

Improving Deep Learning Models for Pediatric Low-Grade Glioma Tumours Molecular Subtype Identification Using MRI-based 3D Probability Distributions of Tumour Location

Canadian Association of
Radiologists Journal
2025, Vol. 76(2) 313–323
© The Author(s) 2024



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08465371241296834
journals.sagepub.com/home/caj



Khashayar Namdar^{1,2,3,4}, Matthias W. Wagner^{1,5,6} ,
Kareem Kudus^{1,2,3}, Cynthia Hawkins⁷, Uri Tabori⁸,
Birgit B. Ertl-Wagner^{1,2,3,5}, and Farzad Khalvati^{1,2,3,4,5,9,10}

Abstract

Purpose: Pediatric low-grade gliomas (pLGG) are the most common brain tumour in children, and the molecular diagnosis of pLGG enables targeted treatment. We use MRI-based Convolutional Neural Networks (CNNs) for molecular subtype identification of pLGG and augment the models using tumour location probability maps. **Materials and Methods:** MRI FLAIR sequences of 214 patients (110 male, mean age of 8.54 years, 143 BRAF fused and 71 BRAF V600E mutated pLGG tumours) from January 2000 to December 2018 were included in this retrospective REB-approved study. Tumour segmentations (volumes of interest—VOIs) were provided by a pediatric neuroradiology fellow and verified by a pediatric neuroradiologist. Patients were randomly split into development and test sets with an 80/20 ratio. The 3D binary VOI masks for each class in the development set were combined to derive the probability density functions of tumour location. Three pipelines for molecular diagnosis of pLGG were developed: location-based, CNN-based, and hybrid. The experiment was repeated 100 times each with different model initializations and data splits, and the Areas Under the Receiver Operating Characteristic Curve (AUROC) was calculated, and Student's *t*-test was conducted. **Results:** The location-based classifier achieved an AUROC of 77.9, 95% confidence interval (CI) (76.8, 79.0). CNN-based classifiers achieved an AUROC of 86.1, 95% CI (85.0, 87.3), while the tumour-location-guided CNNs outperformed the other classifiers with an average AUROC of 88.64, 95% CI (87.6, 89.7), which was statistically significant (*P*-value .0018). **Conclusion:** Incorporating tumour location probability maps into CNN models led to significant improvements for molecular subtype identification of pLGG.

Résumé

Objectif : Le gliome de bas grade est le type de tumeur au cerveau le plus courant chez l'enfant. Son diagnostic moléculaire permet un traitement ciblé. Nous avons recours à des réseaux neuronaux convolutifs (CNN) afin de déterminer le sous-type moléculaire de gliome à partir d'une IRM, et nous améliorons les modèles au moyen de cartes de probabilité de localisation des tumeurs. **Matériel et méthodes :** Des séquences d'IRM de type FLAIR (*fluid-attenuated inversion recovery*) de 214 patients (110

¹ Division of Neuroradiology, Department of Diagnostic & Interventional Radiology, The Hospital for Sick Children (SickKids), Toronto, ON, Canada

² Neurosciences & Mental Health Research Program, SickKids Research Institute, Toronto, ON, Canada

³ Institute of Medical Science, University of Toronto, Toronto, ON, Canada

⁴ Vector Institute, Toronto, ON, Canada

⁵ Department of Medical Imaging, University of Toronto, Toronto, ON, Canada

⁶ Department of Diagnostic and Interventional Neuroradiology, University Hospital Augsburg, Augsburg, Germany

⁷ Department of Paediatric Laboratory Medicine, Division of Pathology, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada

⁸ Department of Neurooncology, The Hospital for Sick Children, Toronto, ON, Canada

⁹ Department of Computer Science, University of Toronto, Toronto, ON, Canada

¹⁰ Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

Corresponding Author:

Farzad Khalvati, Department of Medical Imaging, University of Toronto, 686 Bay Street, Toronto, ON M5G 0A4, Canada.

