This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

An Engineering-optimized, Resource-Constrained nnU-Net v2 Pediatric Baseline for the BraTS-GoAT 2026 Generalizability-Across-Tumors Challenge

Ahad Ali Nazir [1,2][0009-0002-4251-6913] and Absaar Ul Jabbar[1][0000-0003-4438-1418]

[1] School of Interdisciplinary Engineering and Sciences (SINES), National University of Sciences and Technology (NUST), Islamabad, Pakistan

[2] Rehman Medical College, Phase V, Hayatabad, Peshawar, KP, Pakistan

absaar@sines.nust.edu.pk

Abstract. Accurate segmentation of pediatric brain tumors in multiparametric magnetic resonance imaging (mpMRI) is essential for treatment planning and response assessment, yet faces distinctive challenges compared to adult glioma segmentation due to heterogeneous imaging phenotypes and limited pediatric datasets. We present our submission to BraTS-GoAT 2026 Task 3 (Generalizability Across Tumors), targeting, as a single-population baseline within this multi-population challenge, the segmentation of pediatric high-grade gliomas into necrotic core (NCR), peritumoral edema (ED), and enhancing tumor (ET) on the BraTS Pediatric Tumors dataset (1,351 cases; 894 training, 451 held-out test). Our approach leverages nnU-Net v2 in its 3D full-resolution configuration with targeted engineering optimizations: (1) reduction of parallel data-augmentation workers to mitigate memory constraints on resource-limited clusters; (2) disabling PyTorch dynamic compilation to ensure training stability over extended epochs; and (3) ensemble aggregation with test-time augmentation to maximize generalization. Training on four MRI sequences (T1n, T1c, T2w, T2f) on NVIDIA A10G GPUs within strict computational constraints resulted in 5-fold cross-validation Dice scores of 0.824–0.845 (mean 0.831), with strong performance on enhancing tumor (0.861–0.878) and edema (0.849–0.860), while necrotic core remained more variable (0.758–0.795), reflecting the pediatric-specific imaging heterogeneity. The complete deployment pipeline was rigorously validated end-to-end, including containerized inference verification on held-out test cases, ensuring readiness for the challenge evaluation. Held-out test set performance and cross-population generalizability within Task 3 remain to be established through the challenge evaluation and future extension of this work, respectively. On the official BraTS-GoAT 2026 validation leaderboard, our five-fold ensemble with test-time augmentation achieved global Dice Similarity Coefficient (DSC) / Normalized Surface Distance (NSD) scores of 0.769/0.537 for enhancing tumor, 0.814/0.495 for tumor core, and 0.873/0.476 for whole tumor (Submission ID 9773335).

Keywords: Pediatric brain tumor segmentation, nnU-Net, BraTS-GoAT, multiparametric MRI, generalizability.

1 Introduction

Pediatric brain tumors are the most common solid malignancy in childhood and the leading cause of cancer-related mortality in children worldwide [2]. They present a fundamentally different clinical and imaging challenge from adult tumors. The disease spectrum is distinct, covering low-grade gliomas, embryonal tumors, and diffuse midline gliomas, and these show heterogeneous enhancement patterns, variable necrosis, and growth kinetics that frequently diverge from the classic adult glioblastoma phenotype [1]. Pediatric necrotic cores are typically smaller and less consistently ring-enhancing than adult patterns, and pediatric tumors occur far more often in the posterior fossa and brainstem, locations largely absent from adult-dominated training datasets [1,11].

This matters clinically because accurate quantification of enhancing tumor (ET), peritumoral edema (ED), and necrotic core (NCR) informs radiotherapy planning, defines clinical target volumes accounting for potentially infiltrated peritumoral tissue, and guides biopsy site selection. A model trained on adult glioma is not guaranteed to segment these regions correctly in pediatric scans [1]. The BraTS-GoAT (Generalizability Across Tumors) 2026 Challenge was designed to make this generalization gap measurable rather than anecdotal [12]. We report a baseline for the pediatric arm only. We do not evaluate the adult glioma, meningioma, or metastasis arms that make up the rest of Task 3, and we make no claim of demonstrated cross-population generalizability (Sect. 5.3).

