Mureddu *et al. Brain Informatics* (2026) 13:10
https://doi.org/10.1186/s40708-026-00296-z

Brain Informatics

RESEARCH Open Access

U-Net-based transfer learning for automated tumour segmentation enabling fully automated [^{18}F]F-DOPA PET analysis in paediatric gliomas

Michele Mureddu[1*†], Rosella Trò[1†], Federico Giovanni Garau[1], Nicolò Trebino[1], Andrea Bianconi[2], Andrea Rossi[3], Antonia Ramaglia[3], Antonio Verrico[4], Claudia Milanaccio[4], Giovanni Morana[5], Massimiliano Iacozzi[6], Francesco Fiz[6], Arnoldo Piccardo[6†] and Marco Massimo Fato[1†]

Abstract

Background PET imaging with [^{18}F]F-DOPA shows great promise for assessing paediatric gliomas. Manual tumour delineation and parameter extraction are time-consuming and prone to inter-operator variability.

Methods We evaluated whether a deep learning model, leveraging transfer learning from adult glioma datasets, could enable a fully automated pipeline for tumour segmentation and PET parameter extraction. Static and dynamic parameters were compared across three approaches: (i) automatic vs semi-automatic, (ii) automatic vs manual, and (iii) manual vs. semi-automatic. Data from 103 paediatric patients (median age 11 years; 54 females, 49 males) with static and/or dynamic [^{18}F]F-DOPA PET scans (2011–2024) were retrospectively included for fine-tuning the deep learning model. Statistical and survival analyses were performed on 90 subjects; dynamic analysis included 32 patients.

Results The best model achieved a Dice score of 0.82 ± 0.11 and was integrated into the pipeline for extracting static and dynamic indices. Automatic Tumour-to-Striatum ratio showed high reproducibility across comparisons ((i) $p = 0.660$, (ii) $p = 0.342$, (iii) $p = 0.639$), while Tumour-to-Background differed significantly when comparing manual delineations ($p < 0.01$). Dynamic parameters demonstrated good reproducibility with the automatic method ($p > 0.05$). Importantly, both automated static indices correlate significantly with tumour grade, with the overall and progression-free survival ($p < 0.05$).

Conclusions Transfer learning enabled a fully automatic [^{18}F]F-DOPA PET pipeline for paediatric gliomas, providing reproducible static and dynamic parameter extraction and correlating with clinically relevant outcomes. This

†Arnoldo Piccardo and Marco Massimo Fato contributed equally to this work.

†Michele Mureddu and Rosella Trò contributed equally to this work.

*Correspondence:
Michele Mureddu
michele.mureddu@edu.unige.it

Full list of author information is available at the end of the article

approach reduces operator dependence and streamlines analysis, supporting potential integration into routine clinical practice.

Keywords U-Net, Transfer learning, [^{18}F]F DOPA PET, Paediatric gliomas, FLAIR MRI, Automated pipeline

1 Introduction

Gliomas are the most frequent central nervous system (CNS) malignancy in children, accounting for up to 35% of cases [1]. They range from low-grade (LGG) to high-grade (HGG) tumours and exhibit distinct genetic characteristics and treatment responses compared to adult gliomas [2]. Particularly, paediatric glioma stratification relies not only on their pathological characteristics but also on the presence of specific gene mutations [3].

Magnetic resonance imaging (MRI) is the gold standard for evaluating paediatric gliomas, as it offers unmatched soft tissue contrast, evaluation of vascularity, and multiparametric capabilities [4, 5]. However, positron emission tomography (PET) with [^{18}F]F-DOPA has emerged as a complementary modality, offering higher correlation with tumour grade and prognosis, particularly in infiltrative gliomas. Furthermore, [^{18}F]F-DOPA PET can improve the distinction between disease infiltration and oedema, guide the diagnostic biopsy, and help optimize the external beam radiation treatment [6, 7]. Finally, [^{18}F]F-DOPA PET has shown great potential in the non-invasive detection of the H3K27M mutation, a pivotal biomarker in diffuse midline gliomas (DMGs) [8, 9].

Tumour segmentation and clinical metric extraction remain key steps for diagnostic and prognostic evaluation. These tasks are still primarily performed manually, making them time-consuming and prone to inter-observer variability [10]. In PET analysis, static parameters such as the tumour-to-striatum (T/S) and the tumour-to-background (T/N) ratios are commonly used [11]; recently, dynamic parameters, such as the shape, the steepness of the time-activity curve (TAC), have been extracted to provide further metabolic insights [12, 13]. These tasks require high expertise, and manual execution can result in variability between observers [14, 15]. Semi-automatic and automatic methods could address these limitations [16]. A semi-automatic approach has previously been developed [17]. Still, its robustness could be further enhanced by deep learning (DL), which has transformed medical imaging tasks, including segmentation, classification, and feature extraction [18, 19]. In 2021, Futrega et al. [20] proposed an optimized U-Net framework for the brain tumour segmentation in the BraTS21 challenge [21], which was later adapted by Bianconi et al. for glioblastoma segmentation [22].

Within this context, our work aims to employ transfer learning (TL) [23] with the optimized U-Net framework (20, 22) to define a fully automated pipeline to extract static and dynamic [^{18}F]F-DOPA PET parameters from paediatric gliomas. The automatic glioma delineation from the FLAIR modality could allow for the shedding of previously employed semi-automatic methods, such as the growing-region algorithm [24]. We propose validating the fully automated pipeline by comparing the extracted static and dynamic features with their manually and semi-automatically extracted counterparts. Moreover, we investigate the correlation between static indices extracted with a fully automatic method and clinical outcomes, i.e., the grade of the neoplasms, overall survival (OS), and progression-free survival (PFS).

2 Methods

We adapted the U-Net pipeline [20] to automatically delineate paediatric gliomas from FLAIR MRI. Bianconi et al. [22] applied TL to segment glioma sub-regions in adult patients using preoperative (n = 71) and postoperative (n = 166) multimodal MRI data. They used the optimized U-Net defined by Futrega et al. [20] to classify three labels: in the preoperative setting, 1 for tumour core, 2 for enhancing tumour, and 3 for whole tumour; in the postoperative setting, 1 for oedema, 2 for enhancing tumour, and 3 for the resection cavity. We applied the DL in the preoperative scenario and adapted the U-Net to segment the whole tumour rather than sub-regions. Given the smaller size of the paediatric dataset, TL was employed to fine-tune the model to the paediatric cohort. The input/output channels were reduced from 3 to 1: 0 for background and 1 for lesion.

This replaced the semi-automatic segmentation previously performed in MRIcroGL [24]. As shown in Fig. 1, this integration allowed a fully automated workflow from FLAIR segmentation to PET parameter extraction. We did not apply the U-Net model directly to [^{18}F]F-DOPA PET images because of the heterogeneous radiotracer uptake. In rare high-grade neoplasms, uptake is minimal or absent, potentially introducing bias into the model. Instead, FLAIR MRI lesions were used to identify the affected hemisphere and served as a tumour mask in cases where neoplasms exhibited low [^{18}F]F-DOPA uptake (Fig. 1d).

We evaluated model performance using both FLAIR and FLAIR & T1. After model selection, to assess the performance of our algorithm, we compared three segmentation methods: manual, semi-automatic, and automatic. The manual method used the VCAR tool on the GE Healthcare AW server with a region-growing algorithm seeded at maximum PET uptake. Our previously defined pipeline [17] served as the semi-automatic

