Hierarchical nnU-Net Logits for Pediatric Brain Tumor Subregion Segmentation

Haozhou Han[0009−0009−0206−1302], Qian Wang[0009−0003−0416−8118], and Haoyan Ding[0009−0000−4070−5732]

Abstract. Accurate pediatric brain tumor segmentation requires reliable delineation of small and heterogeneous subregions, including enhancing tumor (ET), non-enhancing tumor (NET), and cystic component (CC), that are often underrepresented in voxel-wise training objectives. We study three nnU-Net-based strategies for the MICCAI 2026 Brain Tumor Segmentation in Pediatrics (BraTS-PED) Task 2 under a common preprocessing, inference, and postprocessing protocol: a strong 3D full-resolution nnU-Net baseline, scale-aware deep supervision, and a hierarchy-aware output parameterization. The hierarchy-aware model factorizes the BraTS-PED label structure into whole-tumor, tumor-core, and core-subtype predictions, allowing ET, NET, and CC to compete within the tumor-core branch while preserving standard nnU-Net training and export. On the official validation panel, this model improved ET Dice from 0.532 to 0.567 and CC Dice from 0.123 to 0.218 over the baseline, at the cost of small decreases on the large composite regions. We also report training-set performance. These findings suggest that label-structure-aware output modeling is a useful consideration for robust pediatric brain tumor subregion segmentation.

Keywords: Pediatric brain tumor segmentation · BraTS-PED · nnU-Net · Hierarchical segmentation · Deep learning

1 Introduction

Accurate delineation of brain tumors and their subregions is an important component of quantitative neuro-oncological imaging. Automated segmentation can support tumor-burden assessment, radiotherapy and surgical planning, longitudinal treatment-response monitoring, and the extraction of reproducible imaging biomarkers [1,2,3]. However, manual annotation of three-dimensional magnetic resonance imaging (MRI) is time-consuming, requires specialized expertise, and is subject to inter-observer variability, motivating reliable automated segmentation methods.

The Brain Tumor Segmentation (BraTS) Challenge has played a central role in the development and standardized evaluation of such methods by providing multi-parametric MRI data, expert-derived annotations, common evaluation metrics, and held-out evaluation servers [4,5,6,7]. The BraTS-PED task focuses on pediatric brain tumor segmentation from multi-parametric MRI and forms part of the BraTS 2026 challenge cluster [8,9,10]. In this setting, algorithms must

delineate enhancing tumor (ET), non-enhancing tumor (NET), cystic component (CC), and peritumoral edema (ED), from which composite regions are derived: whole tumor (WT) is the union of all four labels, tumor core (TC) is the union of ET, NET, and CC, and non-enhancing tumor core (NETC) is the NET label alone. These subregions differ substantially in size, frequency, contrast appearance, and clinical interpretation, making pediatric tumor segmentation more than a binary tumor-detection problem.

nnU-Net is a natural starting point for this challenge because it provides a self-configuring and highly competitive framework for biomedical image segmentation [11]. It has also been successful in prior BraTS settings, where BraTS-specific nnU-Net adaptations achieved strong challenge performance [12]. More recent pediatric BraTS work has shown that targeted nnU-Net modifications can further improve difficult pediatric tumor subregions [13]. We therefore use nnU-Net both as a strong baseline and as the common framework for evaluating task-specific variants, rather than introducing an entirely separate segmentation pipeline.

Despite this strong baseline, the conventional nnU-Net formulation treats background, ED, ET, NET, and CC as unrelated classes in a flat multi-class output space. This formulation does not reflect the nested structure of the BraTS-PED labels: WT is decomposed into ED and TC, while TC is further decomposed into ET, NET, and CC. Under a flat softmax formulation, a rare CC voxel must compete directly not only with ET and NET, but also with the much more frequent background and edema classes. This may disadvantage rare tumor-core components even when broader tumor extent and tumor core are correctly localized.

In this work, we investigate whether task-specific supervision and output parameterizations can improve pediatric brain tumor subregion segmentation while retaining the reproducibility and simplicity of the nnU-Net pipeline. We compare three independently trained methods under a shared preprocessing, training, inference, and postprocessing protocol: a standard three-dimensional full-resolution nnU-Net baseline, a hierarchy-aware output parameterization, and a scale-aware deep supervision variant. The hierarchy-aware model explicitly factorizes the BraTS-PED label structure by first estimating whole-tumor membership, then separating tumor core from edema, and finally decomposing tumor-core voxels into ET, NET, and CC.