This work establishes a fully trained 5-fold nnU-Net v2 baseline for the pediatric arm of Task 3 and documents the engineering interventions required for stable training under strict high-performance computing constraints. Our contributions are: (1) reducing parallel data-augmentation workers to prevent memory overflow on resource-limited clusters; (2) disabling PyTorch dynamic compilation to ensure multi-epoch training stability; and (3) validating the deployment pipeline end to end with ensemble test-time augmentation. That the first two fixes are necessary rather than incidental is confirmed in Sect. 3.5, where the default configuration fails outright on identical hardware. We achieve 5-fold cross-validation Dice of 0.824 to 0.845 (mean 0.831), with consistent performance on ET and ED and higher variance on NCR reflecting pediatric imaging heterogeneity.

2 Related Work

Since 2012, the BraTS challenge series has driven most progress in automated glioma segmentation by providing a shared, expert-annotated, multi-institutional dataset and common evaluation framework [6], though almost all of that progress was validated on a single relatively uniform population: adult glioma [5,6]. As the challenge expanded to meningeal tumors, brain metastases, Sub-Saharan African glioma [18], and pediatric cases, models scoring well on adult glioma did not transfer cleanly to the other populations [10], reflecting differences in lesion morphology, in which sub-regions are present at all, in scanner protocols, and in patient demographics. This limits clinical usefulness most where access to specialist neuro-radiological review is already scarce, which is part of why we focus on the pediatric arm.

We build on nnU-Net, whose main contribution was removing the guesswork from 3D segmentation: preprocessing, normalization, patch and batch sizing, network depth, and training schedule are all derived automatically from a dataset's own statistics rather than tuned by hand for each task [13], and newer architectures are still measured against it [19]. We leave it unmodified by design, since establishing what the framework achieves on its own is the control that allows any later architectural gain, for instance one targeting long-range spatial context, to be attributed to the architecture rather than to more careful configuration. Two practical concerns shape the rest of the work: failures that arise when a full 3D pipeline runs on shared infrastructure are rarely documented, so we set ours out in Sect. 3.5; and the benchmark-to-bedside gap is wider in pediatrics than in adult oncology [3], with containerized benchmarking existing because a Dice score says little about whether a model runs correctly elsewhere [4]. We therefore treat deployment verification as a first-class contribution.

3 Methods

3.1 Dataset, Task Definition, and Clinical Rationale for the Sub-region Labels

We used the BraTS Pediatric Tumors dataset (Dataset ID 503), curated as part of the broader BraTS data infrastructure [7,8,9,11]: 1,351 cases in total, of which 894 are available for training and internal cross-validation and 451 are held out by the organizers for testing. Each case provides four co-registered, skull-stripped MRI sequences as separate input channels: channel 0000, T1-weighted (T1n); channel 0001, post-contrast T1-weighted (T1c); channel 0002, T2-weighted (T2w); and channel 0003, T2-FLAIR (T2f). Voxel-wise labels define three foreground sub-regions (label 1, necrotic core, NCR; label 2, peritumoral edema, ED; label 3, enhancing tumor, ET), with all remaining voxels as background.

These labels are not an arbitrary partition of the tumor. Each corresponds to a decision a clinician actually makes. Enhancing tumor is tracked serially to judge treatment response, since post-contrast enhancement marks blood-brain barrier breakdown where tumor is active. Edema extent sets how far beyond the visible margin a radiation field must reach. Necrotic core is less directly actionable, but it still informs grading and steers biopsy planning away from tissue unlikely to yield a usable diagnosis [1]. A segmentation model is therefore reproducing three separate clinical judgments at once, which is worth bearing in mind when one sub-region proves harder to segment than the other.

3.2 Preprocessing

All preprocessing was handled by nnU-Net v2's self-configuration pipeline, unmodified: dataset fingerprinting, resampling of all four channels to a common target spacing, per-channel z-score normalization, and patch and batch geometry derived automatically from the fingerprint. We applied no additional skull-stripping, bias-field correction, or registration beyond what the released challenge data already includes, and we used no external data to supplement training.

