Neuro-Oncology Advances

8(1), vdag024, 2026 | https://doi.org/10.1093/noajnl/vdag024 | Advance Access publication February 13, 2026

Enhancing efficiency in pediatric brain tumor segmentation using a pathologically diverse single-center clinical dataset

Arianna Piffer, Josef Alois Buchner, Antonio Giulio Gennari, Patrice Grehten, Selma Sirin, Elizabeth Ross, Ivan Ezhov, Marcel Rosier, Jan Caspar Peeken, Marie Piraud, Bjoern Menze, Ana Guerreiro Stücklin, Andras Jakab†, and Florian Kofler†

All author affiliations are listed at the end of the article

Corresponding Author: Arianna Piffer, MD, Division of Oncology and Children's Research Center, University Children's Hospital Zurich, Lenngstrasse 30, Zurich 8008, Switzerland (arianna.piffer@kispi.uzh.ch).

†These authors share co-last authorship.

Abstract

Background. Brain tumors are the most common solid malignancies in children, encompassing diverse histological, molecular subtypes, and imaging features and outcomes. Pediatric brain tumors (PBTs), including high- and low-grade gliomas (HGG and LGG), medulloblastomas (MB), ependymomas, and rarer forms, pose diagnostic and therapeutic challenges. Deep learning (DL)-based segmentation offers promising tools for tumor delineation, yet its performance across heterogeneous PBT subtypes and MRI protocols remains uncertain.

Methods. A retrospective single-center cohort of 174 pediatric patients with HGG, LGG, MB, ependymomas, and other rarer subtypes was used. MRI sequences included T1, T1 post-contrast (T1-C), T2, and FLAIR. Manual annotations were provided for 4 tumor subregions: whole tumor (WT), T2-hyperintensity (T2H), enhancing tumor (ET), and cystic component (CC). A 3D nnU-Net model was trained and tested (121/53 split), with segmentation performance assessed using the Dice similarity coefficient (DSC) and compared against intra- and interrater variability.

Results. The model achieved robust performance for WT and T2H (mean DSC: 0.85), comparable to human annotator variability (mean DSC: 0.86). ET segmentation was moderately accurate (mean DSC: 0.75), while CC performance was poor (mean DSC: 0.26). Segmentation accuracy varied by tumor type, MRI sequence combination, and location. Notably, T1, T1-C, and T2 combined produced results nearly equivalent to the full protocol.

Conclusions. DL-based segmentation is feasible for PBTs, particularly for T2H and WT. Challenges remain for ET and CC segmentation, highlighting the need for further refinement. These findings support the potential for protocol simplification and automation to enhance volumetric assessment and streamline pediatric neuro-oncology workflows.

Key Points

- The publicly released model achieves high accuracy in WT and T2H, moderate on ET and limited on CC
- Automatic segmentations reduced manual contouring time by up to 83%
- Optimized MRI protocols preserve performance, enabling shorter scans

PBTs are the most common solid cancers in children and the leading cause of cancer-related child mortality.[1] While they may share certain features with adult brain tumors, PBTs show distinct biological and radiological characteristics, including variations in histologic subregions such as the enhancing tumor core and cystic component. Accurate characterization and

Received June 27, 2025; accepted January 30, 2026.

© The Author(s) 2026. Published by Oxford University Press, the Society for Neuro-Oncology and the European Association of Neuro-Oncology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

Importance of the Study

PBTs require precise segmentation for treatment planning and response assessment. Manual segmentation is labour-intensive and variable, necessitating automated solutions. While DL models have shown success in adult tumors, their performance on pediatric cases is limited due to anatomical differences, fewer cases and tumor heterogeneity. This study presents a DL-based segmentation model trained on single-centric, pathologically diverse pediatric MRI data, demonstrating high accuracy for WT and T2-hyperintensity segmentation. By quantifying time savings and evaluating MRI sequence contributions, our work provides not only a segmentation tool, but also practical guidance for optimizing imaging workflows in the pediatric setting.

segmentation of these subregions are important for treatment planning, response assessment, and for assessing prognosis.[2]

With the advent of novel therapies for PBTs (eg, target therapies), precise and reproducible methods for evaluating treatment response or relapse are essential. Traditional response assessment methods rely on 2D measurements, which assume uniform tumor growth[3]; however, this may not accurately capture the complex, often irregular growth patterns of PBTs, moreover after therapy. Volumetric assessment through tumor segmentation provides a better estimate of the 3D volume and morphology of the growing tumor.[4–6] This is important for PBTs, as they are generally irregularly shaped and often have complex, mixed solid and cystic components.[3] The quantification of these subregions often requires manual segmentation, which is labour-intensive, time-consuming, and subject to interrater variability.[7]

DL-based automatic segmentation has significantly advanced adult brain tumor analysis as reflected by the impact the Brain Tumor Segmentation (BraTS) Challenge made on the field, also leading to open-source software, such as BraTS toolkit and BraTS Orchestrator.[8–16] However, models trained on adult tumors do not generalize well to PBTs due to anatomical differences in the developing pediatric brain and distinct tumor locations and radiological features.[17,18] In adults, peritumoural edema is frequently present and routinely segmented as a distinct region—especially in the context of standardized frameworks such as BraTS. However, in the pediatric setting, peritumoral edema is frequently absent or radiologically negligible. Instead, in pediatric, T2-hyperintensity emerges as the most reliable marker for delineating tumor boundaries. T2H often corresponds to the full tumor extent and can be present even in the absence of contrast enhancement, which is common in several pediatric tumor subtypes. This observation underscores the need for adapted annotation protocols and segmentation strategies tailored to the pediatric population, where traditional adult-centric labels may be insufficient or misleading.

While some studies have proposed automated segmentation methods for PBTs, these models often demonstrate lower performance, are limited to specific histologies, or rely on a minimal number of MRI sequences or segment the whole tumor without considering subregions, restricting their clinical applicability.[19–25] We conclude that there remains a clear need for a comprehensive segmentation approach tailored to PBTs: one that is accurate, efficient, and generalizable across histological types and tumor locations, capable of delineating subregions, and integrating data from multiple MRI sequences. To address these challenges and to improve PBT quantification in the clinical practice, we propose a DL-based segmentation framework tailored specifically for PBTs by training it on a clinically and histologically diverse dataset. We evaluated segmentation accuracy across ages, tumor localizations, and histologies by analyzing specific tumor subregions, comparing performance and reproducibility with human annotators, and assessing the contribution of each MRI sequence to accuracy. The latter step was done by evaluating models trained with different MRI sequence combinations to determine the optimal balance between accuracy and clinical feasibility.[26]

Methods

Patient Cohort

A retrospective collection of MRI scans was conducted, including only pre-treatment scans obtained at initial diagnosis from pediatric patients aged 0-18 years treated at the University Children's Hospital Zurich between 2014 and 2022. Patients with first available scan performed post-biopsy or external ventricular drain placement without significant malformation of brain skull and/or parenchyma were also considered. Patients were excluded from the study if any of the 4 standard scans (native T1-weighted sequence (T1), post-contrast T1-weighted sequence (T1-C), T2-weighted sequence (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR) were missing, or if the scans were not acquired prior to surgical resection. A total of 174 subjects met the inclusion criteria for this study (Figure 1 and Table 1).