2 Methods

2.1 Baseline nnU-Net

We begin with the standard nnU-Net v2 full-resolution 3D configuration as the baseline, retaining its automatically configured preprocessing, augmentation, optimizer, and inference pipeline. This baseline is not treated as a minimal reference implementation, but as a strong challenge model. In our configuration the network was trained on $96 \times 160 \times 160$ voxel patches with batch size 2.

The baseline optimized the standard compound segmentation objective,

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}, \tag{1}$$

where the Dice term addresses foreground imbalance and the cross-entropy term preserves voxel-wise class calibration. We trained the model on the original five-label formulation rather than on overlapping evaluation regions, so the network predicted background, ET, NET, CC, and ED directly. Sliding-window inference and connected-component postprocessing were then applied identically to all methods.

This baseline is important for two reasons. First, it provides a competitive self-configuring architecture whose performance on WT, TC, and NETC is already high, making the remaining errors concentrated in smaller and more ambiguous subregions. Second, it fixes the data pipeline and inference protocol for all subsequent variants.

2.2 Hierarchy-Aware Output Parameterization

The hierarchy-aware model was the main architectural variant and the final validation submission. Its motivation is that the BraTS-PED labels are not a set of unrelated classes. Whole tumor contains tumor core and edema, while tumor core is further decomposed into ET, NET, and CC. A flat five-way softmax ignores this structure: a rare CC voxel must compete directly with background and edema, although the clinically relevant distinction is first whether the voxel belongs to tumor core and only then which core subtype it represents.

As shown in Fig. 1, to encode this structure, we replaced the flat multi class output layer with a hierarchy-aware parameterization. Instead of predicting five mutually competing class logits directly, each decoder output head estimates a whole tumor probability, a tumor core probability conditioned within tumor, and a three-way core subtype distribution:

$$p_{\text{wt}} = \sigma(z_{\text{wt}}), \qquad p_{\text{tc}} = \sigma(z_{\text{tc}}), \qquad \mathbf{q} = \text{softmax}([z_{\text{ET}}, z_{\text{NET}}, z_{\text{CC}}]). \tag{2}$$

The final class probabilities are composed as

$$\begin{aligned} P(\text{bg}) &= 1 - p_{\text{wt}}, & P(\text{ED}) &= p_{\text{wt}}(1 - p_{\text{tc}}), \\ P(\text{ET}) &= p_{\text{wt}} p_{\text{tc}} q_{\text{ET}}, & P(\text{NET}) &= p_{\text{wt}} p_{\text{tc}} q_{\text{NET}}, \\ P(\text{CC}) &= p_{\text{wt}} p_{\text{tc}} q_{\text{CC}}. \end{aligned} \tag{3}$$

This factorization changes the competition space: ET, NET, and CC compete only within tumor core, while ED competes with tumor core within whole tumor. The output therefore follows the same nested structure as the evaluation regions, but still produces the original five voxel-wise labels.

The composed probabilities sum to one. We therefore return $\log P$ in the standard class-channel order, so the softmax operation used by nnU-Net during training and inference recovers the composed probabilities exactly. This design preserves the default Dice plus cross entropy loss, deep supervision at all decoder