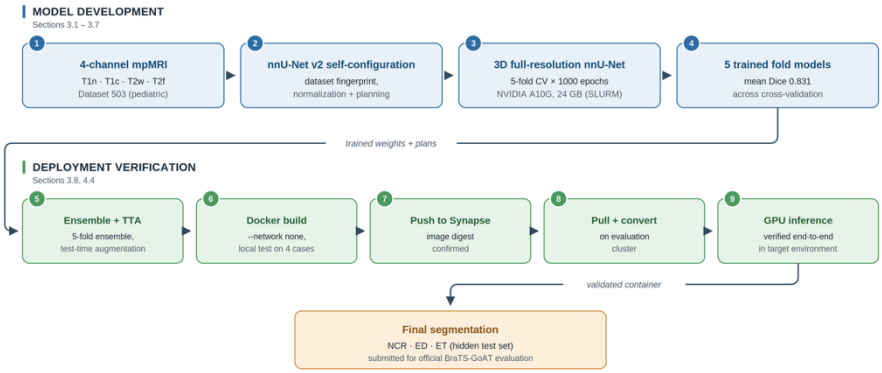

Fig. 1. End-to-end pipeline from multiparametric MRI input through model training to verified containerized deployment. Top: model development (Sect. 3.1–3.7). Middle: deployment verification (Sect. 3.8, 4.4), corresponding to the checklist in the supplementary material. Bottom: final sub-region output on the hidden test set.

3.3 Network Architecture and Loss Function

nnU-Net v2 in its standard 3D full-resolution configuration was utilized with default compound loss and deep supervision without architectural modifications, as detailed in the supplementary material.

3.4 Training Configuration and Hardware

Each of the five folds trained for nnU-Net's standard 1,000 epochs. All training ran on the NUST SINES HPC cluster (School of Interdisciplinary Engineering and Sciences, NUST Islamabad) as SLURM-managed jobs on NVIDIA A10G GPUs with 24 GB of VRAM. A full fold took roughly two and a half days of wall-clock time, close to, but under, the cluster's 72-hour per-job limit. That margin existed only after the fixes described below. Before them, training did not finish at all.

3.5 Critical Engineering Optimizations for Constrained HPC Deployment

Unmodified default configurations failed due to system RAM saturation during data augmentation, confirming the necessity of these adjustments (detailed in supplementary material).

3.6 Inference and Post-processing

Validation-fold inference used nnU-Net's standard sliding-window procedure with Gaussian-weighted patch aggregation, applied to each held-out case within its fold. For the final deployment pipeline, we added test-time augmentation (mirroring across all spatial axes) and multi-fold ensembling to aggregate predictions across all five folds into the final output masks.

3.7 Validation Strategy and Evaluation Metrics

nnU-Net's standard 5-fold cross-validation was used on the 894 training and internal-validation cases, with deterministic, case-level fold splitting so no case appears in both the training and validation partitions of a given fold. For each fold we report mean foreground Dice across all validation cases, along with per-label Dice for NCR, ED, and ET individually, both computed through nnU-Net's built-in evaluation pipeline.

3.8 Containerization and Deployment Verification

A Docker image was built around the trained model and confirmed that it builds cleanly and runs correctly with networking disabled (--network none), matching the isolated execution environment the challenge evaluation is expected to use. The container was tested end to end on four real cases before treating it as working. Inference runtime was measured across these four validation volumes (BraTS-GoAT-00000, -00003, -00004, -00006): 13 s, 7 s, 7 s, and 3 s respectively, averaging approximately 7.5 s per 3D volumetric scan (range 3–13 s, depending on spatial dimensions and tumor burden) on the NVIDIA A10G GPU.

The image was pushed to Synapse, and the resulting digest was confirmed against what was sent; it was then pulled back down and converted successfully on the evaluation cluster. Finally, GPU inference was confirmed on the cluster directly: correct device detection, reasonable runtime, and output matching local results. Each step was checked individually rather than assumed to follow from the one before, since containerized pipelines fail more often at packaging and distribution than at the model itself.

Code, SLURM scripts, and configuration used in this work are publicly available at `https://github.com/ahad21005-cloud/brats-goat-2026-segmentation`.

4 Results

4.1 Training Convergence and Cross-Validation Performance

Training converged reliably across all five cross-validation folds: each completed the full 1000-epoch schedule and internal validation without incident, and none diverged, required early stopping, or needed intervention beyond the configuration described in Sect. 3.

Per-fold mean Dice scores were closely clustered, ranging from 0.824 (Fold 3, the lowest) to 0.845 (Fold 4, the strongest), with Folds 0, 1, and 2 at 0.827, 0.830, and 0.831. Averaged across folds, the resulting five-fold mean Dice of 0.831 constitutes this baseline's headline cross-validation result prior to ensembling; Table 1 reports the full per-fold and per-region performance alongside this mean.

