P. The budget encodes a conservative prior: plausible missed enhancement in this cohort is small and focal, and a large proposal signals that the specialist disagrees with the anatomy and is no longer refining it.

A terminal guard covers the degenerate case in which the primary ensemble finds no tumor core at all while the hybrid ensemble does; the hybrid segmentation is then emitted unchanged. Outside that guard, the assembly satisfies three invariants, which the container asserts at run time for every case: only NETC→ET transitions occur; the TC and WT masks are identical to those of P; and the cystic mask is identical to that of P. Edema cannot be created, because post-processing removes it and the transplant writes only ET.

Three mechanisms, and no learned confidence score, separate a genuine promotion from noise. Selection is by agreement rather than by a threshold on one model's probability, which is why Eq. 2 conditions on H, F, and P jointly. Magnitude is bounded by the whole-case budget. Spatial coherence comes earlier: F passes the same component filters as any other prediction before the intersection is formed, though the intersection itself is not size-filtered. The three thresholds involved, the strict 0.5 decode boundary, the component sizes and the 800-voxel budget, were fixed on development data and are pinned by the synthetic tests of Sec. 3.5.

Nothing balances the frozen anatomy against the specialist during optimization, because the two are never trained together: there is no joint objective in which the anatomy could dominate the corrector's gradient, and no distillation term. The specialist is fitted on its own crops, the ensembles stay frozen, and the two meet only at assembly, where the arbitration is hard rather than learned, one admissible transition under a bounded budget with a terminal guard, which is what makes the invariants provable instead of empirical. Each case ends with the invariant assertions, a geometry restore, and an atomic write of an unsigned 8-bit volume with labels in $\{0, 1, 2, 3\}$. The complete per-case procedure is given as pseudocode in Sec. S1 of the supplementary material.

2.6 Container Construction and Verification

The submitted image is built on a frozen parent, installs no packages, and is gated at build time on the locked dependency versions, the hashes of the trainers and of the single module the inference operators are imported from, an exact manifest of the 25 checkpoints, and the absence of validation, probability, or

oracle artifacts in the image. Parent and child are referenced by immutable digests, so the verified container cannot be silently replaced by a rebuild. The full gate list is in Sec. S2 of the supplementary material, and gates, manifests and digests are released with the code.

2.7 Evaluation

Out-of-fold evaluation pooled all 293 training cases, each scored by a model that had not trained on it. Challenge numbers come from the official server, which scores with an instance-aware framework [10] and reports global DSC and NSD per region for this task; those are therefore the only official metrics we quote, and no validation ground truth is available to us.

3 Results

3.1 Rare-Subtype Specialists: A Negative Result

Severe class imbalance in segmentation is usually attacked at the level of the objective, with reweighted or overlap-based losses such as focal [11], Tversky [14], or generalized Dice [15]. Our anatomical ensembles already use the standard nnU-Net objective, and the residual errors we care about are subtype confusions inside a tumor that has been detected correctly, so we attacked the problem one level up, at the fusion rule.

Before settling on the constrained transplant we trained four dedicated rare-subtype models on out-of-fold errors of the frozen ensembles, each targeting a class that the anatomical models tend to absorb into NETC: a cystic detector built from NETC context, a direct fusion variant, a representation-pretrained variant, and an ET-residual model. Every design was screened before training by a ground-truth oracle that measured how much rare tissue the fusion rule could reach if the corrector were perfect, and each oracle was encouraging. Every design was then evaluated on strictly held-out folds with full-cohort denominators, under acceptance thresholds fixed in advance.

None of the four passed. The failure was the same in each case: the oracle ceiling was large, the learned recovery on held-out data was indistinguishable from zero, and discrimination collapsed between folds, so a variant that looked promising on one fold was uninformative on the next. We also imported a winning model of the 2025 pediatric edition, as a packaged container and not as a reimplementation. Used conservatively, restricted to the frozen tumor core, it stayed well below the acceptance threshold and what little gain it produced came almost entirely from two cases; used without that restriction it reduced CC substantially.