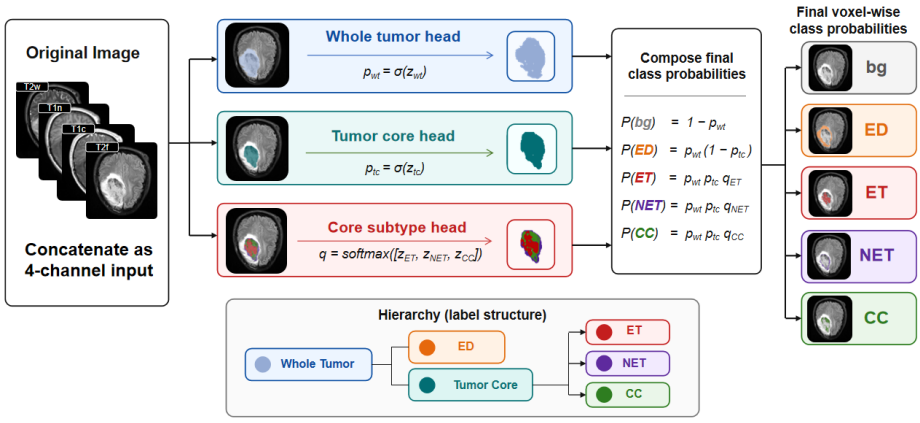

Fig. 1. Hierarchy-aware output parameterization for four-modal MRI segmentation. The four MRI modalities (T2W, T1N, T1C, and T2F) are concatenated as a four-channel input. The decoder feature representation is mapped to three hierarchical output heads that predict whole-tumor probability, tumor-core probability conditioned on whole tumor, and the distribution of tumor-core subtypes. These predictions are then composed into mutually exclusive voxel-wise probabilities for background (bg), ED, ET, NET, and CC. The lower panel summarizes the nested BraTS-PED label structure, in which whole tumor contains edema and tumor core, while tumor core is further divided into ET, NET, and CC.

scales, sliding-window inference, and the challenge export format. The variant is therefore architecture-aware at the output layer, but does not require a cascade, ensemble, auxiliary inference network, or custom submission pipeline.

Two implementation details are worth stating explicitly, because they determine how the hierarchy is supervised. First, the hierarchical head is not restricted to the full-resolution output: every segmentation layer in the decoder, including all deep-supervision layers, is replaced by a hierarchical head. All decoder scales therefore use the same parameterization, and no scale uses a flat five-class output. Second, there are no separate targets for p_{wt}, p_{tc}, or $\mathbf{q}$, and no auxiliary loss terms. The unmodified nnU-Net Dice plus cross-entropy objective of Eq. (1) is applied to the *composed* five-class distribution against the standard five-class one-hot target, independently at each deep-supervision scale with the default nnU-Net scale weights. The three hierarchical quantities are thus learned implicitly through the composed likelihood rather than through explicit region supervision. This is what allows the modification to remain confined to the output layer: the loss, the target construction, and the deep-supervision schedule are all unchanged from the baseline.

Table 1. Class weights used for scale-aware deep supervision. Other foreground classes used weight 1.0 at all decoder scales.

Class	Full resolution	1/2	1/4	1/8 and coarser
CC	1.0	0.7	0.2	0.0
ED	1.0	1.0	0.8	0.5

2.3 Scale-Aware Deep Supervision Variant

As an independent variant, we evaluated whether deep supervision could be adjusted for small structures. Coarse decoder targets are obtained by downsampling the ground truth label map; a small CC region can shrink dramatically or disappear at low resolution. Penalizing the model strongly at those scales may therefore teach the coarse heads that CC is absent. We kept the architecture and inference pipeline fixed, but added per class weights to the deep supervision loss:

$$\mathcal{L}_{\mathrm{DS}} = \sum_{s} \omega_s \left(\mathcal{L}^{(s)}_{\mathrm{CE}}(w_{\cdot,s}) + \mathcal{L}^{(s)}_{\mathrm{Dice}}(w_{\cdot,s}) \right), \tag{4}$$

where ω_s are the standard nnU-Net scale weights and $w_{c,s}$ are class-specific weights at scale s. Other foreground classes retained weight 1.0 at every scale. The CC and ED schedules are summarized in Table 1.

2.4 Training, Prediction, and Postprocessing

The released cohort comprised 294 training cases spanning pre- and post-treatment scans and 91 pre-treatment validation cases, with four co-registered modalities (T1C, T1N, T2F, T2W). Images were used as released without additional brain extraction or registration, and each modality was z-normalized over non-zero voxels. All models were trained with the same full-resolution 3D nnU-Net configuration and the same masked-normalization preprocessing protocol. We used the standard nnU-Net training recipe, including deep supervision, stochastic patch sampling, extensive spatial and intensity augmentation, and the default optimizer and learning-rate schedule. Each of the three methods was trained as an independent model rather than as part of a combined system. This design allowed us to compare the effect of each modification under matched training and inference conditions.