The underlying training dynamics are shown in Fig. 2, which plots training loss, validation loss, and pseudo-Dice for Fold 4 as representative of stable convergence across the full schedule. Training and validation loss track closely together throughout, with no upward drift in validation loss relative to training loss at any point — the standard signature of overfitting is absent. Validation pseudo-Dice rises rapidly over the first ~100 epochs before plateauing in the 0.88–0.90 range. Per-epoch duration held stable

at ~220 s throughout (Sect. 3.5), consistent with the absence of the instability the engineering fixes were designed to prevent.

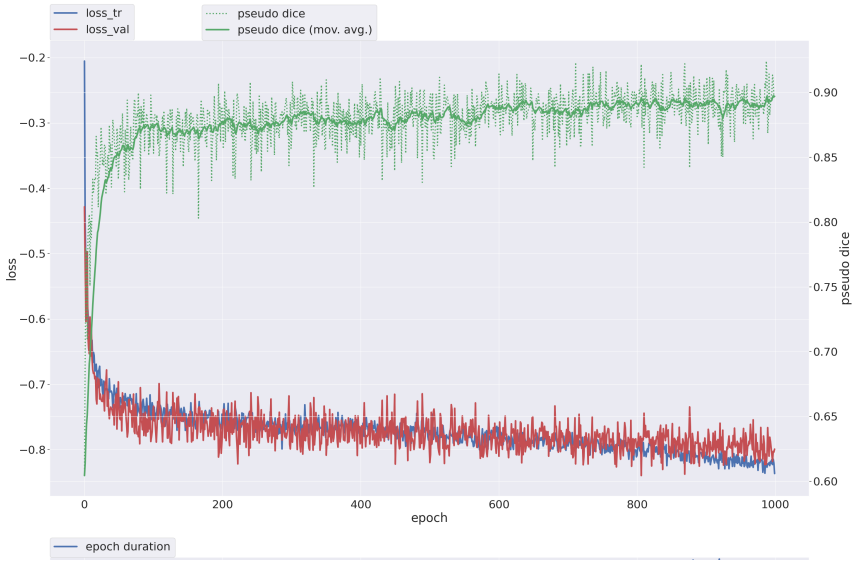

Fig. 2. Training and validation loss and pseudo-Dice over 1000 epochs for Fold 4, showing stable convergence with no divergence or instability.

Table 1. Five-fold cross-validation performance (Dice Similarity Coefficient) of the nnU-Net v2 3D full-resolution baseline on the BraTS Pediatric Tumors dataset (Dataset 503), evaluated across the 894 training/validation cases (the 451 held-out test cases are inaccessible during cross-validation). NCR: necrotic core; ED: peritumoral edema; ET: enhancing tumor. Values are mean ± SD across validation cases within each fold.

Fold	Overall Mean Dice	NCR Dice	ED Dice	ET Dice
0	0.827	0.770 ± 0.286	0.850 ± 0.166	0.862 ± 0.207
1	0.830	0.780 ± 0.291	0.850 ± 0.186	0.861 ± 0.214
2	0.831	0.780 ± 0.278	0.849 ± 0.190	0.864 ± 0.208
3	0.824	0.758 ± 0.300	0.853 ± 0.178	0.862 ± 0.190
4 **(best)**	0.845	0.795 ± 0.276	0.860 ± 0.149	0.878 ± 0.169
Mean (5-fold)	0.831	0.776	0.853	0.866

4.2 Per-Region Performance Analysis

Enhancing tumor and edema are the two most reliably segmented regions, attaining Dice ranges of 0.861–0.878 (mean 0.866) and 0.849–0.860 (mean 0.853), respectively, with negligible fold-to-fold variation. Necrotic exhibited both lower absolute performance and greater variability, ranging from 0.758 to 0.795 (mean 0.776), a markedly wider spread than that observed for either of the other two regions; this behavior is examined further in Sect. 5.1. Notably, the ordering of ET and ED ahead of NCR held without exception across every fold, strongly suggesting that it reflects a structural property of the task rather than an artefact of any individual validation split. The five overall fold scores (0.824, 0.827, 0.830, 0.831, 0.845) cluster tightly around the mean of 0.831. This consistency is expected given that all folds employed an identical configuration, indicating that the residual variation is attributable primarily to the specific cases assigned to each validation split.