Data Preprocessing and Tumor Subregion Segmentation

For all patients, a standardized preprocessing approach was applied to MRI scans. First, files were converted from Digital Imaging and Communications in Medicine (DICOM) format to Neuroimaging Informatics Technology Initiative (NIfTI, gzipped) format, during which all patient identifiers were removed (Figure 1).

Second, a multistep co-registration was conducted. All MRI modalities were registered to their respective T1-C scan, followed by a registration to the 1 mm^3 isotropic SRI-24 atlas space.[27] All co-registrations were conducted as rigid

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

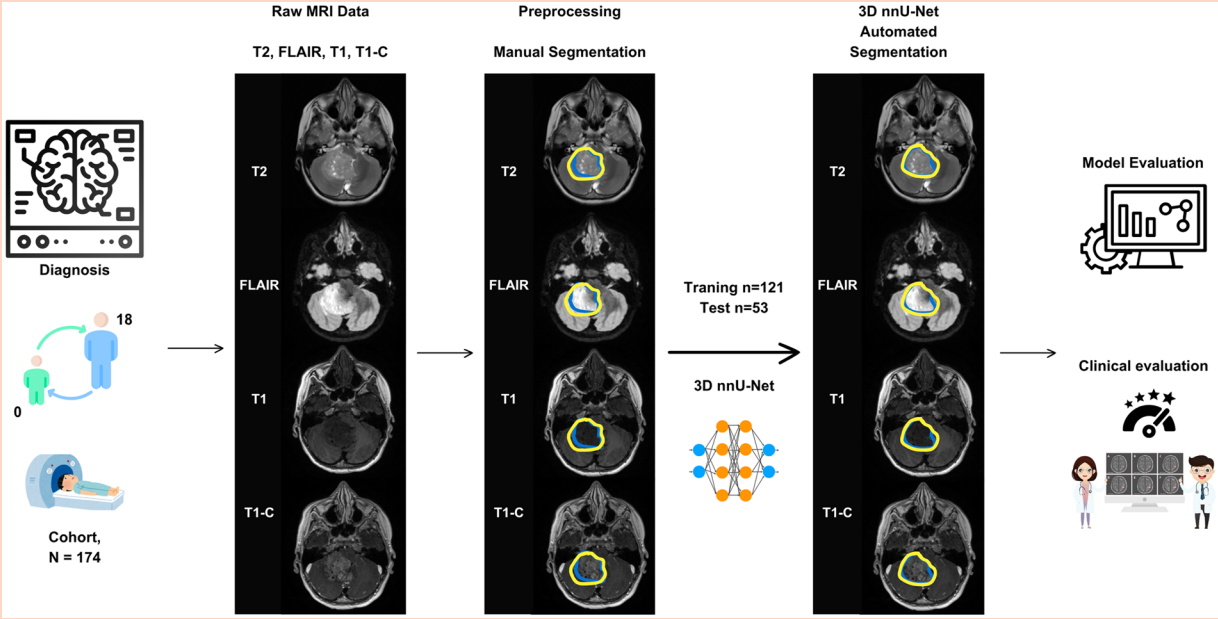

Figure 1. Study workflow, including data collection, preprocessing, expert segmentation of the tumors, model training and testing and model evaluation. Statistical model evaluation includes the Dice score (DSC) as well as volume difference. The model clinical evaluation was based on 4 star rating (rejection, usable with major modifications, usable with minimal modifications, usable).

registrations to enable accurate volumetry. The pre-processing pipeline is publicly available on https://github.com/BrainLesion/preprocessing[28] as part of the BrainLesion suite (https://github.com/orgs/BrainLesion/repositories). Brain extraction by means of skull stripping was intentionally omitted to keep the optic pathway intact and to allow for tumor segmentation within this region.

Manual annotation was performed in 3D Slicer (Version 4.13.0) with three-dimensional axial, sagittal, and coronal views.[29] A medical doctor with experience in pediatric neuro-oncology (AP) carried out the first annotations under the guidance and supervision of 2 board-certified neuro-radiologists (AGG, PG). Next, the neuroradiologists additionally checked and manually adapted the final segmentations.

Due to the diverse histological subtypes and unique MRI characteristics of pediatric tumors, we modified the standard BraTS annotation protocol to better align with the MRI features of the most common tumor types. The annotated tumor subregions included T2H, ET, and CC. The subregions were overlapping in a non-exclusive fashion, thusmeaning a single voxel can be assigned to multiple label groups. ET was characterized by areas with enhancement on T1-C compared to T1, CC typically exhibited a hyperintense signal (bright) on T2 and a hypointense signal (dark) on T1-weighted MRI.

Intrarater and Interrater Variability

To assess the variability in segmentation evaluation among human experts, we measured the intra- and interrater variability using the DSC. We selected a total of 20 of the 174 patients based on tumor type and patient age from both the training and test sets for intrarater and interrater variability analysis. Two radiologists independently segmented these cases, differentiating between the different subregions. Intrarater variability was assessed by having each radiologist segment the same cases at different time points, while interrater variability was determined by comparing segmentations between the 2 radiologists. To quantitatively evaluate pairwise interexpert and intraexpert variability between the 2 annotators and intra variability was evaluated with the DSC to determine if model performance was comparable to intraexpert and interexpert variability.

Neural Network Training

We selected 53 of the 174 patients based on tumor type and patient age to create balanced training and test sets. Patients with lower-quality images (due to poor resolution, motion artifacts, braces, etc.), which resulted in segmentation maps with higher ambiguity, were included exclusively in the training set. We used nnU-Net version 2.5.2 to train our deep-learning segmentation models.[30] After extracting the dataset fingerprint, a convolutional neural network with skip connections is trained according to the selected default experiment planner in the high-resolution 3D configuration. This results in 1000 epoch training runs with a linear learning rate decay.

Our experiments were conducted on a workstation with an Intel 9940X CPU and 2 NVIDIA RTX 8000 GPUs using CUDA version 12.7. This resulted in training times of approximately 67s per epoch using a single GPU, accumulating to a total of ~20 h for 1 training run.

The deep learning (DL) models are publicly available[31] as part of the BrainLesion suite.

Table 1. Overview of patient demographics, tumor histology, and primary tumor locations in the study cohort

Parameter	Patients cohort	
	Training (n = 121)	*Test (n = 53)*
Age (y)		
Median (range)	6 (2.75-10)	6 (2.5-10)
Sex		
Female	51	23
Male	70	30
Tumour types		
Medulloblastoma	20	11
Ependymoma	13	6
High-grade/diffuse midline glioma	14	10
Low-grade glioma	54	18
Other[a]	20	8
Germ cell tumors	*7*	*2*
Craniopharyngioma	*3*	*1*
Meningioma	*1*	*1*
Pinealoblastoma	*1*	*0*
Papillary tumor of the pineal region	*1*	*0*
Choroid plexus papilloma	*1*	*1*
Hemangioblastom	*1*	*1*
Atypical Teratoid/Rhabdoid Tumors (AT/RT)	*1*	*1*
Not specific	*4*	*1*
Primary tumor location		
Brain hemispheres	21	8
Posterior fossa	58	28
Brainstem	14	7
Pinealis	8	3
Other[a]	20	7
Optic	*5*	*2*
Thalamus	*7*	*3*
Chiasma	*4*	*1*
Suprasellar	*4*	*1*

[a]Italic values represent subgroups within the "Other" tumour types and "Other" primary tumor location categories and are reported for descriptive completeness.