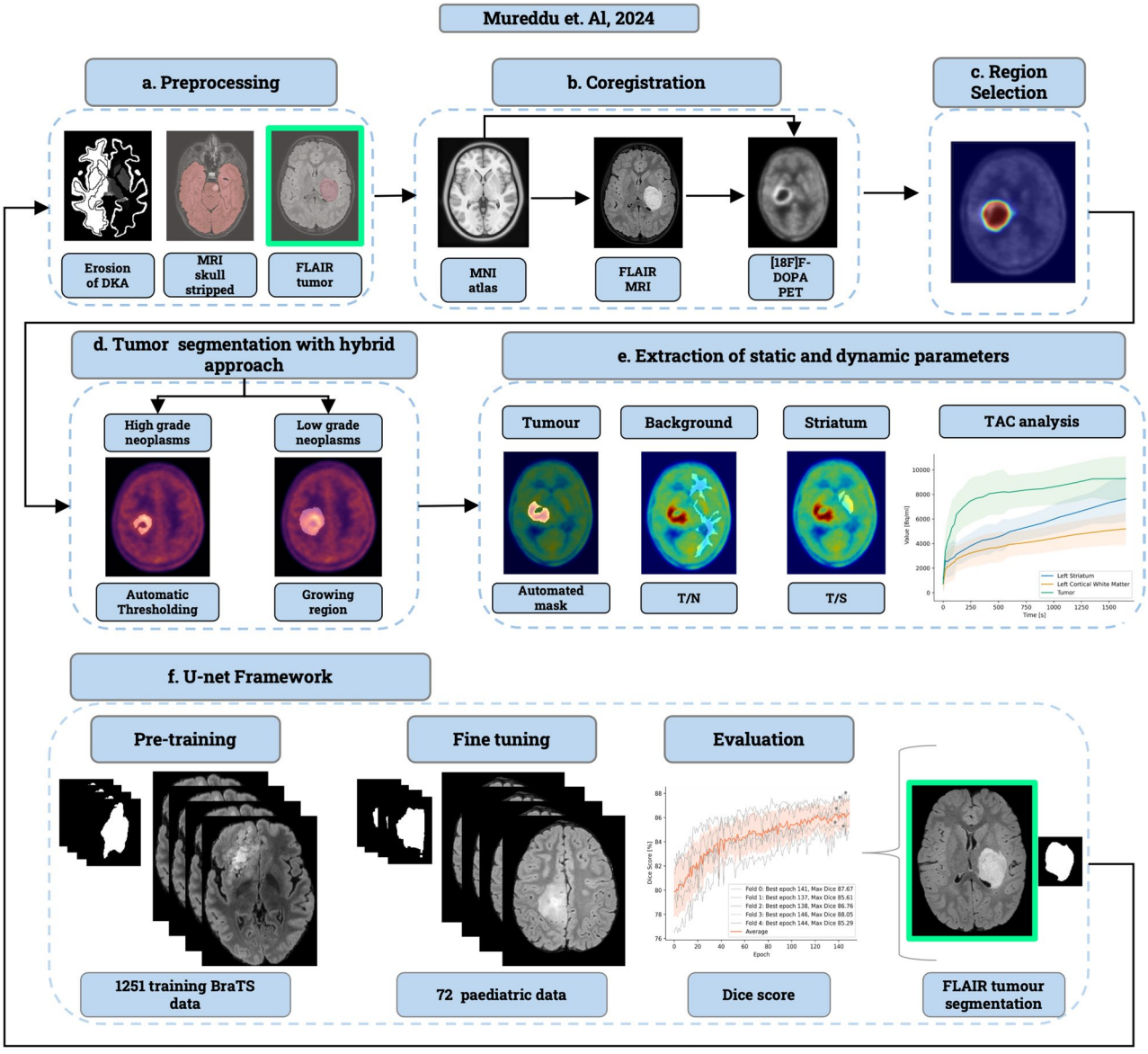

Fig. 1 Comprehensive Workflow for Automated Paediatric Brain Tumour Analysis. This figure illustrates the end-to-end pipeline for processing and analysing brain tumours in paediatric patients. **a** Preprocessing: Initial data preparation includes DKA erosion, MRI skull stripping, and identification of the tumour on FLAIR imaging. **b** Coregistration: Aligns the MNI atlas and FLAIR MRI to the coordinate space of [^{18}F]F-DOPA PET. **c** Region Selection: Selects the contralateral hemisphere for subsequent analysis. **d** Tumour Segmentation: Employs a hybrid approach using automatic thresholding for high-grade neoplasms and a growing region method for low-grade neoplasms to delineate the tumour lesion on PET images. **e** Parameter Extraction: Derives quantitative metrics, including tumour-to-background and tumour-to-striatum ratios, and performs dynamic analysis. These steps of the pipeline are defined in [17]. **f** U-Net Framework: Integrates a deep learning model, pre-trained on the BraTS dataset and fine-tuned on paediatric data, to automatically delineate the tumour on FLAIR MRI, providing a crucial input for (**a**) and subsequent pipeline analysis. This integrated approach, implemented within a GUI, streamlines the entire process from preprocessing to clinical metrics extraction

method. Finally, we investigated the clinical application of the extracted indices. Moreover, a graphical user interface (GUI) has been developed to run the complete pipeline, which is detailed in the supplementary material, and will be freely available on GitHub at the following link: https://github.com/FN-Org/GliAAns-UI and the code documentation can be seen here: https://fn-org.github.io/GliAAns-UI/.

1. Dataset

We retrospectively included paediatric patients referred to our institution (2011–2024) with newly diagnosed, treatment-naive gliomas who underwent FLAIR MRI and [^{18}F]F-DOPA PET (static and/or dynamic) within two weeks. Exclusion criteria included poor image quality, artefacts, multiple active malignancies, missing MRI, or

lack of histological confirmation. Data from 103 subjects meeting these criteria were used to develop the DL algorithm, 73 of whom had been included in a previous study [17]. For subsequent statistical and survival analysis data from 90 subjects were available. The difference in cohort sizes was due to patients being initially referred to other institutions or later transferred to different hospitals.

To pre-train the U-Net, we utilized the BraTS 2023 dataset with 1 251 labelled adult glioma cases for training and validation and an unlabelled test set of 219 patients [22, 25–27]. FLAIR and T1 MRI were consistently used in our cohort. All FLAIR images considered are without gadolinium contrast. Demographic data are detailed in Table 1. The Regional Ethics Committee approved the study (R.P. 006/2019). Written informed consent was obtained from legal guardians, with assent from patients when appropriate.

2. Imaging Protocol

[^{18}F]F-DOPA PET scans were acquired on either a Discovery ST or a Discovery MI PET/CT scanner (GE Healthcare, Milwaukee, WI, USA). PET images were obtained following bolus injection, over 90 s per bed position, covering lower extremities to the vertex. MRI was performed on 1.5 T (Intera Achieva, Philips, The Netherlands) or 3 T (Ingenia Cx, Philips) systems. All patients underwent clinical FLAIR and T1-weighted sequences. All MRI images are considered without contrast enhancement. Specifications of imaging parameters are summarized in Table E1 of the supplementary material.

3. Preprocessing and Augmentation

Since the U-net is pre-trained on the BraTS dataset, we adapted our data in a BraTS-like orientation. The main steps are the following: (i) we applied skull stripping with the Synthstrip tool [28], (ii) we co-registered the FLAIR MRI to the SRI-24 atlas (available at the following link https://www.nitrc.org/projects/sri24) to obtain isotropic voxels with isotropic spacing of 1 mm [29], and (iii) we aligned the orientation of the images in our dataset to match the BraTS orientation, i.e., Left Posterior Superior (LPS).

For co-registering all MRI volumes to the SRI-24 anatomical atlas, which is defined in the same coordinate space as the BraTS dataset, we employed the Advanced Normalization Tools (ANTs) suite. The registration pipeline involved an initial affine transformation followed by the Symmetric Normalization (SyN) deformable algorithm to ensure accurate anatomical alignment. Regarding the resampling process, trilinear interpolation was employed for the MRI. Conversely, nearest-neighbour interpolation was used for all binary lesion masks, thus preserving the binary nature of the segmentations. All images were resampled to an isotropic voxel size, matching the input resolution requirements of the pre-trained U-Net framework.

To ensure our subjects matched the LPS coordinate system used by BRATS, we employed the NiBabel library to align the orientation. To bridge the adult-paediatric domain gap and augment underrepresented small lesions (< 3000 voxels) with non-well-defined edges, we generated 20 samples from 10 subjects using five augmentation techniques. The following methods were applied: (i) flipping around the axial plane with a 50% probability, (ii) elastic deformation, (iii) adding Gaussian noise with a randomly selected standard deviation between 0 and 0.1, (iv) deforming the bias field with a coefficient of 0.5, and (v) varying the parameter gamma to apply gamma correction, to modify the contrast, with values ranging from – 0.3 to 0.3.With this process, we generated 20 new FLAIR MRIs and T1 for each subject. These transformations were applied with random parameters to get new versions of the input image. The Python library PyTorch [30] was applied for this task. In Fig. E3 of the supplementary material we report examples of augmented data.

4. Deep Learning Architecture

The U-Net code used by [20, 22] is a 3D U-Net derived from the nnU-Net framework. We employed this architecture to segment the whole tumour using FLAIR alone and with T1 weighted images. Only these two sequences were considered, as they were consistently acquired for all subjects in the cohort. Pre-training was conducted on 1251 labelled adult BraTS subjects for 150 epochs with fivefold cross-validation. To generate the whole-tumour label, all sub-region masks were merged. Training was performed on a high-performance workstation with a 2.3 GHz 4 × 24-core AMD EPYC 7352 CPU, 251 GB DDR4–3200 RAM, and a 24 GB PNY Quadro RTX 6000

Table 1 Demographics of the study cohort. We report the median age, sex, and histological diagnosis of the patients based on the updated WHO guidelines of 2021

Demographics	Paediatric cohort
Age (years)	
Median	11
Sex n (%)	
Female	54 (52.4%)
Male	49 (47.5%)
Histologic Diagnosis	
Circumscribed astrocytic gliomas	11
Glioneuronal and neuronal tumours	9
Paediatric type diffuse low-grade gliomas	25
Paediatric type diffuse high-grade gliomas	58

GPU. After pretraining, the network was fine-tuned on the paediatric dataset using TL [31].