4.3 Per-Region Performance Analysis

Enhancing tumor and edema are the two most reliably segmented regions, attaining Dice ranges of 0.861–0.878 (mean 0.866) and 0.849–0.860 (mean 0.853), respectively, with negligible fold-to-fold variation. Necrotic exhibited both lower absolute performance and greater variability, ranging from 0.758 to 0.795 (mean 0.776), a markedly wider spread than that observed for either of the other two regions; this behavior is examined further in Sect. 5.1. Notably, the ordering of ET and ED ahead of NCR held without exception across every fold, strongly suggesting that it reflects a structural property of the task rather than an artefact of any individual validation split. The five overall fold scores (0.824, 0.827, 0.830, 0.831, 0.845) cluster tightly around the mean of 0.831. This consistency is expected given that all folds employed an identical configuration, indicating that the residual variation is attributable primarily to the specific cases assigned to each validation split.

4.4 Qualitative Segmentation Examples

Table 1 summarizes overlap accuracy numerically; Figs. 3, 4, and 5 illustrate this performance on individual cases, all drawn from Fold 4 for consistency with the convergence analysis and training-Dice comparison reported above. Fig. 3 presents a representative case (BraTS-GoAT-00228) with overall Dice close to the 5-fold mean, where predicted NCR, ED, and ET boundaries track the ground truth closely, with only minor disagreement at the NCR/ET interface. Fig. 4 shows the lowest-NCR-Dice case in Fold 4 (BraTS-GoAT-00012), a posterior fossa lesion with a very small ground-truth necrotic core that the prediction omits almost entirely while still correctly capturing the surrounding edema and enhancing tumor — a visual instance of the small-object Dice penalty discussed in Sect. 5.1, where a few dozen misclassified voxels produce a large Dice drop even though the overall lesion is well localized. Fig. 5 displays a peak-performance case (BraTS-GoAT-01138), demonstrating high spatial accuracy across all three sub-regions and near-flawless delineation of the necrotic core.

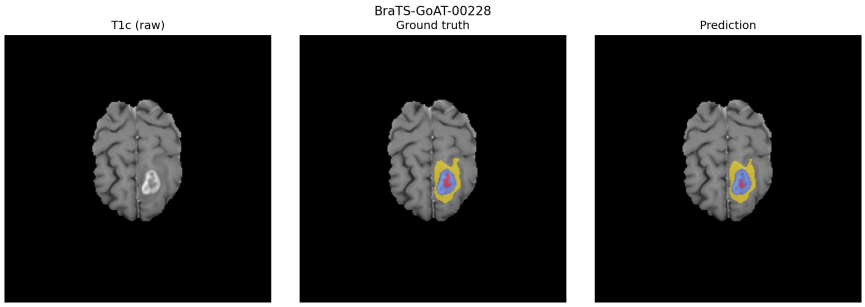

Fig. 3. Representative case (BraTS-GoAT-00228, Fold 4, overall Dice close to the 5-fold mean). Left to right: raw T1c, ground truth, and predicted segmentation (NCR=red, ED=yellow, ET=blue).

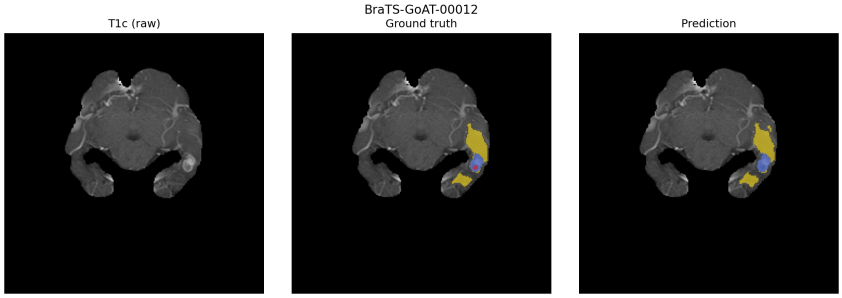

Fig. 4. Lowest-NCR-Dice case in Fold 4 (BraTS-GoAT-00012). The small ground-truth necrotic core (red, center panel) is almost entirely missed by the prediction (right panel), while edema and enhancing tumor are still segmented correctly.