Label postprocessing

To eliminate isolated false positives, particularly in the ET and CC channels, we applied a size-based filtering postprocessing step. Therefore, we applied a connected component analysis (CCA) for all 3 channels using cc3d[32] with 26-connectivity. This defines voxels that touch each other with at least 1 corner as connected.

Subsequently, we removed connected components too small to influence clinical decision-making[33,34] using a threshold of 125 voxels corresponding to 0.125 mL. While we report metrics for the threshold of 125 voxels, the implemented postprocessing step is fully customizable in our public code repository, allowing users to define their own voxel threshold and connectivity rules depending on their application needs (eg, research, clinical trials, and surgical planning).

MRI Sequences Combination Analyses

To evaluate the potential contributions of each MRI sequence to the segmentation accuracy, we trained models for 3 categories of sequence combinations. First, we trained a model for each modality individually to quantify its predictive value. Second, we trained a model for all 4 sequences together. Last we trained models with different combinations of input sequences: T1-C only, T1 only, T2 only, FLAIR only, T1-C + T1 + T2, T1-C + T2 + FLAIR, T1-C + T1, T1-C + T2, T1 + T2, T1 + FLAIR, T1-C + FLAIR, T2 + FLAIR.

Performance Analysis

We evaluate model performance in 2 ways. First, we assess the accuracy of segmentations using volumetric metrics, such as the DSC. Second, we conduct an expert evaluation to assess the clinical utility and applicability of the segmentations.

Volumetric and accuracy assessment

To compare the model predictions with the expert-curated reference segmentation we compute the DSC and volume difference. We evaluate each label channel (T2H, ET, CC) individually and a combination of all, which we call WT inspired by BraTS.[11] We compute similarity metrics using panoptica.[35] To align with radiological nomenclature, we utilized the T2H label, as it provides an excellent evaluation of tumor volume and treatment response in pediatric patients.

Randomized blinded clinical acceptability testing

To assess clinical acceptability[36] model predictions were further evaluated by 3 expert assigning a score from 1 to 4 stars for each segmentation of the test set across different subregions. For each of the 53 cases of the test set, each expert was presented with the 3 3D nn-UNet segmentations: T2H, ET, and CC. Experts (AGG, SS) were given instructions and asked to rate each of the 3 segmentations. The star ratings were defined as follows:

- 1 star: The segmentation is completely incorrect/not in the right location.
- 2 stars: The segmentation is in the correct location but requires significant modifications.
- 3 stars: The segmentation is in the correct location but needs minor adjustments.
- 4 stars: The segmentation is clinically usable and perfect.

Evaluation of contouring time

The time required for segmentation was measured in a sample of 20 patients to estimate the performance improvement enabled by auto-contouring. Since, in clinical practice, AI-generated contours must always be reviewed and potentially modified by a neuroradiologist, the reduction in segmentation time achieved with auto-contouring was calculated as:

$$\textit{Relative reduction in time} = \frac{\left(T(Cman) - T(CAI, adj)\right)}{T(Cman)},$$

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

where *T(Cman)* represents the time required for manual segmentation and *T(CAI, adj)* is the time needed to obtain an adequate segmentation after adjusting the AI-generated contours.[37]

The analysis was conducted using the 3D Slicer software. Specifically, the segmentation time was recorded both without the assistance of autocontouring and separately when modifying the segmentations generated by the neural network to achieve a satisfactory segmentation.

Results

Whole Tumor and Subregion Segmentation

Whole tumor segmentation

The WT segmentation achieved a DSC represented as median (SD) of 0.85 (0.21) on the test set, indicating a strong agreement between the model's predictions and the ground truth segmentations. The low volume difference for WT segmentation (median: 16.3%, SD: 23.5%) suggests that the predicted volumes closely approximated the ground truth volumes. These results suggest that most patients would need little to no manual adjustment of the whole tumor boundary for volume assessment. Performance metrics—DSC and Volume Difference—for the WT, T2H, ET, and CC on the test set are summarized in the boxplot shown in Figure 2A.

T2 hyperintensity segmentation

To align with radiological nomenclature, we utilized the T2H as structure to segment. The model's DSCs for T2H closely matched those of WT segmentation, and the volume difference for T2H segmentation remained consistent with that of the WT (median: 13.5%, SD: 23.5%) (Figure 2A). This reinforces its suitability as an optimal metric for tumor volume calculation and response assessment. As WT is the combination of 3 separate subregion and the T2H region typically has the largest extent, the WT and T2H labels were often identical. Consequently, with a mean DSC of 0.85 and a SD of 0.21 on the test set, T2H segmentation demonstrated the same level of performance as WT segmentation (Table 2 and Figure 2).

Enhancing tumour subregion segmentation

The model demonstrated good performance in enhancing tumor segmentation, achieving a DSC of 0.75 (0.34) and a volume difference of 18% (SD = 29%) on the test set (Figure 2A). However, the variability in performance across the enhancing component could be attributed to the differing levels of enhancement. For instance, the network performance was positively correlated with the intensity of contrast enhancement (Figure 3A), but underperformed with more mildly enhancing regions (Figure 3B), where its segmentation accuracy was lower. This variability is also seen in the inter-variability of our experts.

Cystic component segmentation

The CC demonstrated very low performance, with mean, median, and standard deviation DSCs of 0.26, 0, and 0.35, respectively. However, only 30% of cases contained a cystic component, which included 2 subtypes: ring-enhancing and non-enhancing cysts (Figure 3C and D). The non-enhancing cysts posed greater segmentation challenges, contributing to the overall lower accuracy of the CC. This reduced performance can be attributed to the heterogeneity of the cystic component and its limited representation in the training data.

In the intravariability the median and standard deviation of the DSCs were 0.8, 0.128 and in the intervariability were 0.72, 0.224.

Evaluation of Clinical Utility of Segmentations

We implemented a Likert scale rating system to better reflect the clinical utility of our model's segmentations.[36]

Expert 1 assigned the following scores, presented as mean (SD): for T2H DSCs were 3.5 (0.67), for ET 3.14 (0.87), for CC 1.94.[1] Expert 2's scores were: for T2H DSCs were 3.58 (0.6), for ET 3.12,[1] for CC 2 (1.13) (Figure 4).

Time saving in clinical routine

To assess the clinical impact of automated segmentation, we compared manual and AI-assisted contouring times across all test subjects. For T2H, manual segmentation required a median of 48 minutes and 25 seconds, while AI-assisted segmentation reduced this to 8 minutes and 14 seconds—yielding a time saving of 40 minutes and 11 seconds (83%). For ET, the time was reduced from 26 minutes and 21 seconds to 15 minutes and 12 seconds, corresponding to a saving of 11 minutes and 9 seconds (42%). For CC, segmentation time decreased from 22 minutes and 54 seconds to 16 minutes and 48 seconds, saving 6 minutes and 6 seconds (27%). These results indicate substantial time reductions with AI assistance, particularly for T2H, and support the integration of automated segmentation tools to improve efficiency in clinical workflows.

Segmentation Performance in Various MRI Sequence Combinations

Based on previous work, which investigated the optimal MRI sequence combinations for automatic segmentation, emphasizing the importance of reducing scan time by omitting redundant sequences without compromising segmentation performance,[26] which could be particularly valuable for pediatric imaging.