Training was performed using PyTorch 2.6.0 with CUDA-compatible binaries on RTX 4090. The final predictions for the validation cohort were obtained with nnU-Net sliding-window inference. The same inference settings were used for all three methods, and the hierarchy-aware model was selected as the final validation submission based on official validation-panel performance.

Postprocessing was applied to the final exported label maps in native image space after nnU-Net inference. Component volume was computed from the voxel spacing stored in each image header. All training and validation volumes have $1.0 \times 1.0 \times 1.0$ mm isotropic spacing, which also matches the nnU-Net target spacing. Thus, one voxel corresponds to 1 mm^3, and the ET and CC thresholds of

160 mm^3 and 50 mm^3 correspond exactly to 160 and 50 voxels, respectively. Connected components were defined using 26-connectivity, and components strictly smaller than the corresponding threshold were removed.

The ET and CC thresholds were adopted from the published BraTS-PED winning method rather than tuned on the present validation panel, and the same thresholds were fixed for all variants. No hidden validation or test labels were used for threshold selection. As a post hoc sensitivity analysis, we evaluated minimum component volumes of $\{0, 50, 160, 300, 500, 1000\}$ mm^3 on the training cohort. Moderate filtering at 50–160 mm^3 improved training-cohort Dice over no filtering for both ET and CC across all three models. For CC, 50 mm^3 gave the highest Dice for all three variants. For ET, 160 mm^3 was highest for the hierarchy-aware model and nearly tied with 50 mm^3 for scale-aware deep supervision, whereas the baseline favored 50 mm^3 by approximately 0.011 Dice. Accordingly, the sweep is reported as a post hoc sensitivity analysis rather than as the procedure used to select the thresholds.

3 Results

3.1 Experimental Setup

The three variants were trained with nnU-Net fold `all` for 1000 epochs and evaluated on the official BraTS-PED validation panel using DSC and NSD. All reported results use the final-epoch checkpoint. We selected the hierarchy-aware model because it provided the highest mean ET/CC DSC while keeping WT/TC/NETC within 0.01 DSC of the baseline. This criterion was not prespecified before viewing the validation results and reflects the study's emphasis on the weakest tumor-core subregions rather than the official challenge-ranking objective.

3.2 Training-Set Performance

Because all models were trained on fold `all`, every one of the 294 training cases contributed to model fitting. We therefore distinguish optimization diagnostics from full-volume training-cohort performance. Fig. 2 shows the per-epoch pseudo-Dice and training loss logged during optimization, whereas Table 2 reports full-volume Dice obtained from final-checkpoint inference on all 294 training cases. Neither should be interpreted as held-out generalization performance. The full-volume results indicate substantially stronger fit on the training cohort than performance on the official validation panel, including for ED. Because ED is present in only 66 of the 294 training cases, we additionally computed Dice only among ED-positive training cases. Mean ED Dice on these cases was 0.721 for the baseline, 0.753 for scale-aware deep supervision, and 0.770 for the hierarchy-aware model. Thus, the validation-panel ED score, which rounded to 0.000, cannot be explained by an inability of the models to fit ED on the training subjects, although the available validation feedback does not identify the cause of the generalization failure.

Table 2. Full-volume training-cohort Dice for the three models. Values are voxel-aggregated across all 294 training cases. Because fold `all` was used, these values characterize training-set fit rather than held-out generalization.

Method	WT	TC	NETC	ET	CC	ED
Baseline	0.944	0.941	0.917	0.859	0.881	0.862
Scale-aware DS	0.951	0.949	0.927	0.873	0.891	0.880
Hierarchy-aware	0.951	0.949	0.927	0.873	0.892	0.880