Fig. 5. Peak-performance case in Fold 4 (BraTS-GoAT-01138). The predicted segmentation (right panel) closely mirrors the ground truth (center panel) across all tumors sub-regions, demonstrating highly accurate delineation of the necrotic core (red), edema (yellow), and enhancing tumor (blue).

Table 2. Official BraTS-GoAT 2026 validation leaderboard metrics (Submission ID 9773335) evaluated on the challenge evaluation server. WT: Whole Tumor; TC: Tumor Core; ET: Enhancing Tumor; DSC: Dice Similarity Coefficient; NSD: Normalized Surface Dice.

Region	Dice Similarity Coefficient (DSC)	Normalized Surface Dice (NSD)
Whole Tumor (WT)	0.873	0.476
Tumor Core (TC)	0.814	0.495
Enhancing Tumor (ET)	0.769	0.537

4.5 Deployment Pipeline Verification

Beyond the cross-validation numbers, we treat the deployment pipeline itself as a result worth reporting, since a model that scores well but cannot be packaged and run correctly is not actually a working submission. Sect. 3.8 describes the verification procedure in detail; the full stage-by-stage outcome is summarized in the supplementary material, from the trained 5-fold model through Docker packaging, Synapse push/pull, and GPU inference on the evaluation cluster, all of which were confirmed successful.

4.6 Challenge Leaderboard Validation Results

To evaluate generalization on unseen clinical cases beyond internal cross-validation, the containerized multi-fold ensemble with test-time augmentation was submitted to the BraTS-GoAT 2026 evaluation server. Table 2 reports the official challenge validation leaderboard metrics (Submission ID 9773335): Whole Tumor Dice 0.873 (NSD 0.476), Tumor Core 0.814 (NSD 0.495), and Enhancing Tumor 0.769 (NSD 0.537). These independent server-side metrics confirm the deployment readiness of the containerized pipeline.

5 Discussion

5.1 What the Numbers Mean Clinically

A 5-fold mean Dice of 0.831, with a best fold at 0.845, is a solid result for an unmodified baseline on a population this heterogeneous. For context, an unmodified nnU-Net configuration reported whole-tumor, tumor-core, and enhancing-tumor Dice of 0.89, 0.85, and 0.82 on adult glioma in BraTS 2020, without task-specific tuning [15]. The pediatric numbers here land in a comparable range despite a smaller, more heterogeneous training set and a meaningfully different imaging phenotype, evidence that nnU-Net's strength comes largely from adapting to whatever dataset it is given rather than from assumptions baked in about what a glioma should look like.

The clinical reading of these numbers is arguably more informative than the numbers themselves. Enhancing tumor, at 0.861–0.878, is the best-performing region and also the one most directly tied to how treatment response is judged: in adult and pediatric neuro-oncology alike, serial changes in enhancing tumor volume are the primary imaging signal behind response assessment criteria, making it the number most likely to be used downstream. Edema, at 0.849–0.860, is nearly as strong, and edema extent is what radiation oncologists rely on when defining how far a treatment field must reach beyond the visibly enhancing tumor. Getting both right consistently, across five different validation splits, is a meaningfully different claim than getting a single good average Dice.

Necrotic core, at 0.758–0.795, is the outlier. Unlike the fairly canonical ring-enhancing necrosis of adult glioblastoma, pediatric necrosis is inconsistent in extent, shape, and contrast, and often small and boundary-ambiguous to begin with. Dice is known to penalize small structures harder than large ones for the same absolute boundary error [17], so some of this gap is a property of the metric as much as the model, an explanation that also fits NCR having the widest spread across the five folds, since size-sensitivity makes a metric sensitive to which cases land in a given validation split. BraTS-PEDs 2023 showed a broadly similar pattern, with the smaller, less consistently enhancing sub-regions weakest across top-ranked nnU-Net-based methods, though differing label definitions make direct numerical comparison unmeaningful [16]. Clinically, this is also where segmentation error matters least: necrotic core is not typically a direct radiotherapy target the way enhancing tumor and edema are, so a baseline around 0.78 is a reasonable starting point rather than a red flag.

5.2 Constrained Training and Reproducible Deployment

The engineering fixes in Sect. 3 are a small part of the paper by word count and a large part of it by actual time spent. Multi-day, 1000-epoch training on 24GB GPUs under a strict wall-time cap is a completely ordinary situation for an academic group without a dedicated GPU cluster, and the specific failures we hit, augmentation-related out-of-memory crashes, dynamic-compilation instability, are, in our experience, rarely written down anywhere a new group training nnU-Net would find them before hitting the same wall. We would rather this be one page of a short paper than another few days lost to the same debugging cycle for someone else.