We evaluated the DSCs for different tumor subregions and WT segmentation across various MRI sequence combinations (Supplementary Figure S1 and Supplementary Table S1).

In summary, the combination of T1-C+T1+T2 showed the best performance for segmenting WT, T2H, and ET subregions, with reasonably high DSCs (WT: 0.84, T2H: 0.84, ET: 0.69).

Segmentation Performance: Comparison with BraTS-PED-Trained Model

Pre-trained BraTS models showed limited generalization when applied to our pediatric cohort. Qualitative examples revealed frequent segmentation failures, including

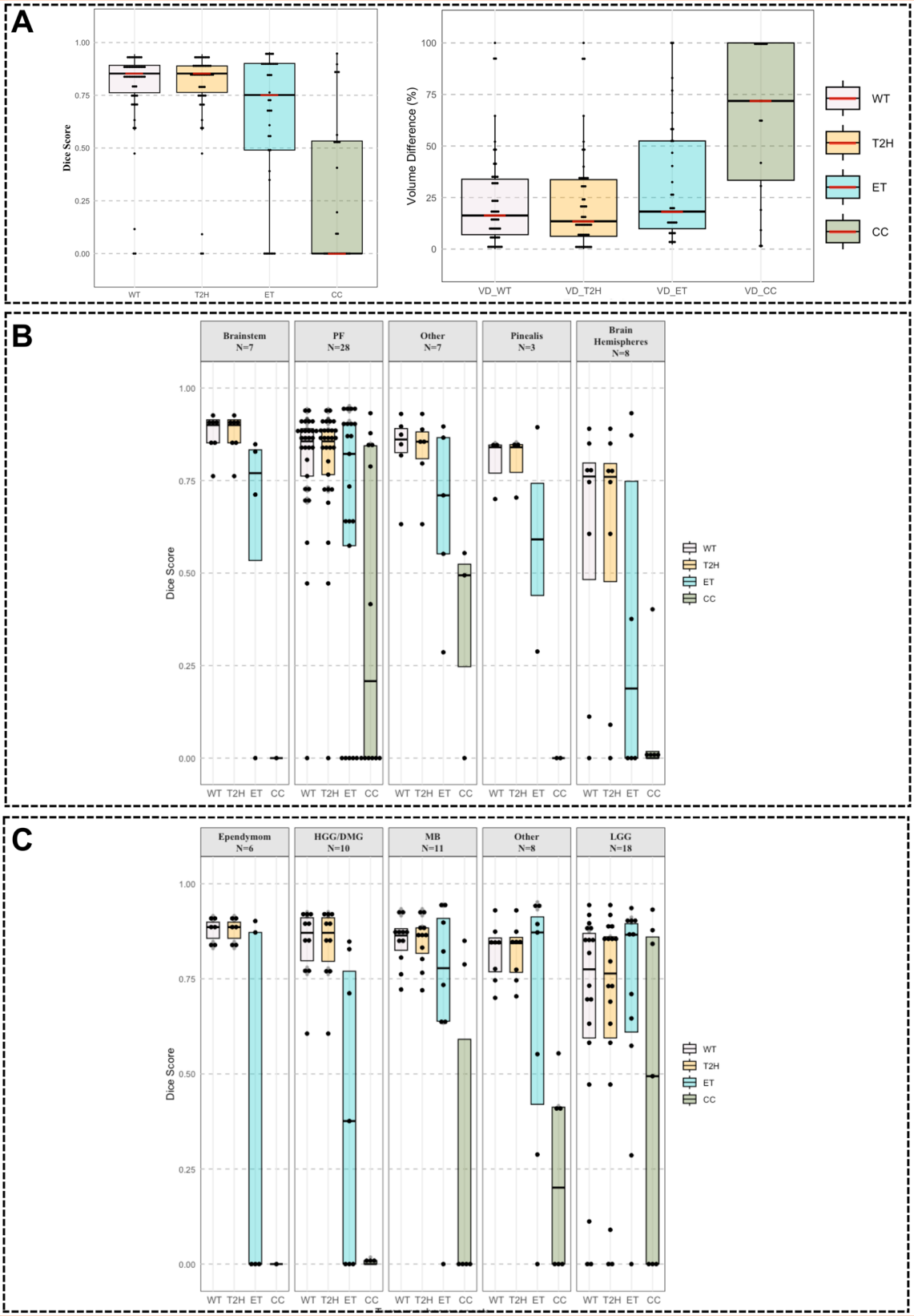

Figure 2. Box plots summarizing the segmentation performance on the test set for T2H, ET and CC. (A) DSC distribution for different tumor subcomponents (left) and volume difference (right), averaged across all tumor types and locations. (B) DSC performance stratified by tumor localization. (C) DSC performance stratified by tumor type. Abbreviations: CC, cystic component; ET, enhancing tumor component; T2H, T2-hyperintense tumor region; WT, whole tumor (combined mask).

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

Table 2. DSC values (median and standard deviation) of Whole Tumor, T2-Hyperintensity and Enhancing Tumor segmentation of the 3D U-Net compared to intraoperator and interoperator variability, stratified by tumor histology and location

	DSC (Median (SD))		
Whole Tumour	U Net	Intrarater variability	Interrater variability
Tumour Types			
Medulloblastoma	0.86 (0.07)	0.85 (0.12)	0.83 (0.09)
Ependymoma	0.89 (0.03)	0.87 (0.06)	0.87 (0.07)
High-grade/diffuse midline glioma	0.87 (0.1)	0.88 (0.04)	0.81 (0.08)
Low-grade glioma	0.82 (0.31)	0.74 (0.2)	0.72 (0.08)
Other	0.85 (0.07)	0.82 (0.08)	0.78 (0.11)
Primary tumor location			
Brain hemispheres	0.76 (0.34)	0.88 (0.1)	0.86 (0.07)
Posterior fossa	0.86 (0.19)	0.85 (0.2)	0.83 (0.09)
Brainstem	0.91 (0.06)	0.89 (0.09)	0.87 (0.09)
Pinealis	0.84 (0.08)	0.82	0.81
Other	0.88 (0.1)	0.78 (0.03)	0.78 (0.09)
T2-Hyperintensity			
Tumor types			
Medulloblastoma	0.86 (0.07)	0.84 (0.13)	0.83 (0.09)
Ependymoma	0.89 (0.03)	0.88 (0.08)	0.87 (0.07)
High-grade/diffuse midline glioma	0.87 (0.1)	0.88 (0.08)	0.81 (0.09)
Low-grade glioma	0.8 (0.31)	0.74 (0.2)	0.72 (0.08)
Other	0.85 (0.07)	0.82 (0.08)	0.78 (0.11)
Primary tumor location			
Brain Hemispheres	0.76 (0.35)	0.88 (0.1)	0.86 (0.07)
Posterior Fossa	0.86 (0.19)	0.85 (0.2)	0.83 (0.09)
Brainstem	0.91 (0.06)	0.89 (0.09)	0.87 (0.09)
Pinealis	0.84 (0.08)	0.82	0.81
Other	0.86 (0.1)	0.78 (0.03)	0.78 (0.09)
Enhancing tumor			
Tumor Types			
Medulloblastoma	0.76 (0.32)	0.82 (0.06)	0.85 (0.4)
Ependymoma	0 (0.5)	0.66 (0.3)	0.6 (0.43)
High grade/diffuse midline glioma	0.49 (0.38)	0.45 (0.4)	0.65 (0.4)
Low-grade glioma	0.87 (0.3)	0.84 (0.2)	0.72 (0.38)
Other	0.88 (0.19)	0.7 (0.12)	0.62 (0.39)
Primary tumor location			
Brain hemispheres	0.52 (0.34)	0.68 (0.4)	0.38 (0.4)
Posterior fossa	0.8 (0.39)	0.76 (0.2)	0.68 (0.2)
Brainstem	0.8 (0.42)	0.83 (0.04)	0.75 (0.4)
Pinealis	0.8 (0.43)	0.68	0.78
Other	0.7 (0.3)	0.74	0.6