Of the 103 patients, 72 were randomly selected for fine-tuning. The remaining 31 subjects were reserved as a blind test set, kept entirely unseen during training and utilized only after fine-tuning to select the best-performing model. The selected 72 subjects were included as input with the lesion mask obtained from the growing region of MRIcroGL.

To adapt the model, only the first two encoder layers were frozen. Training parameters, including epochs and cross-validation settings, followed the same configuration used in the BraTS pretraining phase.

5. Model selection

We selected the optimal U-Net architecture based on its performance within the independent blind test set (n = 31), which was excluded from both pre-training and fine-tuning. Segmentation performance was evaluated across five cross-validation folds, comparing models trained on FLAIR-only versus FLAIR & T1, with and without data augmentation. To quantify accuracy, we calculated Dice scores using the semi-automatic masks as reference.

Furthermore, to assess the robustness and generalisability of the framework, we performed a comparative analysis of segmentation metrics and PET-derived indices between the blind test set and the full cohort (Supplementary Tables E4 and E5).

6. Pipeline validation: comparison between automatic, semi-automatic and manual methods

We validated the automated pipeline by comparing parameters extracted from static and dynamic PET using automatic, semi-automatic, and manual methods. Static parameters were compared on a cohort of 90 subjects and dynamic data were available only for 32 subjects. The difference in cohort sizes in the statistical analysis, with respect to model training, was due to patients initially being referred to other institutions or later transferred to different hospitals. Therefore, we included all patients for whom both grade and survival data were available.

To ensure that the segmentation inference phase remained independent of the training process, we introduced data perturbations to the 72 subjects previously used for fine-tuning. This step was designed to evaluate the performance of the model on unseen variations and data. Specifically, FLAIR images were modified using: (i) adding random Gaussian noise with a standard deviation between 0 and 0.05 and (ii) elastic deformation. To verify that these perturbations did not alter the final PET results, we performed a sensitivity analysis. We compared the PET-derived indices extracted from both the original and perturbed datasets using the Wilcoxon signed-rank test across the entire cohort (n = 103), see Table E2 in the supplementary material.

Moreover, to evaluate our proposed pipeline against manual analysis we performed manual segmentation, reproducing the clinical workflow by using the region-growing algorithm implemented in the PET VCAR tool, seeded at the point of maximum uptake. This method was applied to delineate both the tumour lesion and the contralateral striatum. T/S ratios were calculated as the ratio between the maximum uptake values within the lesion and the contralateral striatum VOI. T/N ratios were obtained by normalizing the maximum uptake of the tumour lesion to that of the reference region, defined as the contralateral centrum semiovale. This reference region was delineated using a spherical VOI encompassing both grey and white matter.

In the proposed pipeline, the entire white matter was selected as the background region. We employed this comprehensive volume over the centrum semiovale to ensure a more representative sampling of background activity, giving confidence that the true maximum values are captured, for example in both tumour and background measurements.

To maintain methodological consistency and allow for direct comparison with established clinical workflow, we also performed a harmonization analysis utilizing the centrum semiovale as the background region across all three methods (see Supplementary Material, Table E3).

For dynamic PET, tumour VOIs were drawn on the final frame and projected across previous frames. For semi-automatic analysis, we used our previously defined pipeline [17]. The dynamic parameters included Time-To-Peak (TTP), defined as the time point of maximum TAC uptake, and the tumour and striatum slopes. These slopes were calculated using linear regression of mean VOI values from the second minute until the end of the acquisition. Finally, the Dynamic Slope Ratio (DSR) was derived as the ratio of the tumour slope normalized to the striatum slope.

We compared extracted indices using statistical tests including normality assessments, Wilcoxon tests, and correlation analyses. p-values derived from the Wilcoxon tests were adjusted using the Bonferroni correction to account for the three analysis methods. For each statistical analysis we also report the 95% confidence interval and the effect size (Cohen's D), which is a parameter to quantify the magnitude of the differences between groups. Moreover, we report Bland Altman plots in Fig. E5 of the supplementary material.

Table 2 Dice and Hausdorff scores for the pre-training and fine-tuning with and without data augmentation

Modality	Pre-training phases	Fine-tuning without augmentation	Fine-tuning with augmentation
Flair			
Dice	91.97 ± 0.41	91.33 ± 0.90	86.68 ± 1.22
Hausdorff	5.78 ± 0.55	6.06 ± 1.2	10.45 ± 2.26
Flair & T1			
Dice	92.496 ± 0.3415	91.608 ± 0.9423	86.458 ± 1.704
Hausdorff	5.482 ± 0.5337	5.774 ± 0.849	9.744 ± 2.380

Results are presented for models trained using only FLAIR MRI and for models trained with both FLAIR and T1 weighted MRI

7. Correlation of automatically extracted static indices with grade of neoplasms, overall survival and progression-free survival

Data from 90 subjects were employed to investigate the clinical applicability of the automatic method by assessing the correlation between T/S and T/N with clinical outcomes including the grade of the neoplasm, the OS, and the PFS.

To evaluate the correlation with grade, subjects were classified into High-Grade (HG, n = 49) and Low-Grade (LG, n = 41) groups. The normality of the distributions was tested using the Shapiro–Wilk test, and equality of variances was assessed using the Levene test. Finally, Mann–Whitney U test was applied to assess the statistical significance of the differences in static indices between HG and LG tumours.

Kaplan–Meier survival analyses were performed to investigate the relationship between T/S and T/N with OS and PFS. The time interval was defined as the number of days between the date of diagnosis and the date of the last follow-up. For OS, events were coded as 0 for alive and 1 for deceased subjects; for PFS, events were coded as 0 for stable disease and 1 for progressive disease. Subjects were dichotomized into high and low groups based on median T/S and T/N values; survival differences were then evaluated using log-rank tests.

3 Results

1. Segmentation Performance Across Validation Scenarios: FLAIR vs. Multimodal MRI, with and without Augmentation

We trained the U-Net framework using fivefold cross-validation over 150 epochs both during the pre-training and the adult & paediatric fine-tuning process. Table 2 presents the mean Dice scores and Hausdorff distances obtained in the 5-folds cross validation. We compare the FLAIR-only and FLAIR & T1 models across both the pre-training and fine-tuning phases. Results are reported for the scenarios with and without data augmentation. Furthermore, in Figs. E1 and E2 of the supplementary we illustrate the results summarized in the Table 2.

2. Segmentation Performance on Paediatric Test Set Using Fine-tuned U-Net Models with and without Data Augmentation

The reporting results were obtained using a U-Net model pre-trained on the BraTS dataset and fine-tuned with data from our paediatric cohort. Furthermore, as BraTS does not release the segmentation mask for the test set, we considered our paediatric blind-test set of 31 subjects to select the model for subsequent statistical analysis. The blind test set was completely excluded from the pre-training and fine-tuning. We calculated Dice scores to compare the masks inferred by the U-Net with those obtained via the MRIcroGL region-growing algorithm [17]. This followed the same evaluation protocol used for the validation results (Table 2). To evaluate the impact of data augmentation on fine-tuning, we compared the model without augmentation against the augmented model. This comparison assesses whether increasing data diversity improves the robustness and generalization of the model. We report the mean and standard deviation of the dice scores in Table 3.

Moreover, we selected the fine-tuned model that was trained using only FLAIR MRI with augmentation (fold 3). This model achieved the highest Dice score in the blind test set (see Table 3) and demonstrated superior performance across the full cohort (n = 103). This model obtained a Dice score of 0.79 ± 0.15, compared with 0.73 ± 0.27 for the FLAIR + T1 model (see Supplementary Material, Table E4).

In the following paragraphs, we will report the results obtained with the model fine-tuned with only FLAIR (fold 3). In Fig. 2, we present three example subjects from

Table 3 Dice scores obtained during evaluation of the model within the blind test set

Fold	FLAIR with no augmentation	FLAIR with augmentation	FLAIR & T1 with no augmentation	FLAIR & T1 with augmentation
Fold 0	0.78 ± 0.20	0.81 ± 0.14	0.78 ± 0.17	0.79 ± 0.18
Fold 1	0.79 ± 0.20	0.82 ± 0.14	0.78 ± 0.17	0.79 ± 0.17
Fold 2	0.78 ± 0.20	0.81 ± 0.14	0.78 ± 0.17	0.77 ± 0.18
Fold 3	0.8 ± 0.18	0.82 ± 0.11	0.78 ± 0.17	0.78 ± 0.19
Fold 4	0.78 ± 0.19	0.80 ± 0.13	0.78 ± 0.18	0.79 ± 0.19