The pattern is worth stating plainly. An oracle ceiling measures how much rare tissue is *reachable* when the answer is already known; it says nothing about how much is *predictable* from the imaging evidence the model actually sees. For rare pediatric subregions the two quantities came apart almost completely

Table 1. Internal five-fold validation of the trained components, pooled over folds: 293 cases for the ensemble members and, for the specialist, the 258 pre-treatment timepoints whose out-of-fold tumor core is non-empty. Mean per-case Dice follows nnU-Net's own convention and excludes cases in which the region is absent from both reference and prediction; pooled Dice is $2\sum \mathrm{TP}/(2\sum \mathrm{TP} + \sum \mathrm{FP} + \sum \mathrm{FN})$ over the cohort. These are component ensembles, not the assembled pipeline, and neither block is the challenge metric.

Metric	Component	ET	NETC	CC	TC	WT
Mean per-case Dice	Primary member XL_{500}	.509	n/a	.176	.855	.874
	Shared member XL_{os5}	.502	n/a	.184	.846	.868
	Shared member L	.503	n/a	.196	.853	.872
	ET specialist, TC ROI	.579	.795	.203	n/a	n/a
Pooled Dice	Primary member XL_{500}	.738	n/a	.411	.852	.881
	Shared member XL_{os5}	.729	n/a	.446	.851	.881
	Shared member L	.735	n/a	.460	.853	.885
	ET specialist, TC ROI	.737	.815	.430	n/a	n/a

in our experiments, and the screening step we had trusted to select promising designs turned out to select nothing. The submitted system follows from that: a small, budgeted, invariant-preserving transplant driven by agreement between independently trained ensembles, with no learned corrector at all.

3.2 Components on the Training Partition

Table 1 reports the internal five-fold validation of the four trained components on the training partition, pooled across folds so that every case is scored by a model that did not train on it. Three qualifications apply. These are the individual components, not the assembled output, which was never run over the training partition at all; the metric is nnU-Net's internal Dice over the regions each model is supervised on rather than the instance-aware metric the challenge computes, so this table and Table 2 are not directly comparable, and NETC and ED have no channel of their own in the region-based ensembles; and normalized surface Dice does not appear, because the internal evaluator computes Dice and IoU only. The split is at case level, so where a patient contributes more than one timepoint those timepoints may fall in different folds; that bears on this table alone, since the official results come from a disjoint 91-case cohort and Sec. 3.3 remains the only end-to-end measurement of the submitted system.

3.3 Challenge Validation

The frozen Docker candidate's official 91-case results are the bold row of Table 2: DSC .588, .917, .204, .000, .944, .944 and NSD .473, .652, .185, .000, .670, .671 for

ET, NETC, CC, ED, TC, and WT. TC, WT and the cystic mask are invariant by construction; NETC also scored high but is the one region the transplant may change, since a promotion takes voxels out of it; ET was moderate, CC remained difficult, and ED was not recovered because the post-processing of Sec. 2.3 removes it.

The transplant fires rarely, by design, and its effect is small and focal. Of the 91 cases, 29 produced a non-empty proposal totalling 6944 voxels in the frozen reference run; 27 were admitted, adding 4057 ET voxels, or 3.0% of the 134,818 ET voxels submitted. Two were rejected in full, at 1726 and 1161 voxels, and the terminal guard fired once. Since the transplant can only relabel NETC as ET, the ET voxels gained equal the proposal size, checked directly from the two mask sets rather than from the container's own accounting. Per case the median admitted proposal is 2.2% of that case's final ET volume but reaches 54.1% where the primary ensemble found almost no enhancement, so the budget bounds a correction absolutely, not relatively. One case per regime, on the T1C image, is shown in Sec. S6 of the supplementary material.

3.4 Ablation on the Official Validation Set

Table 2 compares six of our own submissions, all scored by the official server on the same 91 cases, so the comparison rests on no local evaluation choice and no ground truth of ours. The anatomical model accounts for most of the difference: the 500-epoch configuration, the only change between the first two rows, raises NSD by 0.007 for NETC, TC, and WT, whereas everything the rare-subtype stage adds afterwards is an order of magnitude smaller. The invariants are visible in the table rather than merely asserted, since rows two to four differ only in the transplant and their CC, TC, and WT entries are identical in both metrics; the official scorer thus confirms, on data we never had access to, the guarantees the assembly rule is supposed to provide. Row five, the restricted prior-winner fusion, is the only variant above the submitted system on any DSC axis, by 0.002 CC against an acceptance threshold fixed in advance at 0.010, with 95.5% of that gain from two cases, which we read as noise rather than transfer; unrestricted, the same model collapses CC to 0.076, and although its NSD rises slightly for NETC, TC and WT, it is 0.343 worse summed over the twelve official axes.