substantial undersegmentation and recurrent mislocalization, most commonly involving ventricular structures rather than tumor tissue. In contrast, the pediatric-trained model produced contours that were more spatially consistent with the reference annotations. Quantitatively, Dice scores obtained with the BraTS models were low overall and consistently poor across all tumor histologies and anatomical locations. This pattern was observed even though the evaluated BraTS model corresponded to the top-ranked method of the BraTS-PED Challenge 2023 and 2024. These results indicate that the limited performance of BraTS-derived models reflects a systematic lack of generalization to pediatric brain tumors rather than deficiencies restricted to specific tumor subtypes or locations (Supplementary Figure S2).

Discussion

This study introduces a DL-based framework for the segmentation of PBTs, addressing a critical gap in automated segmentation techniques specifically tailored to pediatric patients. Due to the distinct anatomical and radiological characteristics of pediatric tumors, models trained on adult datasets often fail to generalize effectively. Our approach,

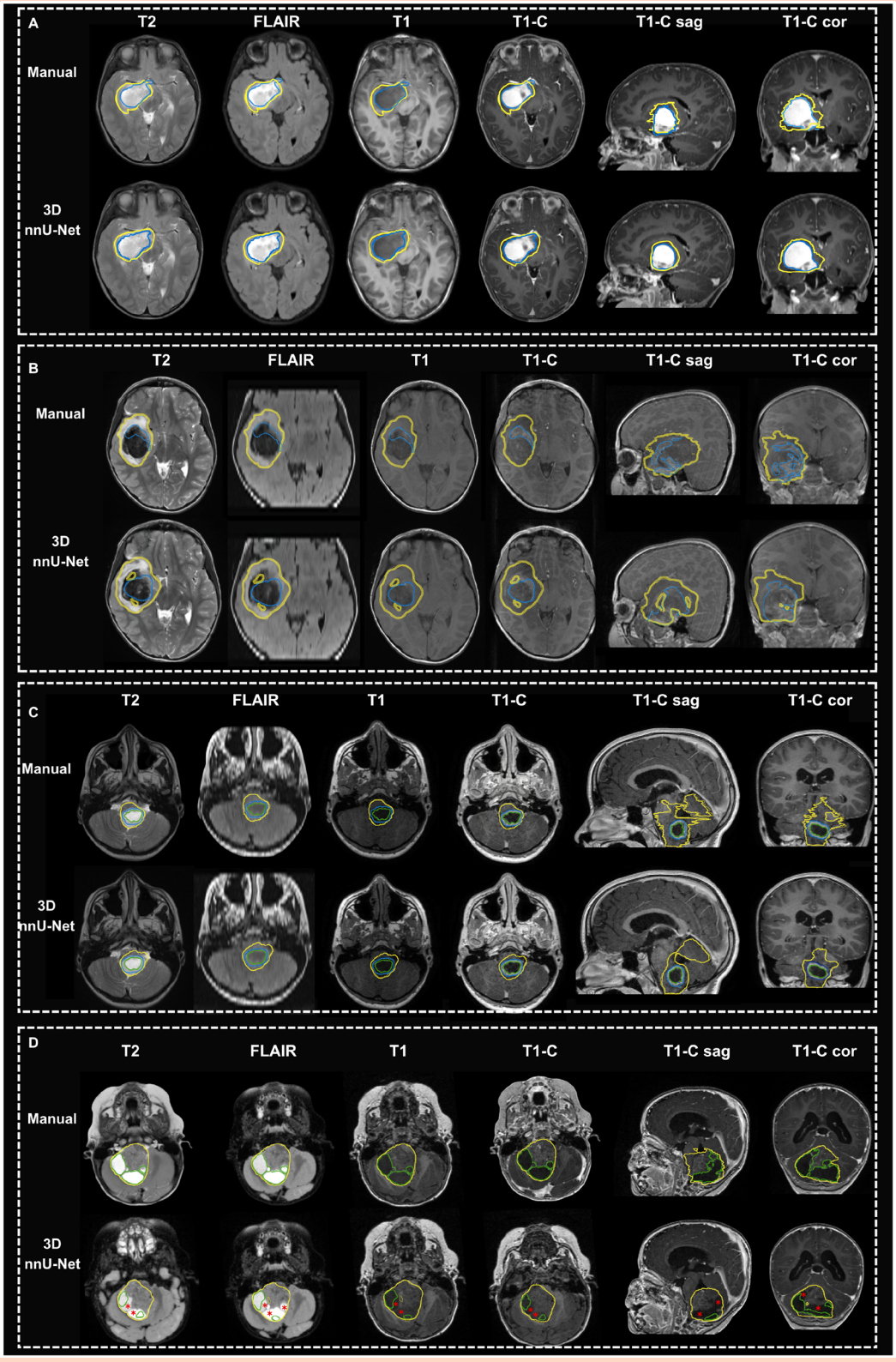

Figure 3. Example tumor segmentations: cystic and enhancing component segmentation. Panel A shows enhancing components, with bright enhancement, manual annotations (top row) with 3DU-Net predictions (bottom row). Panel B shows enhancing components, with mild enhancement, manual annotations (top row) with 3DU-Net predictions (bottom row). The manual segmentations display a characteristic "Christmas tree effect" in the coronal and sagittal views—a visual artifact where isolated voxels appear stacked or scattered across slices, resembling a conical or tree-like pattern. Abbreviations: CC, cystic component; ET, enhancing tumor component; T2H, T2-hyperintense tumor region; WT, whole tumor (combined mask).

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

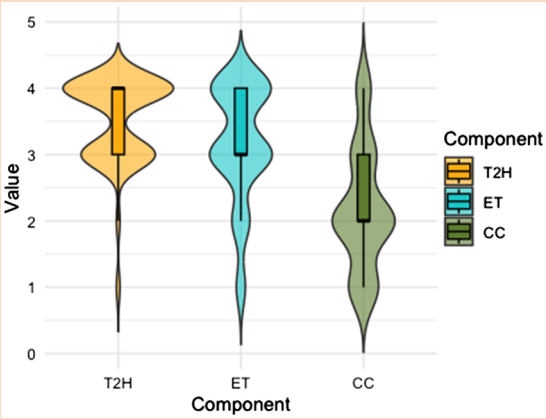

Figure 4. Violin plot with overlaid boxplots showing the distribution of segmentation metrics for the individual tumor components T2H, ET, and CC. Abbreviations: CC, cystic component; ET, enhancing tumor component; T2H, T2-hyperintense tumor region.

leveraging multiparametric MRI sequences (T1,T1-C,T2, and FLAIR), demonstrates human-like performance in whole tumor and tumor subregion segmentation.