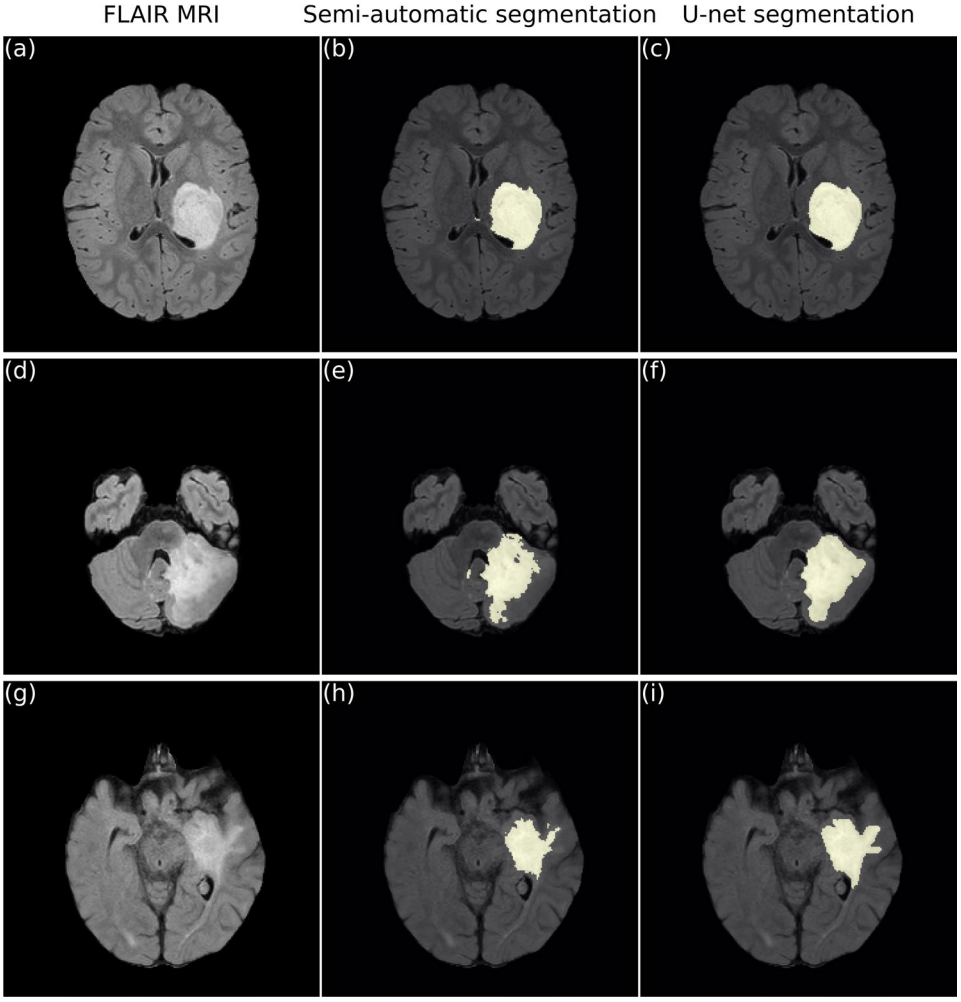

Fig. 2 Comparison of the semi-automatic and automatic segmentation. Each figure represents the axial plane of FLAIR MRI acquired with no contrast. **a**, **b** and **c** are reporting a ten-year-old male affected by glioblastoma (WHO Stage IV) within the right hemisphere. The second and third columns depict the two segmentation methods. **d**, **e** and **f**: A seventeen-year-old female affected by an intermediate-grade glial lesion (WHO Stage III) at the level of the right cerebellum. In rows **g**, **h**, and **i**, we report a one-year-old female affected by a ganglioglioma (WHO Stage I), comparing the two lesions obtained semi-automatically and automatically

the blind test set, comparing the semi-automatic lesion and the U-Net-based segmentation.

3. Agreement between PET parameters obtained with and without perturbated FLAIR MRI

The validation of the fully automated pipeline was assessed by employing the model, which had been pre-trained on BraTS data and fine-tuned with the paediatric augmented datasets (fold 3), on the entire cohort. The data of 72 subjects already used in the fine-tuning process were altered with perturbations to ensure that the same data was not fed into the trained network. Before evaluating our pipeline against semi-automatic and manual methods, we conducted a statistical analysis to verify that these perturbations did not significantly affect the extraction of PET parameters. The comparison between indices derived from original and perturbed images yielded no statistically significant differences with p-value > 0.05 for all static and dynamic parameters (Table E2 of the supplementary material).

4. Validation of the Proposed Pipeline Against Manual and Semi-automatic static and dynamic Indices

We applied the proposed pipeline to a cohort of 90 subjects and evaluated its performance by comparing the results with those obtained manually and with the semi-automatic approach. Statistical analysis for static indices was performed for all included subjects, while dynamic PET scans were available only for 32 patients.

As the Shapiro–Wilk test confirmed that none of the indices followed a normal distribution, the three extraction methods were compared using the Wilcoxon test:

(i) automatic versus semi-automatic, (ii) automatic versus manual, and (iii) semi-automatic versus manual. We applied the Bonferroni correction to the *p*-value.

In Table 4, we summarized all the results of the statistical analysis, including the confidence interval and the D effect size.

For the T/S comparison, no statistically significant differences were observed in any case, as shown in Fig. 3b (i) $p = 0.66$, (ii) $p = 0.342$, (iii) $p = 0.639$. There is a significant difference in T/N ($p < 0.005$) when comparing the manual versus semi-automatic and the manual versus automatic approaches (Fig. 3a). Within these comparisons we got (ii) bias = − 0.252 and (iii) bias = − 0.224, see Figure E5. Conversely, there is no statistical significance between the semi-automatic and automatic methods, $p = 0.207$; with Bonferroni correction, $p = 0.414$.

No statistical significance was found across all methods for T/N when the same background region, i.e. the centrum semiovale, was considered (see Table E3).

Dynamic indices were compared across the three methods in 32 subjects. As several parameters did not meet the normality assumption (Shapiro–Wilk test), comparisons were performed using the Wilcoxon signed-rank test. No significant differences were observed for striatum slope or TTP among the methods. However, Bland–Altman analysis revealed a systematic bias in TTP relative to the manual approach, with mean biases of 2.464 for the automatic method and 1.589 for the semi-automatic method (Figure 5). Significant differences between the manual and semi-automatic methods were found for tumour slope ($p = 0.008$, D = 0.18) and Dynamic Slope Ratio (DSR) ($p = 0.012$, D = − 0.22) (Table 4). Similar differences were observed when comparing the automatic and manual estimates of tumour slope and DSR. However, these effects did not remain significant after Bonferroni correction. Despite this, Bland–Altman analysis indicated residual biases of 0.246 for tumour slope and 0.695 for DSR. Figure 4 shows raincloud plots illustrating indices distributions across the three methods.

Finally, we applied the Spearman correlation to further explore the relationship among the extracted metrics. In Fig. 5, we present the scatter plots for all the extracted indices considering the three methods of analysis. The correlation coefficients are summarized in Table 4. For all the comparisons we got a p-value < 0.05.

5. Consistency of PET static indices between the blind test set and the 90-patient cohort

Table 4 Statistical results of the Wilcoxon Test applied for comparing the three methods: manual, semi-automatic and automatic

Static parameters	*p*-value	Bonferroni correction	D effect size	Confidence interval 95%	Spearman correlationcoefficient
(i) automatic vs semi-automatic					
T/S	0.660	1.0	− 0.022	[− 0.029, 0.004]	0.98
T/N	0.207	0.622	− 0.056	[− 0.054, − 0.005]	0.99
(ii) automatic vs manual					
T/S	0.342	1.0	− 0.039	[− 0.047, 0.070]	0.93
T/N	$p < 0.01$	$p < 0.01$	− 0.460	[− 0.300, − 0.200]	0.92
(iii) manual vs semi-automatic					
T/S	0.639	1.0	− 0.016	[− 0.033, 0.0170]	0.96
T/N	$p < 0.01$	$p < 0.01$	− 0.410	[− 0.270, − 0.180]	0.93
Dynamic parameters	***p*-value**	**Bonferroni correction**	**D effect size**	**Confidence interval**	**Spearman correlation coefficient**
(i) automatic vs semi-automatic					
Tumour Slope	0.711	1.0	− 0.021	[− 0.074, 0.016]	0.99
Striatum Slope	0.860	1.0	0.002	[− 0.024, 0.031]	0.99
Time To Peak	0.500	1.0	0.08	[− 49.68, 168.75]	0.88
Dynamic Slope Ratio	0.900	1.0	0.03	[− 0.150, 0.400]	0.98
(ii) automatic vs manual					
Tumour Slope	0.017	0.050	0.18	[− 0.090, 0.420]	0.88
Striatum Slope	0.490	1.0	− 0.08	[− 0.170, 0.040]	0.95
Time To Peak	0.068	0.20	0.21	[− 6.290, 300.625]	0.73
Dynamic Slope Ratio	0.018	0.056	0.25	[0.050, 1.650]	0.83
(iii) manual vs semi-automatic					
Tumour Slope	0.008	0.023	− 0.20	[− 0.440, − 0.107]	0.88
Striatum Slope	0.612	1.0	0.08	[− 0.030, 0.180]	0.95
Time To Peak	0.237	0.71	− 0.14	[− 276.65, 84.37]	0.66
Dynamic Slope Ratio	0.012	0.04	− 0.22	[− 1.310, − 0.130]	0.80