There is also a less technical point worth making here. BraTS-GoAT's inclusion of Sub-Saharan African data specifically reflects an intent to broaden who this kind of research is accessible to, not just which patients it serves [18]. A pipeline that only works on a dedicated, purpose-built GPU cluster quietly excludes exactly the groups that motivation is aimed at including. Writing down concrete, low-cost fixes for common training failures on shared infrastructure is a small thing, but it is a small thing in the right direction.

Reproducible containerization is a critical component of clinical translation, and Sect. 3 and Sect. 4.3 reflect this priority. A model can produce excellent cross-validation Dice and still fail as a submission if the container does not build reproducibly, if the image does not survive being pushed and pulled from a registry, or if GPU inference silently falls back to CPU or produces different output than local testing did. None of that shows up in a Dice table. Verifying each stage separately, build, offline test, push

with digest confirmation, pull and conversion, GPU inference, caught issues that a single end-to-end smoke test would likely have papered over, and reflects a broader push in medical AI toward treating reproducible, independently verifiable deployment as part of the actual contribution, not an implementation detail [3,4].

5.3 Limitations

Our 5-fold cross-validation results (Sect. 4.1) are internal, computed on the 894 training cases; the official validation leaderboard results now reported in Sect. 4.5 provide an external check on a portion of these numbers but are themselves distinct from, and not a substitute for, performance on the 451-case held-out test set the organizers control, which remains undetermined until final challenge evaluation. This baseline addresses only the pediatric arm of Dataset 503. This scope was a direct consequence of available compute and time: training access was limited to single-node A10G GPUs under a 72-hour per-job SLURM ceiling (Sect. 3.4), and a full 5-fold, 1000-epoch cross-validation across all four Task 3 populations was not feasible within the submission deadline. Rather than distribute this budget thinly across four populations, we chose to establish one population baseline with full methodological rigor complete cross-validation, ensembling, and end-to-end deployment verification. Consequently, this work does not, by itself, demonstrate generalizability across the full set of tumor populations Task 3 evaluates; we see it as a necessary first step toward that broader claim, not a substitute for it, and extending evaluation to the remaining populations is stated as future work in Sect. 6.

Training-set Dice (Sect. 4.1, supplementary material) was computed on a 23-case subset rather than the full 894-case training split, due to compute and time constraints; a full-scale training-set evaluation is left for future work. A direct empirical comparison of single-fold versus 5-fold ensemble+TTA performance on local validation data was not completed for this submission, as ensembled prediction outputs were not retained from the Sect. 3.8 deployment verification run. We report ensemble+TTA as the configuration used for final deployment (Sect. 3.6, supplementary material) but have not quantified its performance delta, we leave this comparison for future work.

6 Conclusion and Future Work

A self-configuring nnU-Net v2 baseline has been developed and validated for pediatric brain tumor sub-region segmentation under BraTS-GoAT 2026 Task 3, employing five-fold cross-validation on the NUST SINES HPC cluster. All five folds completed successfully, yielding a five-fold mean Dice of 0.831 with per-region performance that aligns with clinical priorities: strong and consistent results for enhancing tumor and edema, the two regions most directly employed in treatment planning and response monitoring, together with a lower but nonetheless competitive necrotic core score attributable to genuine biological variability rather than to a deficiency in the model. The two engineering modifications that enabled multi-day training to complete on this cluster have been documented, and the deployment pipeline was verified end to end, encompassing image build, offline testing, Synapse push with digest confirmation, pull and conversion, and GPU inference, rather than treating containerization as a secondary

concern. Evaluation on the official BraTS-GoAT validation leaderboard (Submission ID 9773335) yielded Whole Tumor, Tumor Core, and Enhancing Tumor Dice scores of 0.873, 0.814, and 0.769, respectively, confirming deployment readiness on unseen challenge data.