Our results are significantly higher than other pediatric DL models,[24,25] which achieved DSC ranging from 0.71 to 0.76; and align with a prior pediatric-specific segmentation studies.[26] Unlike several existing models that focus solely on specific histologies or tumor locations—thereby limiting their clinical applicability—our model successfully segments a diverse range of tumor histologies and locations, demonstrating adaptability beyond single tumor types. A fundamental challenge in pediatric neuro-oncology remains the scarcity of large, annotated imaging datasets, which hinders the development of robust and generalizable DL models.[25] Despite this, our model effectively segments a wide spectrum of tumor subtypes and anatomical locations, showcasing its potential for broader clinical use.

A key contribution of our work is the development of a new annotation protocol specifically tailored to the clinical utility of pediatric neuro-oncology. Unlike the BraTS framework, which includes enhancing tumor, necrosis and oedema—features less relevant for PBTs—we prioritized subregions that are critical for clinical decision-making in children, such as T2H. T2H segmentation demonstrated nearly identical performance to WT (mean DSC: 0.85), validating it as a reliable surrogate for volumetric tumor assessment (Figure 2A). This shift in labelling emphasis not only improves segmentation accuracy but also enhances clinical relevance, particularly in longitudinal monitoring scenarios where contrast enhancement may fluctuate or be absent altogether. As such, the adoption of T2H-based segmentation represents a necessary evolution in the development of automated tools for pediatric neuro-oncology.

Despite the robust training framework, segmentation in pediatric neuro-oncology presents unique challenges due to anatomical variability and radiological heterogeneity. To mitigate model sensitivity to false positives—particularly in ET and CC—we implemented a voxel thresholding step during postprocessing. We excluded segmentations smaller than 125 voxels (0.125 mL), which are typically non-clinically relevant and often represent noise. This small-volume threshold, in combination with 26-connectivity filtering, ensures that only spatially coherent and volumetrically significant tumor regions are retained.[33,34] Our segmentation framework enables users to adapt this thresholding step based on specific application needs, providing flexibility for clinical versus research use cases.

Another novel aspect of this study is our investigation into how segmentation performance varies by tumor type and patient age. Our analysis revealed that segmentation accuracy differed significantly across tumor histologies and anatomical locations, with ependymomas showing the highest accuracy and LGGs the lowest (Figure 2B). This likely reflects the anatomical variability, less pronounced enhancement, and heterogeneous morphology typical of LGGs. Age-related variability also suggests that developmental anatomy influences model generalizability, underscoring the need for large, diverse training sets. Additionally, similar performance variations were observed across anatomical locations, reflecting structural and imaging complexities unique to each brain region (Figure 2C). This variability highlights the heterogeneous nature of PBTs and the complexity of their radiological features, which can make consistent segmentation more challenging.[2,30] It also underscores the need for models capable of handling a wide range of tumor types and brain locations, which this study addresses by training the model on a diverse cohort of patients with different tumor histologies.

In terms of specific tumor subregions, the segmentation of T2H showed excellent performance (Figure 2). These results are particularly relevant given that only 63% of patients exhibited enhancing tumor components, and merely 31% presented with a cystic component. The use of T2H ensures a more comprehensive assessment of tumor burden and facilitates better clinical decision-making in the pediatric population.

ET segmentation performed reasonably well, with a mean DSC of 0.75, although some variability was observed due to the differing levels of enhancement across tumors, and also seen in the intervariability measurement. In contrast, CC segmentation showed lower performance, with a mean DSC of 0. This poor performance can be attributed to the heterogeneous nature of cystic components, which may vary significantly in appearance across different tumor types and patient cohorts.[2]

These challenges are consistent with findings from Familiar et al, who highlight the inherent difficulties in distinguishing mild enhancement from non-enhancing tissue and delineating CC. Pediatric tumors often exhibit mixed solid-cystic characteristics and bright yet mild enhancement, which complicates segmentation and can introduce interobserver and intraobserver variability[38]. As illustrated in Figure 3, variability is minimal in clearly enhancing tumors but increases substantially in mildly enhancing cases. Similarly, the cystic components exhibit diverse imaging appearances, complicating consistent segmentation. We also observed a "Christmas tree effect" in manual segmentations—discontinuous voxels visible in coronal and sagittal planes due to axial-only annotations—underscoring the limitations of manual approaches and the need for automated methods ensuring 3D coherence. This artifact—where isolated voxels appear scattered or stacked across slices,

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

forming a tree-like structure—is likely due to the fact that manual segmentations are typically performed in the axial plane which underscores the limitations of two-dimensional manual approaches and the need for 3D approaches to ensure spatial coherence[39]. These variations contribute to the difficulty in training the model, leading to segmentation failures. When evaluating cases with lower DSCs in ET segmentation, 2 key trends emerged[1]: the model occasionally predicted a few voxels of enhancing tumor when the ground truth contained none, resulting in a DSC of zero, and[2] mildly enhancing regions in the ground truth were sometimes missed by the model. These findings underscore the need for improved classification techniques and more precise ground truth labelling. Familiar et al also stress the need for standardized imaging protocols and annotation criteria. Variability in imaging acquisition across institutions limits model generalizability, reinforcing the importance of multi-institutional collaboration and standardized data curation.

Our interrater and intrarater analysis supports the model's reliability, with T2H and WT segmentation showing concordance with human performance.[40] The models' lower segmentation performance for ET and CC might be explained by greaterambiguity in segmenting these subregions, a phenomenon also observed in human segmentations as reflected by the lower interrater and intrarater scores for these channels. These findings highlight the potential of AI to reduce variability and improve reproducibility—crucial for clinical adoption.

Manual segmentation of these tumor subregions remains the gold standard; however, this approach is time-consuming, labor-intensive, and subject to interoperator variability.[7,41] DL-based automation offers a scalable solution, enabling improved tumor tracking, treatment response assessment, surgical planning, and potentially radiomics. Despite promising developments, the clinical use of automated pediatric segmentation has been constrained by limited training datasets.[25]

One of the most significant advantages of automated segmentation is the potential reduction in manual contouring time. Our study quantified these time savings, demonstrating an 83% reduction in segmentation time for T2H, 42% for ET, and 27% for CC. These efficiencies are particularly relevant in clinical practice, where radiologists must balance segmentation accuracy with workflow constraints. By minimizing manual intervention while maintaining segmentation quality, DL models can enhance productivity and facilitate real-time clinical decision-making. This is especially beneficial for longitudinal tumor monitoring and treatment response assessment, where volumetric changes over time need to be accurately quantified.

We also systematically explored the contribution of different MRI sequences to segmentation accuracy, evaluating multiple sequence combinations to identify clinically efficient and diagnostically effective inputs. Our results highlight that combining T1, T1-C, and T2 sequences offers an optimal balance of segmentation performance. This sequence triad delivered high DSC across all subregions, offering guidance for protocol optimization in routine clinical settings, especially when scan time is a concern. Notably, we found that for volume assessment and delineation of tumor boundaries, the inclusion of T1-C did not significantly impact performance—the segmentation accuracy ofT2H and WT was essentially the same with or without T1-C. This underscores the limited value of contrast-enhanced sequences forT2-based volumetric assessment, which could further streamline imaging protocols and reduce contrast agent exposure in pediatric patients.