We report the results for both the static and dynamic indices. We also report for each comparison the D effect size, the confidence interval and the Spearman correlation coefficient

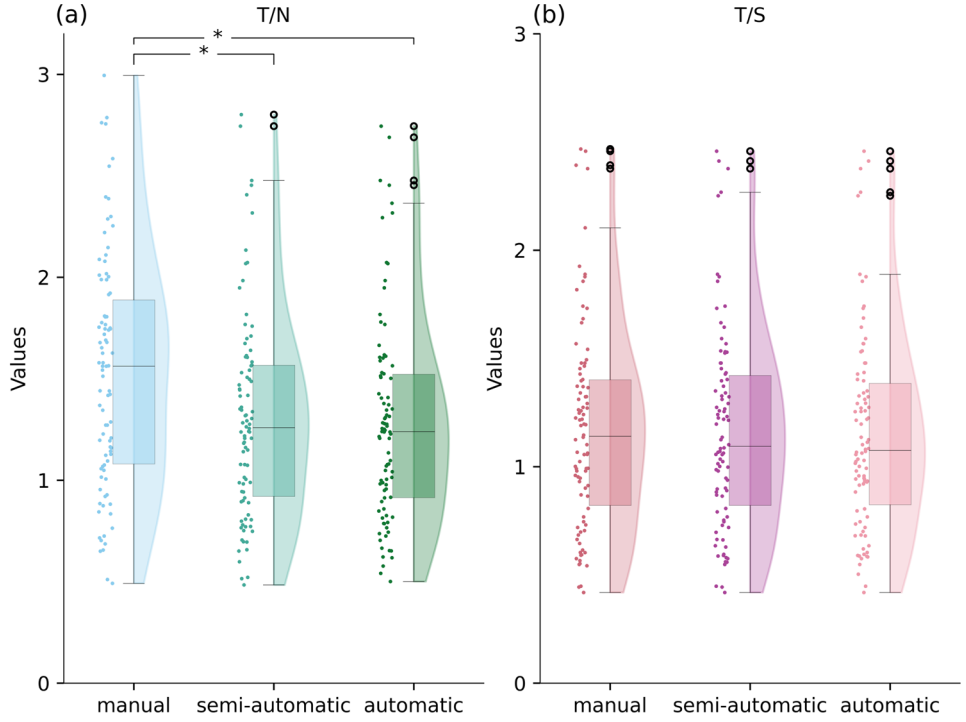

Fig. 3 Distribution of Static [^{18}F]F-DOPA PET Indices. Raincloud plots showing the distribution of **a** T/N and **b** T/S ratios across manual, semi-automatic, and automatic methods. Statistically significant differences ($p < 0.05$) were observed between the manual reference and both the semi-automatic and automatic pipelines for T/N ratios, as indicated by the '*'. No significant differences were found for T/S ratios across the three methods. Black dots denote outliers

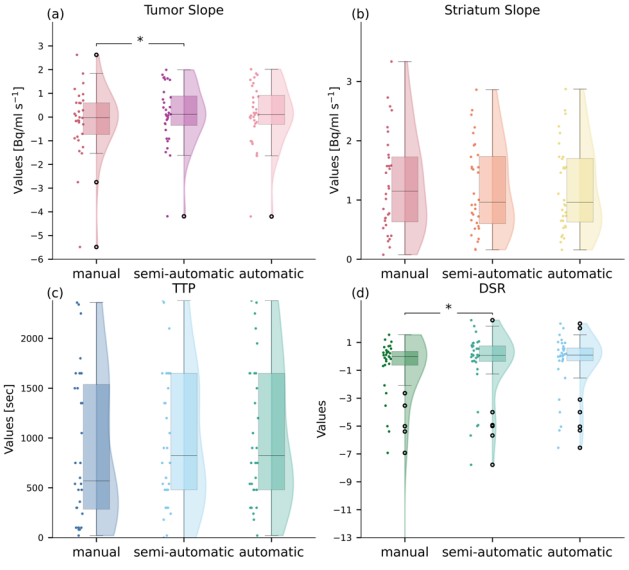

Fig. 4 Distribution of Dynamic [^{18}F]F-DOPA PET Metrics. Raincloud plots comparing **a** Tumour Slope, **b** Striatum Slope, **c** Time-to-Peak (TTP), and **d** Dynamic Slope Ratio (DSR) across manual, semi-automatic, and automatic methods. Statistically significant differences ($p < 0.05$) were observed between the manual and semi-automatic methods for Tumour Slope and DSR, as indicated the '*'. No significant differences were found across the three approaches for Striatum Slope or TTP. Black dots denote outliers

Table E5 (Supplementary Material) reports the Wilcoxon test results comparing T/S and T/N values obtained from the 90-subject cohort and the independent blind test set (n = 31). We obtained no significant differences for T/S across all methods, while T/N yielded significant differences ($p < 0.05$) when considering the manual method.

6. Investigation of clinical utility of the static parameters extracted with the automatic pipeline

The T/S and T/N values extracted from 90 subjects were classified according to tumour grade (HG vs LG). Since the data did not meet the assumptions of normality, the non-parametric Mann–Whitney U test was applied in all cases. For T/S, we obtained a p-value < 0.05 and an effect size (D) of −1.05, while for T/N we found a p-value < 0.05 with D equal to − 0.79. The corresponding raincloud plots of these distributions are shown in Fig. 6a.

Over the same cohort, we studied the correlation between T/S and T/N and survival outcomes. To define the high and low T/S and T/N groups, we considered the median values as a cut-off. The median T/S is 1.08 and the median value for T/N is 1.24. The log-rank test resulted in $p < 0.05$ between the survival curves of high and low groups of both T/S and T/N. The survival curves are represented in Fig. 6b and d for OS and PFS based on

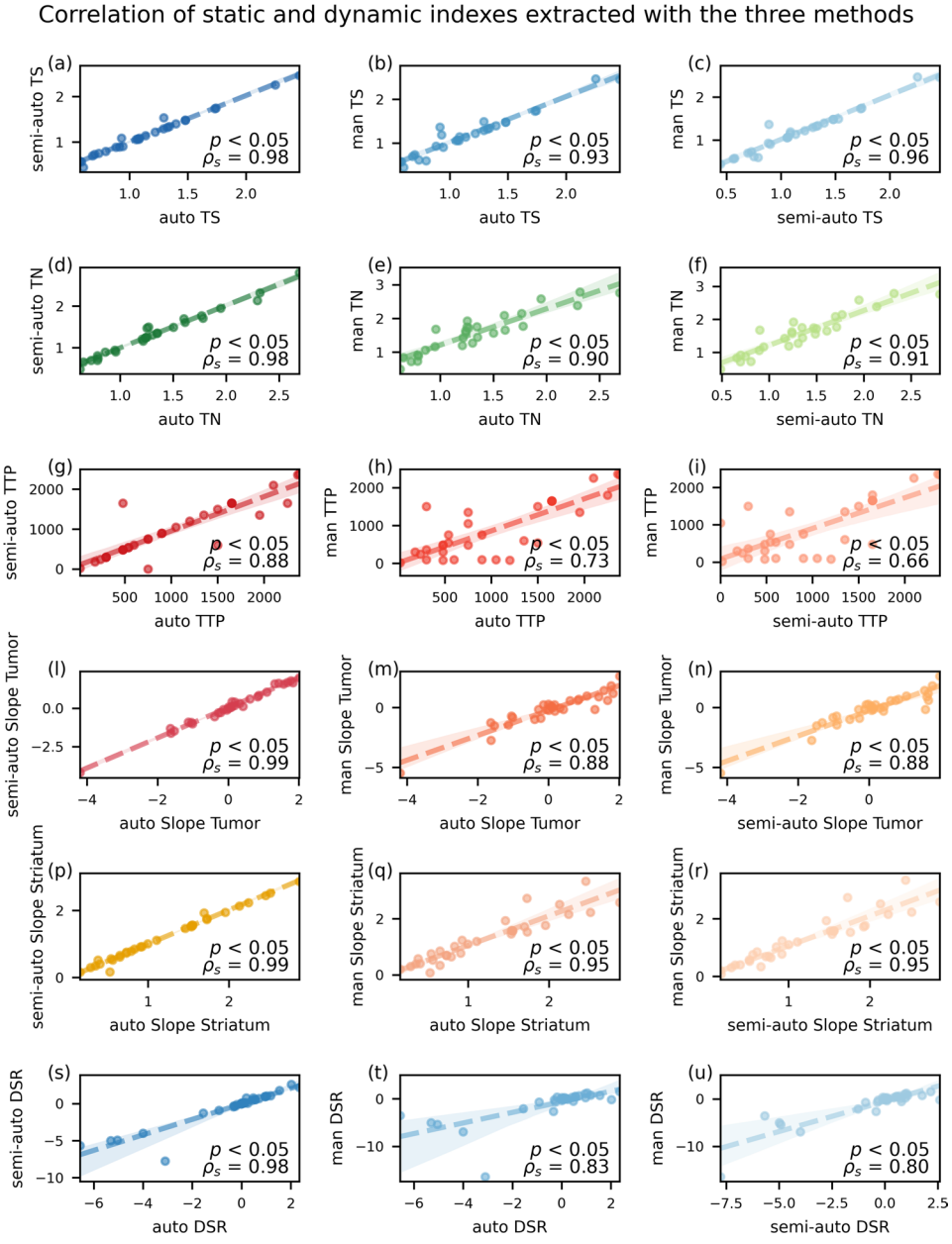

Fig. 5 Correlation analysis across extraction methods. Scatter plots displaying Spearman correlation coefficients (ρ) and associated *p*-values for manual, semi-automatic, and automatic pipelines. Rows: **a**–**c** T/S; **d**–**f** T/N; **g**–**i** Time-To-Peak; **l**–**n** Tumour Slope; **p**–**r** Striatum Slope; and **s**–**u** Dynamic Slope Ratio. First Column: semi-automatic vs. automatic; Second columns: manual vs. automatic; and third column: manual vs. semi-automatic. All comparisons yielded statistically significant results ($p < 0.05$), with values ranging from 0.66 to 0.99

T/S, and in Fig. 6c and e for OS and PFS based on T/N, respectively.