The five trained folds were subsequently combined into a single ensemble incorporating test-time mirroring and finalized within the verified container for evaluation. Future work will explore integrating State-Space Model (SSM) components [14] into the nnU-Net bottleneck. This direction is motivated by a known limitation of convolutional encoders: their effective receptive field is bounded by kernel size and network depth, which constrains their ability to model long-range spatial context and can disproportionately affect small or boundary-ambiguous structures [21]. Mamba-based SSM layers [21,22] have been shown to capture such long-range dependencies in 3D biomedical segmentation at linear computational cost, including within U-Net-style architectures directly comparable to those used here. Given that the necrotic core is both the smallest and most boundary-ambiguous sub-region in this work (Sect. 5.1), and that this class of limitation is architecture-level rather than data-level, SSM integration represents a targeted next step rather than a speculative one; empirical validation is left to future work. Subject to available time and computational resources, this evaluation is additionally intended to be extended to the remaining tumor populations covered by Task 3, such that the claim of generalizability may ultimately be tested directly rather than inferred from performance on a single population.

Acknowledgments. We acknowledge the NUST SINES High-Performance Computing (HPC) cluster for the computational resources used in this work, and the BraTS-GoAT 2026 organizers for curating and distributing the Task 3 dataset and evaluation infrastructure.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Nafe, R., Porto, L., Samp, P.F., You, S.J., Hattingen, E.: Adult-type and Pediatric-type Diffuse Gliomas: What the Neuroradiologist Should Know. Clin. Neuroradiol. 33(3), 611–624 (2023).
2. Su, Z., Lu, J., Shi, Y., Li, T., Qi, B., Guo, Z.: Global, regional, and national childhood brain and central nervous system cancer burden: an analysis based on the Global Burden of Disease Study. Trop. Med. Health. 53(1), 130 (2025).
3. Tozzi, A.E., Fabozzi, F., Eckley, M., Croci, I., Dell'Anna, V.A., Colantonio, E., Mastronuzzi, A.: Gaps and Opportunities of Artificial Intelligence Applications for Pediatric Oncology in European Research: A Systematic Review of Reviews and a Bibliometric Analysis. Front. Oncol. 12, 905770 (2022).
4. Karargyris, A., Umeton, R., Sheller, M.J., Aristizabal, A., George, J., Wuest, A., Pati, S., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nat. Mach. Intell. 5, 799–810 (2023).

5. Baid, U., Ghodasara, S., Mohan, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021).
6. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34(10), 1993–2024 (2015).
7. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data 4, 170117 (2017).
8. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al.: Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017).
9. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al.: Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. The Cancer Imaging Archive (2017).
10. Bakas, S., Reyes, M., Jakab, A., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv:1811.02629 (2018).
11. Kazerooni, A.F., Khalili, N., Liu, X., et al.: The Brain Tumor Segmentation (BraTS) Challenge 2023: focus on pediatric brain tumors (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv:2305.17033 (2023).
12. Baid, U., Conte, G.M., Bakas, S., et al.: BraTS-GoAT 2026: Segmentation Generalizability of Brain Tumor Sub-regions Across Tumors in Pre-operative MRI. BraTS 2026 Challenge Description (2026).
13. 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. Nat. Methods 18, 203–211 (2021).
14. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. arXiv:2312.00752 (2023).
15. Isensee, F., Jaeger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnU-Net for Brain Tumor Segmentation. arXiv:2011.00848 (2020).
16. Kazerooni, A.F., Khalili, N., Familiar, A., et al.: BraTS-PEDs: Results of the Multi-Consortium International Pediatric Brain Tumor Segmentation Challenge 2023. arXiv:2407.08855 (2024).
17. Bertels, J., Robben, D., Vandermeulen, D., Suetens, P.: Optimization with soft Dice can lead to a volumetric bias. In: International MICCAI Brainlesion Workshop (BrainLes 2019), LNCS 11992, pp. 89–97. Springer (2020).
18. Adewole, M., Rudie, J.D., Gbdamosi, A., et al.: The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa). arXiv:2305.19369 (2023).
19. Ždímalová, M., Boratková, K., Sitár, V., Sebö, Ľ., Lehotská, V., Trnka, M.: 3D Medical Image Segmentation with 3D Modelling. Bioengineering (Basel) 13(2), 160 (2026).
20. Paszke, A., Gross, S., Massa, F., et al.: PyTorch: An imperative style, high-performance deep learning library. arXiv:1912.01703 (2019).
21. Ma, J., Li, F., Wang, B.: U-Mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv:2401.04722 (2024).
22. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. arXiv:2401.13560 (2024).