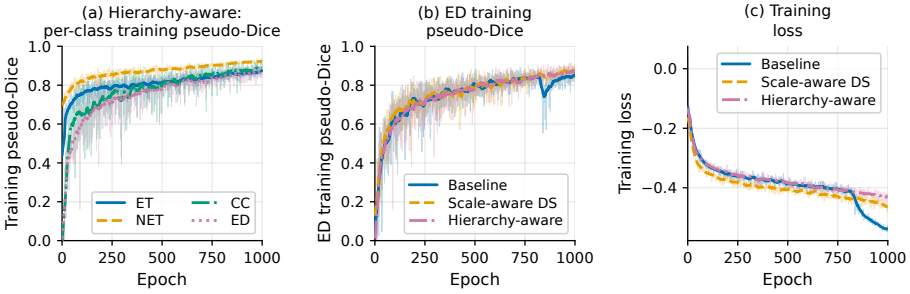

Fig. 2. Training optimization over 1000 epochs. Faint lines show per-epoch values and bold lines show 25-epoch moving averages. (a) Per-class training pseudo-Dice for the hierarchy-aware model. (b) ED training pseudo-Dice for the baseline, scale-aware deep-supervision, and hierarchy-aware models. (c) Training loss for the three models.

The baseline exhibited a transient late-training decrease in pseudo-Dice, particularly for CC and ED, despite a continuous 1000-epoch run and an uninterrupted learning-rate schedule; its performance recovered toward the end of training. The other two variants did not show a comparable late-epoch fluctuation.

3.3 Official Validation Results

Table 3 reports official validation DSC. Relative to the baseline, the hierarchy-aware model improved ET from 0.532 to 0.567 and CC from 0.123 to 0.218, while scale-aware deep supervision achieved the highest WT, TC, and NETC DSC. ED rounded to 0.000 for all three variants.

The hierarchy-aware model also had the highest unweighted six-region mean DSC (0.592 versus 0.573 and 0.574) and the highest ET/CC mean (0.393 versus 0.328 and 0.324), whereas scale-aware deep supervision led on the WT/TC/NETC mean.

For NSD (Table 4), hierarchy-aware output improved ET from 0.416 to 0.455 and CC from 0.113 to 0.166. Scale-aware deep supervision matched the baseline ET NSD (0.416) and increased CC NSD to 0.138.

Table 3. Official validation-panel DSC by region for the three independently trained methods. Best values are shown in bold.

Method	WT	TC	NETC	ET	CC	ED
Baseline	0.938	0.938	0.909	0.532	0.123	0.000
Scale-aware DS	**0.941**	**0.940**	**0.913**	0.515	0.132	0.000
Hierarchy-aware	0.932	0.932	0.905	**0.567**	**0.218**	0.000

Table 4. Official validation-panel NSD for the three submitted variants. Best values are shown in bold.

Method	WT	TC	NETC	ET	CC	ED
Baseline	**0.658**	**0.658**	**0.642**	0.416	0.113	0.000
Scale-aware DS	0.657	0.656	0.641	0.416	0.138	0.000
Hierarchy-aware	0.653	0.652	0.640	**0.455**	**0.166**	0.000

3.4 Treatment-Status Comparison

Table 5 compares the full 294-case cohort with the 262-case pre-treatment-only sensitivity analysis.

The effect was method- and region-dependent. Because restriction also removed 32 training cases, this analysis does not isolate treatment status, and it was not used for final-model selection.

As a qualitative validation example, Fig. 3 shows a representative validation-slice prediction overlaid on the FLAIR image for the three evaluated methods. The visualization illustrates the same trend as the quantitative results: the hierarchy-aware model produces more visible tumor-core subregion predictions, particularly for ET and CC, whereas the baseline and scale-aware variants emphasize the larger tumor extent more conservatively.

Finally, ED scored below 0.003 for all three submitted variants. Export audit identified no obvious missing-label failure, but ED was predicted very sparsely on the validation cohort. In the training cohort, ED was present in only 66 of 294 cases and consisted predominantly of small components, suggesting that class rarity and component size may contribute to the observed generalization gap. Because validation annotations are unavailable, the cause cannot be determined definitively. The similar failure of the flat-softmax baseline indicates that it is not specific to the hierarchy-aware parameterization.

4 Discussion

The standard nnU-Net baseline is already highly competitive for large composite regions, but the difficult small tumor-core subregions benefit from a more structured output space. This is consistent with the nested organization of the BraTS-PED labels: a flat softmax assigns each voxel through a single unrestricted competition among background, ED, ET, NET, and CC, which is unfavorable for small core classes that must compete with background and edema as well

Table 5. Official validation-panel performance after training on the full 294-case cohort or the pre-treatment-only 262-case subset. Each entry reports DSC / NSD.