4 Discussion

Several works have demonstrated the applicability of transfer learning to segment paediatric brain tumours using adult data. Liu et al. [32] showed that pre-training DL models on open-source adult glioma datasets like BraTS significantly enhances segmentation performance on rare paediatric tumours such as DMG, achieving a dice score of up to 0.880 ± 0.072 using nnU-net over 1000 epochs. Similarly, Boyd et al. [33] proposed an in-domain transfer learning framework for pLGGs, reaching a median dice similarity coefficient (DSC) of 0.877 and outperforming human experts in blinded testing. Ladefoged et al. [34] applied transfer learning with artificial neural networks for [^{18}F]FET PET/MRI-based delineation of paediatric CNS tumours, obtaining high agreement with manual annotations and robust clinical tracking (median Dice score 0.93). These findings highlight how adult

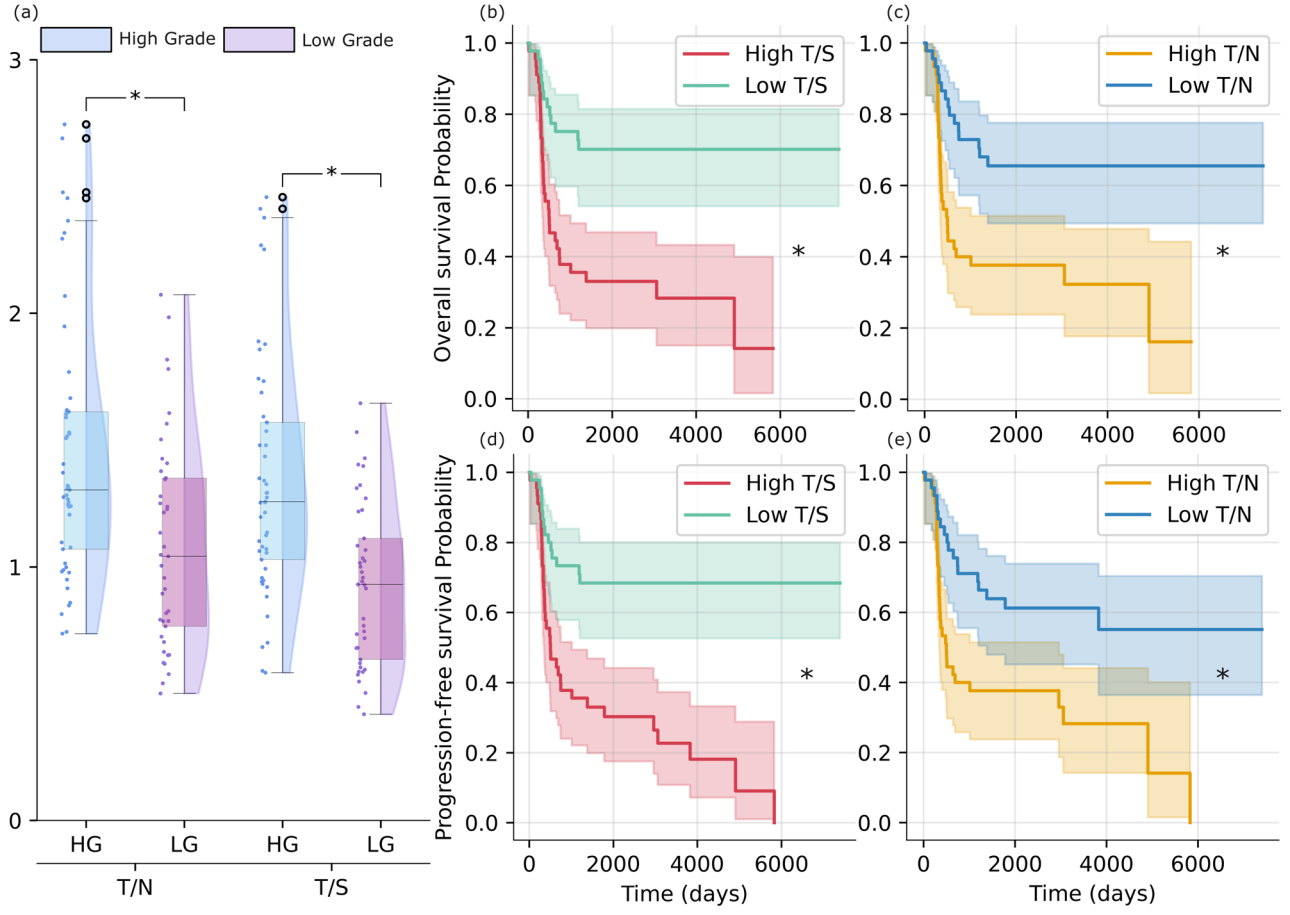

Fig. 6 Clinical Validation of Automated Static [^{18}F]F-DOPA PET Indices. **a** Distribution of T/N and T/S ratios across High-Grade and Low-Grade gliomas. Both metrics demonstrate statistically significant differences between tumour grades ($p < 0.05$). **b**–**e** Kaplan–Meier curves evaluating the prognostic value of T/S and T/N for (**b**), **c** Overall Survival (OS) and **d**, **e** Progression-Free Survival. Subjects were stratified using median values (T/S = 1.08; T/N = 1.24). All comparisons indicate significant associations between automated PET indices and clinical outcomes ($p < 0.05$), as indicated by '*'

data can support segmentation tasks in rare paediatric tumours, overcoming data scarcity.

Our study evaluated the feasibility of a U-Net-based model for automatic glioma segmentation on FLAIR MRI in 103 paediatric patients, aiming to introduce an end-to-end pipeline for [^{18}F]F-DOPA PET analysis.

We pre-trained the U-Net model by employing BRATS data from the 2023 BRATS challenge. For fine-tuning we employed a cohort 72 paediatric patient and an augmented cohort of 272 paediatric patient. Our input to the U-net were the clinical images and segmented mask obtained with the growing region of MRIcroGL.

A critical observation emerged during the fine-tuning and testing phase. Fine-tuning with augmentation appears to yield lower Dice scores and higher Hausdorff scores than fine-tuning without augmentation. However, the results of the blind test set in Table 3 suggest an improvement when augmentation is applied. As reported in Fig. E1c and d of the supplementary material, Dice scores and Hausdorff distances, despite performing better for the model trained without augmentation, show a flat trend over 150 epochs compared to the pre-training phase. This pattern suggests that during the fine-tuning process the weights were not updated, thereby preventing further adaptation to the specificities of paediatric tumours. Consequently, while these models achieved high metrics during internal cross-validation (Table 2), their performance dropped on the blind test set (Table 3), indicating the presence of overfitting to the validation cohort.

Data augmentation was specifically applied to shift data distribution towards more complex scenarios, e.g., tumours with small, poorly demarcated tumours. As illustrated in Fig. E1e and f, this approach transformed the training dynamics: the learning curves showed an active optimization trend, with Dice scores steadily increasing and Hausdorff distances decreasing. Although the increased complexity of the augmented dataset led to apparently lower cross-validation metrics (Table 2), it acted as a robust regulariser. By forcing the model to learn more generalised morphological features rather than memorizing cohort-specific patterns,

the augmentation effectively bridged the gap between the adult BRATS cohort and the paediatric domain.

This is confirmed by the blind test set results (Table 3), where the augmented FLAIR-only model was the only one consistently exceeding a Dice score of 0.8 (0.81 ± 0.13), aligning with Boyd et al.'s findings. In contrast, all models fine-tuned on the original 72-patient cohort without augmentation scored below 0.80, failing to generalise to unseen subjects.

Interestingly, the model trained with both FLAIR and T1 on the augmented dataset followed a similar trend in the fine-tuning process but achieved a lower average Dice score (0.78 ± 0.18) in the blind test set.