3.5 Verification, Reproducibility, and Deployment

Reproducibility was measured on the artifact that will be scored: the submitted image, pulled from the registry by digest, re-ran its build gate and the synthetic tests that pin every decision boundary of Sec. 2.3 and of the assembly (Sec. S3 of the supplementary material), then reprocessed all 91 cases, of which ten differed from the frozen reference by twelve voxels in total, with presence, component counts and instance topology unchanged in every region and a minimum per-case agreement Dice of 0.99960. Replaying the frozen probability volumes inside the image reproduces all 91 reference segmentations exactly, so the post-processing

Table 2. Ablation on the official challenge validation set (91 cases). Every row is one of our own submissions, scored by the official server. The submitted system is in bold. Rows two to four differ only in the enhancing-tumor transplant, and their CC, TC, and WT entries are identical, which is the assembly invariant confirmed externally.

Metric	Variant	ET	NETC	CC	ED	TC	WT
Global DSC	Hybrid ensemble (XL_{250}+XL_{os5}+L)	.584	.916	.203	.000	.943	.943
	Primary ensemble (XL_{500}+XL_{os5}+L)	.586	.916	.204	.000	.944	.944
	+ ET transplant, no budget	.588	.917	.204	.000	.944	.944
	+ 800-voxel budget (submitted)	**.588**	**.917**	**.204**	**.000**	**.944**	**.944**
	+ prior-winner fusion, restricted	.588	.917	.206	.000	.944	.944
	+ prior-winner fusion, unrestricted	.536	.903	.076	.000	.930	.930
Global NSD	Hybrid ensemble (XL_{250}+XL_{os5}+L)	.467	.645	.183	.000	.663	.664
	Primary ensemble (XL_{500}+XL_{os5}+L)	.471	.652	.185	.000	.670	.671
	+ ET transplant, no budget	.472	.652	.185	.000	.670	.671
	+ 800-voxel budget (submitted)	**.473**	**.652**	**.185**	**.000**	**.670**	**.671**
	+ prior-winner fusion, restricted	.473	.652	.187	.000	.670	.671
	+ prior-winner fusion, unrestricted	.448	.663	.050	.000	.684	.685

arithmetic is bit-exact and the residue is GPU convolution non-determinism at the decision boundary; we report a measured voxel count and claim no determinism. Two further replay measurements, and the measured resource usage against the Task 2 limits, are in Secs. S4 and S5 of the supplementary material.

4 Discussion

The central result is not that a specialist network outperformed nnU-Net. For rare pediatric subregions the reliable gain came from limiting what a correction may do rather than from making the corrector stronger: four learned correctors with large oracle ceilings transferred nothing, whereas a transplant confined to NETC→ET, to voxels two independently trained ensembles agree on, and to a bounded budget, improved ET and left TC, WT, and CC untouched (Sec. 3.4).

Editing the enhancing-tumor label in post-processing is established practice in pediatric BraTS, but our transition runs opposite to the usual one. The winning BraTS-PEDs 2023 entry, a different system from the 2025 model discussed in Sec. 3.1, relabels ET as necrosis or edema whenever predicted ET falls below 4% of the whole-tumor volume, on the grounds that ET is absent from the reference in many pediatric cases [2]; our transplant promotes NETC to ET, because the errors we measured were ET absorbed into NETC rather than ET hallucinated. The two are complementary rather than competing, theirs suppressing ET a model over-calls and ours recovering ET a model under-calls, and they differ in kind: their criterion is relative to whole tumor, ours an absolute per-case budget, which is why our correction is bounded in voxels yet unbounded as a share of a case's ET volume. Relabeling ET as edema also moves a voxel out of

the tumor core, so it cannot leave TC and WT invariant, whereas NETC→ET does.

A domain gap remains: CC stays at 0.204 Dice and ED is not recovered at all, since prevalence and phenotype mixture differ between our development split and the validation cohort, and discarding predicted edema, the right deployment choice here, caps ED at zero.