Method	Cohort	WT	TC	NETC	ET	CC	ED
Baseline	Full	.938/.658	.938/.658	.909/.642	.532/.416	.123/.113	.000/.000
Baseline	Pre-only	.929/.663	.929/.662	.901/.642	.493/.390	.104/.088	.000/.000
Scale-aware	Full	.941/.657	.940/.656	.913/.641	.515/.416	.132/.138	.000/.000
Scale-aware	Pre-only	.941/.679	.940/.678	.912/.657	.532/.424	.161/.161	.000/.000
Hierarchy-aware	Full	.932/.653	.932/.652	.905/.640	.567/.455	.218/.166	.000/.000
Hierarchy-aware	Pre-only	.932/.675	.932/.674	.904/.650	.529/.429	.180/.176	.000/.000

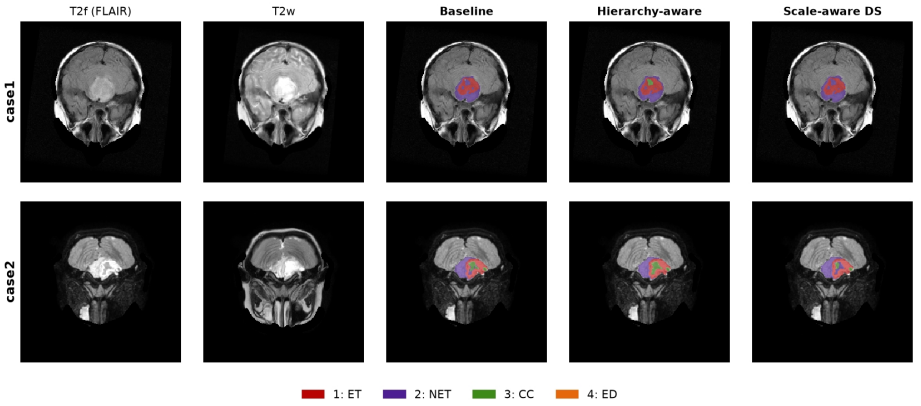

Fig. 3. Qualitative comparison on a selected validation slice. Predictions after post-processing are overlaid on the FLAIR image. Label colors correspond to ET (1) in red, NET (2) in violet, CC (3) in green, and ED (4) in orange.

as with the other core subtypes. Decomposing the prediction into whole-tumor, tumor-core, and core-subtype probabilities restricts ET, NET, and CC competition to the tumor-core branch, preserving mutually exclusive voxel labels while matching the structure of the evaluation regions.

Scale-aware deep supervision behaved complementarily. Down-weighting CC at coarse decoder scales was intended to avoid penalizing the model when small cystic regions vanish under target downsampling, and its strongest effect was on the large regions; it did not improve ET DSC, and its CC gain was smaller than the hierarchy-aware model's. Loss reweighting across scales can therefore stabilize supervision, but appears insufficient when the binding constraint is competition among labels in the output space. The CC gain of $+0.009$ came with an ET DSC loss of -0.017 even though the reweighting never mentions ET, plausibly because removing the coarse-scale penalty on CC lets those voxels be absorbed into the adjacent enhancing rim; we did not verify this with a per-scale diagnostic, and the unchanged ET NSD alongside the lower ET DSC indicates that the difference is not consistent across overlap- and surface-based metrics. Since no single modification improved every region, a hybrid that preserves both effects is a natural next step.

Several limitations remain. Each method was trained once without an explicitly specified random seed, so run-to-run variability cannot be separated from the observed differences between variants. Because fold `all` used all training cases and the official server returned only aggregate results, we could not perform held-out cross-validation, paired statistical testing, or case-level failure analysis. Finally, the postprocessing thresholds were fixed independently of the hierarchy-aware model rather than jointly optimized with it.

Overall, our findings indicate that label-structure-aware output modeling is a promising direction for improving rare pediatric tumor-core subregions within an otherwise standard nnU-Net workflow, and that the improvement is obtained without altering the loss, the training schedule, or the inference pathway. The remaining weakness is the poor ED performance observed on the official validation panel. Its cause cannot be determined from the available aggregate validation feedback, although the rarity and small component size of ED motivate further investigation of sampling and supervision strategies for rare subregions.