This suggests that despite augmentation, increasing the number of input modalities within a limited sample size made the multi-modal approach more prone to overfitting compared to the FLAIR-only model. The addition of T1-weighted images might have introduced redundant information that limited the generalisability in the blind test set. Moreover, the superior performance of the FLAIR-only model highlights the critical role of this sequence in paediatric tumour segmentation, where neoplasms typically exhibit marked hyperintensity compared to T1-weighted images.

After model selection, we investigated the clinical utility of proposed pipeline by applying the DL across the full cohort ($n = 103$).

We applied perturbations (rotations, Gaussian noise, and elastic deformation) to include the whole cohort without passing the exact same data already used in the fine-tuning process. Furthermore, we demonstrated that perturbations did not affect PET parameters extraction (Table E2 of the supplementary material). Importantly, the selected model demonstrated consistent performance across both the independent blind test set and the full patient cohort ($n = 103$, $n = 90$), in terms of lesion mask inference and PET static indices (Tables E4 and E5 of the supplementary material). These findings support the robustness and generalisability of the proposed approach within this monocentric paediatric cohort.

Statistical and survival analysis were performed over a cohort of 90 subjects due to the availability of survival data. The comparison of static indices across the three segmentation methods demonstrated a good overall consistency. T/S showed no significant differences across any comparison ($p > 0.3$ in all cases) and high correlation coefficients ($\rho \geq 0.93$), confirming the robustness of this parameter and its low dependence on the segmentation approach. A strong agreement was observed also for T/N between semi-automatic and automatic methods. In contrast, the T/N ratio differed significantly in comparisons considering the manual case scenario ($p < 0.01$) with moderate effect sizes ($D = -0.46$ and $D = -0.41$). These discrepancies were attributed to the different background region: while the clinical workflow utilized the centrum semiovale, the automated and semi-automated methods employed the entire white matter. This observation was further supported by comparing T/N and T/S when the centrum semiovale was used consistently as the background VOI across all methods. The statistical differences previously observed were no longer evident under these conditions (see Supplementary Table E3 and Fig. E2). This result confirms the robustness of our tumour segmentation approach and that the observed variability was primarily driven by the background selection. Although absolute T/N values were affected by the choice of background VOI, correlation analysis demonstrated strong agreement across methods ($\rho \geq 0.91$). These findings support the reliability of the automated pipeline in identifying tumour regions and metabolic peaks. Furthermore, the high correlation between automated and manual indices indicates that, despite shifts in absolute values, as shown in the Bland Altman plots (Fig. E5 panels e and f), patient ranking remained stable, thereby preserving the clinical utility of the approach.

The automatic T/S and T/N showed clinical significance when correlated with tumour grade and survival outcomes. The non-parametric Mann–Whitney U test revealed highly significant differences between high-grade and low-grade gliomas for both indices. Furthermore, survival analysis demonstrated that both T/S and T/N significantly discriminated between high- and low-risk groups for OS and PFS. Collectively, these findings support the reliability of the proposed pipeline for quantitative [^{18}F]F-DOPA PET static analysis in paediatric gliomas. Information on the aggressiveness and prognosis of paediatric gliomas is indeed key in charting the correct therapeutic course; molecular imaging allows for the non-invasive acquisition of this information across the entire tumour volume and over repeated examinations.

Dynamic analysis results, though limited to 32 patients due to scan availability, are reported to demonstrate the feasibility of the fully automatic pipeline approach. No significant differences were observed for TTP and striatum slope across any comparison ($p > 0.2$), with ρ exceeding 0.9 across all methods for the striatum slope. TTP between manual and semi-automatic methods resulted with the lowest correlation of $\rho = 0.66$. Statistically significant differences were found for tumour slope and DSR when comparing manual and semi-automatic methods. However, the corresponding effect sizes were small, indicating limited practical relevance.

Comparing the manual and the automatic method, tumour slope and DSR showed higher correlations ($\rho = 0.88$ and $\rho = 0.83$, respectively). However, after Bonferroni correction, the resulting p-values were just above the threshold of statistical significance (p-value $= 0.050$ and p-value $= 0.056$, respectively).

The Bland–Altman analysis (Fig. E5) indicates that dynamic parameters are particularly sensitive to variations in tumour delineation. No systematic bias was observed between the semi-automatic and automatic methods. However, TTP, DSR, and tumour slope showed greater bias relative to the manual method (Fig. E5). This increased dispersion is likely attributable to the inherently lower anatomical detail of dynamic PET images, which makes manual segmentation especially challenging. Even small voxel-level differences in tumour delineation can modify the shape of TACs, thereby affecting the derived dynamic metrics. Consequently, although dynamic parameters demonstrated high correlations across methods, their susceptibility to segmentation variability, together with the limited number of dynamic datasets available, constrains the generalisability of these findings. These observations warrant validation in a larger cohort.

While the findings are encouraging, several limitations must be considered. The primary constraint is the relatively small cohort size, a common challenge in paediatric neuro-oncology due to the rarity of these tumours. While we mitigated this through data augmentation techniques and transfer learning from the BraTS dataset, the limited sample size prevented a robust stratification analysis.

Moreover, most scans were performed on the newer PET system; therefore, a subgroup analysis would have been statistically precluded by the substantial imbalance in sample sizes. Further subdivision of the sample (from 103 to 90 subjects for static metrics, and even fewer for the dynamic subgroup) would have significantly reduced statistical power, potentially leading to unreliable results. Nevertheless, we expect the impact of scanner-related differences to be mitigated by our preprocessing and PET parameters extraction, which relies on ratio-based indices rather than absolute values.

The limitation related to sample size is particularly evident in the statistical analysis of 4D PET data, and the results concerning dynamic parameters should therefore be considered preliminary. Owing to the restricted subset of patients with dynamic acquisitions, these findings should be considered exploratory rather than a definitive validation.

Lastly, in our work we employ semi-automatic masks, generated via MRIcroGL and expert-reviewed, rather than purely manual segmentations. This choice was dictated by the practical complexity of manual delineation in this specific clinical context and the absence of pre-existing manual ground truths for this dataset. While these masks provided a consistent reference for fine-tuning and evaluation, we acknowledge that manual segmentation remains the theoretical gold standard for calculating dice scores.

Future efforts must focus on expanding the dataset through multicentric collaborations with larger and more diverse cohorts to confirm the generalisability of our pipeline. This will be fundamental in performing robust longitudinal studies for therapy monitoring and fully establish the clinical added value of dynamic PET acquisitions in the paediatric population. Additionally, integrating the framework with multiparametric MRI sequences could further enhance its diagnostic and prognostic accuracy.

Overall, our findings present a robust pipeline for the automated segmentation and metric extraction of [^{18}F]F-DOPA PET in paediatric gliomas. While the dynamic analysis remains exploratory, the fully automatic method demonstrates significant potential for extracting static indices, offering a more reproducible and streamlined clinical workflow for paediatric neuro-oncological assessments.

Abbreviation

[^{18}F]F DOPA 18 Fluoro Di-Oxy-Phenylalanine

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40708-026-00296-z.

Supplementary Material 1

Acknowledgements

This work was supported by the Italian Ministry of Health and by Regione Liguria "Ricerca Finalizzata di Rete NET-2019-12371188 GLI-HOPE.

Author contributions

MM, RT, and MMF conceptualized the study. Data were curated by AB, AR (Andrea Rossi), AR (Antonia Ramaglia), AV, CM, GM, MI, FF, and AP. Formal analysis was performed by MM, RT, FGG, and NT. Funding acquisition was provided by AP. The investigation was conducted by MM, RT, FGG, and NT. Resources were provided by AB, AR (Andrea Rossi), AR (Antonia Ramaglia), AV, CM, GM, MI, FF, and AP. The methodology was developed by MM, RT, FGG, NT, and MMF. Project administration was carried out by MMF. Software development, validation, and visualization were performed by MM, RT, FGG, and NT. Writing—original draft preparation was carried out by MM, RT, FGG, and NT. Writing—review and editing were performed by MM, RT, FGG, NT, AB, AR (Andrea Rossi), AR (Antonia Ramaglia), AV, CM, GM, MI, FF, AP, and MMF. All authors read and approved the final manuscript.

Funding

This study received funding from the project GLI-HOPE NET-2019-12371188.

Data availability

The dataset supporting the conclusions of this article is not publicly available due to privacy restrictions of clinical data imposed by the Galliera Hospital's administration but is available from the corresponding author on reasonable request. The code for our pipeline is available at [https://github.com/MicheleMureddu/Pediatric_fdopa_pipeline] (https://github.com/MicheleMureddu/Pediatric_fdopa_pipeline) . The GUI is available here: [https://github.com/FN-Org/GliAAns-UI] (https:/github.com/FN-Org/GliAAns-UI) with related documentation: [https://fn-org.github.io/GliAAns-UI/] (https://fn-org.github.io/GliAAns-UI/) Current software has been designed and tested under Python 3.11, the U-NET has been tested with high-performance workstation with a 2.3 GHz 4×24-core AMD EPYC 7352 CPU, 251 GB DDR4–3200 RAM, and a 24 GB PNY Quadro RTX 6000 GPU. The Graphical User Interface will be available on the same page of the pipeline code. Data from the BRATS dataset, employed

in this work to pre-trained the U-Net model are available here: [https://www.synapse.org/Synapse:syn51156910/files/] (https://www.synapse.org/Synapse:syn51156910/files).