Because the pipeline is assembled from frozen artifacts, most of the practical risk lay in provenance rather than modeling: a wrong checkpoint or a stale operator yields a plausible but different system, and neither announces itself in the scores. Hash-locked manifests and build-time gates turn such mistakes into build failures.

Four limitations remain: CC and especially ED are data-limited; the budget and component thresholds are constants fitted on development data; the transplant promotes only tissue already labeled NETC, so it cannot recover rare tissue predicted as background or edema; and two five-fold ensembles plus a specialist are expensive. Phenotype-stratified folds, calibrated presence estimation, dedicated edema modeling and distillation are the natural next steps, provided the invariants survive them.

5 Conclusion

Two frozen nnU-Net ensembles provide tumor anatomy and a specialist trigger, a budgeted transplant promotes NETC to ET only where both ensembles and the specialist agree, and a terminal guard covers the degenerate case. Restricting the transition graph and bounding the whole-case budget make TC, WT and cystic preservation explicit rather than empirical, at Dice 0.588, 0.917, 0.204, 0.000, 0.944 and 0.944 on the challenge validation set. Alongside it we report a negative result: a large ground-truth oracle ceiling did not translate into learnable transfer in any of four specialist designs, nor in an imported challenge winner, so oracle screening should not be trusted to select rare-subtype correctors. CC and ED remain the next target.

Acknowledgements

Data were obtained through the BraTS 2026 Challenge, Synapse ID syn74274097. We thank the BraTS organizing committee, the clinical consortia, the annotators and ISSAI, Nazarbayev University for the computing resources.

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

Declaration on Generative AI

The authors used generative AI assistants for code assistance and language refinement, verified the manuscript against the implementation and the official

results, and take full responsibility for its content, claims, references and software. No generative-AI system is listed as an author.

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021). https://doi.org/10.48550/arXiv.2107.02314
2. Capellán-Martín, D., Jiang, Z., Parida, A., Liu, X., Lam, V., Nisar, H., Tapp, A., Elsharkawi, S., Ledesma-Carbayo, M.J., Anwar, S.M., Linguraru, M.G.: Model ensemble for brain tumor segmentation in magnetic resonance imaging. In: Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation (BraTS 2023, CrossMoDA 2023), Lecture Notes in Computer Science, vol. 14669. pp. 221–232. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-76163-8_20
3. 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
4. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnU-Net for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2020), Lecture Notes in Computer Science, vol. 12659. pp. 118–132. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72087-2_11
5. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jäger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI 2024), Lecture Notes in Computer Science, vol. 15009. pp. 488–498. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_47
6. Karargyris, A., Umeton, R., Sheller, M.J., Aristizabal, A., George, J., Wuest, A., Pati, S., Kassem, H., Zenk, M., Baid, U., 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
7. Kazerooni, A.F., Khalili, N., Liu, X., Gandhi, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., et al.: The brain tumor segmentation in pediatrics (BraTS-PEDs) challenge: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv preprint arXiv:2404.15009 (2024). https://doi.org/10.48550/arXiv.2404.15009
8. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., et al.: The brain tumor segmentation (BraTS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv preprint arXiv:2305.17033 (2023). https://doi.org/10.48550/arXiv.2305.17033
9. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., et al.: BraTS-PEDs: Results of the multi-consortium international pediatric brain tumor segmentation challenge 2023. arXiv preprint arXiv:2407.08855 (2024). https://doi.org/10.48550/arXiv.2407.08855

10. Kofler, F., Möller, H., Buchner, J.A., de la Rosa, E., Ezhov, I., Rosier, M., Mekki, I., Shit, S., Negwer, M., Al-Maskari, R., et al.: Panoptica – instance-wise evaluation of 3D semantic and instance segmentation maps. arXiv preprint arXiv:2312.02608 (2023). https://doi.org/10.48550/arXiv.2312.02608
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007 (2017). https://doi.org/10.1109/ICCV.2017.324
12. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., 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
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Lecture Notes in Computer Science, vol. 9351. pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: Machine Learning in Medical Imaging (MLMI 2017), Lecture Notes in Computer Science, vol. 10541. pp. 379–387. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_44
15. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA 2017), Lecture Notes in Computer Science, vol. 10553. pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
16. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019). https://doi.org/10.1016/j.neucom.2019.01.103