In conclusion, our DL-based segmentation framework represents a significant advancement in pediatric brain tumor segmentation. Importantly, the trained models and preprocessing pipelines have been publicly released to ensure reproducibility and foster future research. By integrating multiparametric MRI and leveraging an advanced 3D nnU-Net architecture, our approach offers a clinically viable tool for volumetric assessment, treatment planning, and response monitoring.

Despite its strong performance, several limitations must be considered. Although trained on a diverse pediatric cohort, the dataset size remains a constraint, potentially affecting the model's generalizability to rarer tumor subtypes or atypical cases. Additionally, segmentation of cystic components and enhancing tumors remains challenging, emphasizing the need for improved ground truth annotations and more refined learning strategies, particularly for tumors with subtle enhancement or heterogeneous cystic structures. Another key challenge is the variability in imaging protocols across institutions, which may impact external validation and highlights the need for multi-institutional studies. While our model significantly reduces manual segmentation time, expert review is still required for quality control, meaning further validation in real-world clinical settings is necessary before widespread implementation.

Future research should focus on expanding training datasets, optimizing model architectures for more challenging tumor subregions, and conducting extensive clinical validation to support broader adoption in pediatric neuro-oncology.

Supplementary Material

Supplementary material is available online at *Neuro-Oncology Advances* (https://academic.oup.com/vdaf241).

Keywords

deep learning | MRI segmentation | nnU-Net | pediatric brain tumours | tumor subregions

Author Contributions

Conceptualization and study design: A.P., A.G.S., B.M., F.K., and A.J. Data curation and annotation: A.P., A.G.G., P.G., S.S., E.R. Model development and evaluation: A.P., J.A.B., I.E., M.R., M.P., J.C.P., F.K. Statistical analysis: A.P., J.A.B., F.K., A.J. Supervision: A.G.S., B.M., F.K., A.J. Manuscript writing—original draft: A.P. Manuscript review and editing: All authors (A.P., J.A.B., A.G.G., P.G., S.S., E.R., I.E., M.R., J.C.P., M.P., B.M., A.G.S., A.J., F.). Shared last authorship: F.K., A.J.

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026

Conflict of Interest Statement

None declared.

Funding

This work was supported by the Swiss National Science Foundation (A.P.) and Forschungszentrum für das Kind Universität Zürich (A.P.). This work was supported by Anna Valentina Lioba Eleonora Claire Javid Mamasani and the Gemeinnützige Hertie Stiftung (F.K.). A.J. received funding from the Prof. Max Cloetta Foundation, EMDO Foundation and Vontobel Foundation. The publication fee was covered by UZH Zurich.

Ethics Statement

This retrospective study was approved by the Cantonal Ethics Committee Zurich (KEK Zürich), BASEC No. 2020-00801 (project title: "Long term outcome of children and adolescents with central nervous system tumors"; principal investigator: Dr. med. Dr. sc. nat. Ana Guerreiro Stüklin). The study was conducted in accordance with the Declaration of Helsinki and applicable Swiss regulations (Human Research Act and Ordinance on Human Research). Written informed consent was obtained from patients or their legal guardians when required. For patients diagnosed and/or treated before 2015 for whom written consent documentation was unavailable, KEK Zürich granted approval for the secondary use of existing clinical and imaging data for retrospective analysis.

Data Availability

The segmentation models and pre-processing pipeline are publicly available via the BrainLesion GitHub repository (https://github.com/BrainLesion; https://github.com/BrainLesion/PeTu). Anonymized imaging data and annotations used in this study may be made available from the corresponding author upon reasonable request for non-commercial academic purposes.

Affiliations

Division of Oncology and Children's Research Center, University Children's Hospital Zurich, Zurich, Switzerland (AP., A.G.S.); Department of Radiation Oncology, TUM School of Medicine and Health, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany (J.A.B., J.C.P.); Department of Neuropaediatrics, University Children's Hospital Zurich, Zurich, Switzerland (A.G.G.); Center for MR- Research, University Children's Hospital Zurich, Zurich, Switzerland (A.G.G., A.J.); Department of Diagnostic Imaging, University Children's Hospital Zurich, Zurich, Switzerland (P.G., S.S.); Geisel School of Medicine at Dartmouth, Georgia Institute of Technology, Hanover, United States (E.R.); Department of Computer Science, TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany (I.E.); TranslaTUM—Central Institute for Translational Cancer Research, Technical University of Munich, Munich, Germany (I.E., F.K.); Helmholtz AI, Helmholtz Zentrum Munich, Munich, Germany (M.R., M.P., F.K.); Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland (M.R., B.M., F.K.); Faculty of Medicine, University of Zürich, Zurich, Switzerland (A.J.); Institute of Radiation Medicine (IRM), Helmholtz Zentrum, Oberschleißheim, Germany (J.C.P.); Partner Site Munich, German Consortium for Translational Cancer Research (DKTK), Munich, Germany (J.C.P.); Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany (F.K.); Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany (F.K.)

References

1. Pollack IF, Agnihotri S, Broniscer A. Childhood brain tumors: current management, biological insights, and future directions: JNSPG 75th anniversary invited review article. *J Neurosurg Pediatr*. 2019;23:261-273. https://doi.org/10.3171/2018.10.PEDS18377
2. Familiar AM, Fathi Kazerooni A, Vossough A, et al. Towards consistency in pediatric brain tumor measurements: challenges, solutions, and the role of artificial intelligence-based segmentation. *Neuro Oncol*. 2024; 26:1557-1571.
3. Bhatia A, Sabin ND, Fisher MJ, Poussaint TY. Review of imaging recommendations from response assessment in pediatric Neuro-Oncology (RAPNO). *Pediatr Radiol*. 2023;53:2723-2741. https://doi.org/10.1007/s00247-023-05780-w
4. Gahrmann R, van den Bent M, van der Holt B, et al. Comparison of 2D (RANO) and volumetric methods for assessment of recurrent glioblastoma treated with bevacizumab—a report from the BELOB trial. *Neuro-Oncol*. 2017; 19:853-861. https://doi.org/10.1093/neuonc/now311
5. von Reppert M, Ramakrishnan D, Brüningk SC, et al. Comparison of volumetric and 2D-based response methods in the PNOC-001 pediatric low-grade glioma clinical trial. *Neurooncol Adv*. 2024;6:vdad172. https://doi.org/10.1093/noajnl/vdad172
6. Ellingson BM, Kim GHJ, Brown M, et al. Volumetric measurements are preferred in the evaluation of mutant IDH inhibition in non-enhancing diffuse gliomas: evidence from a phase I trial of ivosidenib. *Neuro Oncol*. 2022;24:770-778. https://doi.org/10.1093/neuonc/noab256
7. Singh MK, Singh KK. A review of publicly available automatic brain segmentation methodologies, machine learning models, recent advancements, and their comparison. *Ann Neurosci*. 2021; 28:82-93. https://doi.org/10.1177/0972753121990175
8. Kofler F, et al. (2025). *BrainLesion Suite: A Flexible and User-Friendly Framework for Modular Brain Lesion Image Analysis* arXiv preprint arXiv:2507.09036 https://github.com/BrainLesion/preprocessing. Accessed February 22, 2026.
9. Morin O, Vallières M, Jochems A, et al. A deep look into the future of quantitative imaging in oncology: a statement of working principles and proposal for change. *Int J Radiat Oncol Biol Phys*. 2018;102:1074-1082. https://doi.org/10.1016/j.ijrobp.2018.08.032
10. Rathore S, Bakas S, Pati S, et al. Brain Cancer Imaging Phenomics Toolkit (brain-CaPTk): an interactive platform for quantitative analysis of glioblastoma. In: Crimi A, Bakas S, Kuijf H, Menze B, Reyes M, eds.

Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer International Publishing; 2018:133-45. Available from: http://link.springer.com/10.1007/978-3-319-75238-9_12

11. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34:1993-2024. https://doi.org/10.1109/TMI.2014.2377694
12. Mitra S. Deep learning with radiogenomics towards personalized management of gliomas. *IEEE Rev Biomed Eng*. 2023;16:579-593. https://doi.org/10.1109/RBME.2021.3075500
13. Vossough A, Khalili N, Familiar AM, et al. Training and comparison of nnU-Net and DeepMedic methods for autosegmentation of pediatric brain tumors. *AJNR Am J Neuroradiol*. 2024; 45:1081-1089. https://doi.org/10.3174/ajnr.A8293
14. Kofler F, Berger C, Waldmannstetter D, et al. BraTS toolkit: translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. *Front Neurosci*. 2020; 14:125. https://doi.org/10.3389/fnins.2020.00125
15. Kofler F. BraTS orchestrator: Democratizing and disseminating state-of-THE-art brain tumor image analysis. *arXiv*; 2025. https://doi.org/10.48550/arXiv.2506.13807.
16. Canisius J. GlioMODA: Robust glioma segmentation in clinical routine. *medRxiv*. 2025. https://doi.org/10.1101/2025.11.12.25339968.
17. Bhatia A, Khalvati F, Ertl-Wagner BB. Artificial intelligence in the future landscape of pediatric neuroradiology: opportunities and challenges. *AJNR Am J Neuroradiol*. 2024 May;45:549-553. https://doi.org/10.3174/ajnr.A8086
18. Zhang S, Edwards A, Wang S, Patay Z, Bag A, Scoggins MA. A prior knowledge based tumor and tumoral subregion segmentation tool for pediatric brain tumors. arXiv; 2021. http://arxiv.org/abs/2109.14775. Date accessed Jun 14, 2025.
19. Nalepa J, Adamski S, Kotowski K, et al. Segmenting pediatric optic pathway gliomas from MRI using deep learning. *Comput Biol Med*. 2022; 142:105237. https://doi.org/10.1016/j.compbiomed.2022.105237
20. Peng J, Kim DD, Patel JB, et al. Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors. *Neuro Oncol*. 2022; 24:289-299. https://doi.org/10.1093/neuonc/noab151
21. Vafaeikia P, Wagner MW, Hawkins C, Tabori U, Ertl-Wagner BB, Khalvati F. Improving the segmentation of paediatric low-grade gliomas through multitask learning. *Annu Int Conf IEEE Eng Med Biol Soc*. 2022;2022:2119-2122. https://doi.org/10.1109/EMBC48229.2022.9871627
22. Artzi M, Gershov S, Ben-Sira L, et al. Automatic segmentation, classification, and follow-up of optic pathway gliomas using deep learning and fuzzy c-means clustering based on MRI. *Med Phys*. 2020; 47:5693-5701. https://doi.org/10.1002/mp.14489
23. Fathi Kazerooni A, Arif S, Madhogarhia R, et al. Automated tumor segmentation and brain tissue extraction from multiparametric MRI of pediatric brain tumors: a multi-institutional study. *Neuro-Oncol Adv*. 2023;5:vdad027. https://doi.org/10.1093/noajnl/vdad027
24. Boyd A, Ye Z, Prabhu SP, et al. Stepwise transfer learning for expert-level pediatric brain tumor MRI segmentation in a limited data scenario. *Radiol Artif Intell*. 2024; 6:e230254. https://doi.org/10.1148/ryai.230254
25. Kazerooni AF, Khalili N, Liu X, et al. The Brain Tumor Segmentation in Pediatrics (BraTS-PEDs) Challenge: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv; 2024. http://arxiv.org/abs/2404.15009. Date accessed June 14, 2020.
26. Buchner JA, Peeken JC, Etzel L, et al. Identifying core MRI sequences for reliable automatic brain metastasis segmentation. *Radiother Oncol*. 2023; 188:109901. https://doi.org/10.1016/j.radonc.2023.109901
27. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. 2010 May;31:798-819. https://doi.org/10.1002/hbm.20906
28. Kofler F, et al. BrainLesion Suite: A flexible and user-friendly framework for modular brain lesion image analysis. *arXiv*; 2025. https://doi.org/10.48550/arXiv.250709036
29. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*. 2012;309:1323-1341. https://doi.org/10.1016/j.mri.2012.05.001
30. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021; Feb18:203-211. https://doi.org/10.1038/s41592-020-01008-z
31. Piffer A, et al. Enhancing efficiency in paediatric brain tumour segmentation using a pathologically diverse single-center clinical dataset. *arXiv*; 2025. https://arxiv.org/abs/2507.22152 .
32. Silversmith M. cc3d: connected components on multilabel 3D & 2D images [Internet]. https://github.com/seung-lab/connected-components-3d. Date accessed Feb 22, 2026.
33. Kickingereder P, Burth S, Wick A, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*. 2016; 280:880-889. https://doi.org/10.1148/radiol.2016160845
34. Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017; 4:170117. https://doi.org/10.1038/sdata.2017.117
35. Kofler F, Möller H, Buchner JA, et al. Panoptica—instance-wise evaluation of 3D semantic and instance segmentation maps. 2023. http://arxiv.org/abs/2312.02608. Date accessed Jun 14, 2025.
36. Kofler F, Ezhov I, Isensee F, et al. Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. *Melba*. 2023; 2:27-71. https://doi.org/10.59275/j.melba.2023-dg1f
37. Hosny A, Bitterman DS, Guthier CV, et al. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. *Lancet Digit Health*. 2022; 4:e657-e666. https://doi.org/10.1016/S2589-7500(22)00129-7
38. Familiar A. Towards consistency in pediatric brain tumor measurements: Challenges, solutions, and the role of artificial intelligence-based segmentation. *Neuro-Oncology*. 2024;26;1557-1571. https://doi.org/10.1093/neuonc/noae093
39. Kaltenecker D. Virtual reality-empowered deep-learning analysis of brain cells. *Nat Methods*. 2024;21:1306-1315. https://doi.org/10.1038/s41592-024-02245-2
40. Kofler F, Wahle J, Ezhov I, et al. Approaching peak ground truth. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI); 2023; Cartagena, Colombia. pp. 1-6. https://doi.org/10.1109/ISBI53787.2023.10230497
41. Buchner JA, Kofler F, Etzel L, et al. Development and external validation of an MRI-based neural network for brain metastasis segmentation in the AURORA multicenter study. *Radiother Oncol*. 2023; 178:109425. https://doi.org/10.1016/j.radonc.2022.11.014

Downloaded from https://academic.oup.com/noa/advance-article/doi/10.1093/noajnl/vdag024/8483039 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 25 March 2026