Declarations

Ethics approval and consent to participate

The Regional Ethics Committee approved the study (R.P. 006/2019). Written informed consent was obtained from legal guardians, with assent from patients when appropriate.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Via all'Opera Pia 13, 16145 Genoa, Italy
[2]U.O. Clinica Neurochirurgica e Neurotraumatologica, IRCCS Ospedale Policlinico San Martino, Largo Rosanna Benzi 10, 16132 Genoa, Italy
[3]NeuroRadiology Unit, IRCCS Istituto Giannina Gaslini, Via Gerolamo Gaslini 5, 16147 Genoa, Italy
[4]Neuro-Oncology Unit, IRCCS Istituto Giannina Gaslini, Via Gerolamo Gaslini 5, 16147 Genoa, Italy
[5]Department of Neurosciences, University of Turin, Via Cherasco 15, 10126 Turin, Italy
[6]Nuclear Medicine Unit, Ente Ospedaliero Ospedali Galliera, Mura delle Cappuccine 14, 16128 Genoa, Italy

Received: 13 November 2025 / Accepted: 6 March 2026
Published online: 25 March 2026

References

1. Hauser P (2021) Classification and treatment of pediatric gliomas in the molecular era. Children 8:739. https://doi.org/10.3390/children8090739
2. Aggarwal P et al (2022) Pediatric versus adult high grade glioma: immunotherapeutic and genomic considerations. Front Immunol 13:1038096. https://doi.org/10.3389/fimmu.2022.1038096
3. Louis DN et al (2021) The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro Oncol 23:1231–1251. https://doi.org/10.1093/neuonc/noab106
4. Leach JL et al (2020) MR imaging features of diffuse intrinsic pontine glioma and relationship to overall survival: report from the International DIPG Registry. Neuro Oncol 22:1647–1657. https://doi.org/10.1093/neuonc/noaa140
5. Broniscer A, Gajjar A (2004) Supratentorial high-grade astrocytoma and diffuse brainstem glioma: two challenges for the pediatric oncologist. Oncologist 9:197–206. https://doi.org/10.1634/theoncologist.9-2-197
6. Morana G et al (2014) Value of 18F–3,4-dihydroxyphenylalanine PET/MR image fusion in pediatric supratentorial infiltrative astrocytomas: a prospective pilot study. J Nucl Med 55:718–723. https://doi.org/10.2967/jnumed.113.125500
7. Misch M et al (2015) 18F-FET-PET guided surgical biopsy and resection in children and adolescents with brain tumors. Childs Nerv Syst 31:261–267. https://doi.org/10.1007/s00381-014-2552-y
8. Piccardo A et al (2019) Advanced MR imaging and 18F-DOPA PET characteristics of H3K27M-mutant and wild-type pediatric diffuse midline gliomas. Eur J Nucl Med Mol Imaging 46:1685–1694. https://doi.org/10.1007/s00259-019-04333-4
9. Morana G et al (2020) Correlation of multimodal 18F-DOPA PET and conventional MRI with treatment response and survival in children with diffuse intrinsic pontine gliomas. Theranostics 10:11881–11891. https://doi.org/10.7150/thno.50598
10. Di Ieva A et al (2021) Application of deep learning for automatic segmentation of brain tumors on magnetic resonance imaging: a heuristic approach in the clinical scenario. Neuroradiology 63:1253–1262. https://doi.org/10.1007/s00234-021-02649-3
11. Piccardo A et al (2022) Joint EANM/SIOPE/RAPNO practice guidelines/SNMMI procedure standards for imaging of paediatric gliomas using PET with radiolabelled amino acids and [18F]FDG: version 1.0. Eur J Nucl Med Mol Imaging 49:3852–3869. https://doi.org/10.1007/s00259-022-05817-6
12. Fiz F et al (2023) Diagnostic and dosimetry features of [64Cu]CuCl2 in high-grade paediatric infiltrative gliomas. Mol Imaging Biol 25:391–400. https://doi.org/10.1007/s11307-022-01769-3
13. Ginet M et al (2020) Integration of dynamic parameters in the analysis of 18F-FDopa PET imaging improves the prediction of molecular features of gliomas. Eur J Nucl Med Mol Imaging 47:1381–1390. https://doi.org/10.1007/s00259-019-04509-y
14. Kubben PL, Postma AA, Kessels AGH, van Overbeeke JJ, van Santbrink H (2010) Intraobserver and interobserver agreement in volumetric assessment of glioblastoma multiforme resection. Neurosurgery 67:1329. https://doi.org/10.1227/NEU.0b013e3181efbb08
15. Grabowski MM et al (2014) Residual tumor volume versus extent of resection: predictors of survival after surgery for glioblastoma. J Neurosurg 121:1115–1123. https://doi.org/10.3171/2014.7.JNS132449
16. Ghaffari M, Sowmya A, Oliver R (2020) Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the BraTS 2012–2018 challenges. IEEE Rev Biomed Eng 13:156–168. https://doi.org/10.1109/RBME.2019.2946868
17. Mureddu M et al (2024) A new tool for extracting static and dynamic parameters from [18F]F-DOPA PET/CT in pediatric gliomas. J Clin Med 13:6252. https://doi.org/10.3390/jcm13206252
18. Jan M, Spangaro A, Lenartowicz M, Mattiazzi Usaj M (2024) From pixels to insights: machine learning and deep learning for bioimage analysis. BioEssays 46:2300114. https://doi.org/10.1002/bies.202300114
19. Shen D, Wu G, Suk HI (2017) Deep learning in medical image analysis. Annu Rev Biomed Eng 19:221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442
20. Futrega M, Milesi A, Marcinkiewicz M, Ribalta P (2021) Optimized U-Net for brain tumor segmentation. *arXiv preprint* arXiv:2110.03352. https://doi.org/10.48550/arXiv.2110.03352
21. Baid U et al. (2021) The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint* arXiv:2107.02314. https://doi.org/10.48550/arXiv.2107.02314
22. Bianconi A et al (2023) Deep learning-based algorithm for postoperative glioblastoma MRI segmentation: a promising new tool for tumor burden assessment. Brain Inform 10:26. https://doi.org/10.1186/s40708-023-00207-6
23. Tataei Sarshar N et al (2023) Glioma brain tumor segmentation in four MRI modalities using a convolutional neural network and based on a transfer learning method. In: Iano Y, Saotome O, Kemper Vásquez GL, Cotrim Pezzuto C, Arthur R, de Gomes Oliveira G (eds) Proceedings of the 7th Brazilian Technology Symposium (BTSym'21). Springer, Cham, pp 386–402. https://doi.org/10.1007/978-3-031-04435-9_39
24. Rorden C (2025) From MRIcro to MRIcron: the evolution of neuroimaging visualization tools. Neuropsychologia 207:109067. https://doi.org/10.1016/j.neuropsychologia.2025.109067
25. Bakas S et al (2017) Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 4:170117. https://doi.org/10.1038/sdata.2017.117
26. Menze BH et al (2015) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34:1993–2024. https://doi.org/10.1109/TMI.2014.2377694
27. Bakas S et al (2017) Segmentation labels for the pre-operative scans of the TCGA-GBM collection. Cancer Imaging Arch. https://doi.org/10.7937/k9/tcia.2017.klxwjj1q
28. Hoopes A, Mora JS, Dalca AV, Fischl B, Hoffmann M (2022) SynthStrip: skull-stripping for any brain image. Neuroimage 260:119474. https://doi.org/10.1016/j.neuroimage.2022.119474
29. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A (2010) The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp 31:798–819. https://doi.org/10.1002/hbm.20906
30. Pérez-García F, Sparks R, Ourselin S (2021) TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput Methods Programs Biomed 208:106236. https://doi.org/10.1016/j.cmpb.2021.106236
31. Hadjiiski L et al (2023) AAPM task group report 273: recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Med Phys 50:e1–e24. https://doi.org/10.1002/mp.16188
32. Liu X et al. (2023) Automatic segmentation of rare pediatric brain tumors using knowledge transfer from adult data. In: 2023 IEEE 20th international

symposium on biomedical imaging (ISBI), pp 1–4. https://doi.org/10.1109/ISBI53787.2023.10230757
33. Boyd A et al (2023) Expert-level pediatric brain tumor segmentation in a limited data scenario with stepwise transfer learning. medRxiv. https://doi.org/10.1101/2023.06.29.23292048
34. Ladefoged CN et al (2022) Automatic detection and delineation of pediatric gliomas on combined [18F]FET PET and MRI. Front Nucl Med 2:960820. https://doi.org/10.3389/fnume.2022.960820