Declarations

Funding No funding was received for conducting this study.

Ethics Approval and Consent to Participate Not applicable. This study used de-identified data provided through the BraTS-PED challenge and involved no direct recruitment of human participants by the authors.

Code availability. The implementation is available at `https://github.com/naiQAQian/Brats-2026-Task2`.

Acknowledgments. Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097). We thank the BraTS-PED challenge organizers and data contributors for providing the dataset, evaluation platform, and challenge infrastructure.

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

References

1. Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., Brugnara, G., Schell, M., Kessler, T., Foltyn, M., et al.: Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. The Lancet Oncology **20**(5), 728–740 (2019). `https://doi.org/10.1016/S1470-2045(19)30098-1`
2. Kickingereder, P., Neuberger, U., Bonekamp, D., Piechotta, P.L., Götz, M., Wick, A., Sill, M., Kratz, A., Shinohara, R.T., Jones, D.T.W., et al.: Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. Neuro-Oncology **20**(6), 848–857 (2018). `https://doi.org/10.1093/neuonc/nox188`

3. Kickingereder, P., Götz, M., Muschelli, J., Wick, A., Neuberger, U., Shinohara, R.T., Sill, M., Nowosielski, M., Schlemmer, H.P., Radbruch, A., et al.: Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment response. Clinical Cancer Research **22**(23), 5765–5771 (2016). https://doi.org/10.1158/1078-0432.CCR-16-0702
4. 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
5. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
6. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018). https://doi.org/10.48550/arXiv.1811.02629, https://arxiv.org/abs/1811.02629
7. Karargyris, A., Umeton, R., Sheller, M.J., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence **5**(7), 799–810 (2023). https://doi.org/10.1038/s42256-023-00652-2
8. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., Bagheri, S., Baid, U., Bergquist, T., Borja, A.J., Calabrese, E., Chung, V., Conte, G.M., Dako, F., Eddy, J., Ezhov, I., Familiar, A., Farahani, K., Haldar, S., Iglesias, J.E., Janas, A., Johansen, E., Jones, B.V., Kofler, F., LaBella, D., Lai, H.A., Van Leemput, K., Li, H.B., Maleki, N., McAllister, A.S., Meier, Z., Menze, B., Moawad, A.W., Nandolia, K.K., Pavaine, J., Piraud, M., Poussaint, T., Prabhu, S.P., Reitman, Z., Rodriguez, A., Rudie, J.D., Sanchez-Montano, M., Shaikh, I.S., Shah, L.M., Sheth, N., Shinohara, R.T., Tu, W., Viswanathan, K., Wang, C., Ware, J.B., Wiestler, B., Wiggins, W., Zapaishchykova, A., Aboian, M., Bornhorst, M., de Blank, P., Deutsch, M., Fouladi, M., Hoffman, L., Kann, B., Lazow, M., Mikael, L., Nabavizadeh, A., Packer, R., Resnick, A., Rood, B., Vossough, A., Bakas, S., Linguraru, M.G.: 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, https://arxiv.org/abs/2305.17033
9. Kazerooni, A.F., Khalili, N., Liu, X., Gandhi, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., 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, https://arxiv.org/abs/2404.15009
10. Bakas, S., Linguraru, M., Kazerooni, A.F., Farahani, K., Astaraki, M., Maleki, N., Menze, B., Baid, U., Yordanov, N., Kofler, F., You, S., Chung, V., Mangul, S., Velichko, Y., Gandhi, D., Jiang, Z., Conte, G.M., Moawad, A., LaBella, D., De Verdier, M.C., Aboian, M., Wiestler, B., Huse, J., Huang, R.: BraTS 2026 cluster of challenges. https://doi.org/10.5281/zenodo.19714728 (2026). https://doi.org/10.5281/zenodo.19714728, version v1

11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
12. 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
13. Li, X., Xu, Z.Q.J., Ren, Y., Qiu, T., Wang, X.: Towards reliable pediatric brain tumor segmentation: Task-specific nnU-Net enhancements. arXiv preprint arXiv:2511.00449 (2025). https://doi.org/10.48550/arXiv.2511.00449, https://arxiv.org/abs/2511.00449