Email: farzad.khalvati@utoronto.ca

hommes, âge moyen de 8,54 ans; 143 gliomes porteurs d'une fusion de BRAF et 71 gliomes porteurs d'une mutation V600E de BRAF), prises entre janvier 2000 et décembre 2018, ont été utilisées dans le cadre de cette étude rétrospective approuvée par le CÉR. Les segmentations des tumeurs (volumes d'intérêts, VOI) ont été fournies par un *fellow* en neuroradiologie pédiatrique, et vérifiées par un neuroradiologiste spécialisé en pédiatrie. Les patients ont été divisés de façon aléatoire en deux ensembles : « développement » et « test », selon un ratio de 80/20. Les masquages 3D binaire des VOI de chaque classe de l'ensemble « développement » ont été combinés afin d'obtenir des fonctions de probabilité d'emplacement des tumeurs liée à la densité. Trois pipelines de diagnostic moléculaire de gliome de bas grade chez l'enfant ont été mis au point : diagnostic fondé sur l'emplacement, diagnostic fondé sur les CNN, et diagnostic hybride. L'expérience a été répétée 100 fois, chacune au moyen de modèles d'initialisation et de divisions des données différentes. L'aire sous la courbe de fonction d'efficacité du récepteur (AUROC) a été calculée, et le test de Student a été réalisé. **Résultats :** Le trieur fondé sur l'emplacement a obtenu une AUROC de 77,9 et un intervalle de confiance (IC) de 95 % (76,8 - 79,0). Le trieur fondé sur les CNN a obtenu une AUROC de 86,1 et un IC de 95 % (85,0 - 87,3). Le trieur hybride a surpassé les autres trieurs; son AUROC moyenne était de 88,64, avec un IC de 95 % (87,6 - 89,7), ce qui est significatif sur le plan statistique ($P = 0,0018$). **Conclusion :** L'incorporation de cartes de probabilité d'emplacement des tumeurs dans des modèles de CNN a permis d'améliorer de manière significative la détermination des sous-types moléculaires des gliomes de bas grades chez l'enfant.

Keywords

pediatric low-grade glioma, brain tumour, CNN, machine learning, tumour location

Introduction

Brain tumours are the most common solid cancer among children, with pediatric Low-Grade Glioma (pLGG) being the most frequent.¹⁻³ The advent of targeted therapies such as BRAF proto-oncogene, serine/threonine kinase (BRAF) inhibitors^{4,5} has improved therapeutic outcomes of pLGG, but successful treatment planning for pLGG is governed by identifying tumour type and molecular subtype.^{6,7} Currently, the standard of care for molecular subtype identification of pLGG is tissue diagnosis through biopsy or surgery, which carries inherent risks, and sometimes is not feasible due to a tumour's location.⁸⁻¹¹

While MRI visualizes the tumour in its entirety and could represent a non-invasive alternative to biopsy for tumour classification, determining the molecular subtype of a tumour based on MRI remains a challenging task.¹² The feasibility of Machine Learning (ML) algorithms to identify genetic markers of pLGG has been demonstrated,¹³ but there remain important gaps and opportunities warranting further improvement. The performance of MRI-based pLGG subtype identification pipelines in the literature is currently suboptimal and tumour location has been shown to be a significant predictor.^{14,15}

We, therefore, aimed to establish a tumour-location—and a Convolutional Neural Network (CNN) based pipeline and a merged CNN pipeline with tumour location probability maps, and to evaluate their respective performance to identify molecular subtypes of pLGG based on MRI. The motivation for the proposed tumour-location-guided CNN algorithm was using regions outside the manual segmentation to improve the classification performance.

Materials and Methods

Dataset

The local institutional research ethics board approved this retrospective study waiving the need for informed consent. The internal dataset from The Hospital for Sick Children (Toronto, Ontario,

Canada) included MR images of patients with the 2 most common molecular subtypes of pLGG, BRAF fusion and BRAF p.V600E mutation. Patients were identified using the electronic health record (EHR) database of the hospital from January 2000 to December 2018. Inclusion criteria were an age of 0 to 18 years, histopathological/molecular confirmation of BRAF gene status, and a diagnosis of BRAF fusion or BRAF p.V600E mutation. Exclusion criteria were imaging artifacts precluding assessment, absence of an axial FLAIR sequence, and a molecular diagnosis other than BRAF fusion or BRAF p.V600E mutation.

All patients underwent MRI of the brain at field strengths of 1.5T or 3T, using MRI scanners from various vendors (Signa, GE Healthcare; Achieva, Philips Healthcare; Magnetom Skyra, Siemens Healthineers). We only used the axial FLAIR sequence (3-5 mm slice thickness; 0-2 mm gap) in order to maximize the sample size. Segmentation of volumes of interest (VOIs) was performed by a neuroradiology fellow using a semi-automated approach on FLAIR images with the Level-Tracing-Effect tool in the 3D Slicer library (Version 4.10.2, <https://www.slicer.org/>). In terms of reproducibility and robustness, the semi-automatic process has been confirmed to surpass multi-user manual delineation.¹⁶ The final VOIs were confirmed by a pediatric neuroradiology fellowship-trained and board-certified radiologist with 7 years of neuroradiology research experience.

The preprocessing pipeline included labelling, resampling, normalization, skull stripping, bias correction, and registration to the SRI24¹⁷ atlas for each image volume. SRI24 is an MRI atlas based on normal adult human brain anatomy, which is a well-known option for preprocessing brain MRI.¹⁸ It should be highlighted that registration is a key step in the preprocessing pipeline. Without proper registration, tumour location may become imprecise, and thus, not useful for the pLGG molecular diagnosis.

Location-Based Analysis

The proposed location-based pipeline only uses the manual segmentation mask of pLGG tumours, ignoring the rest of the

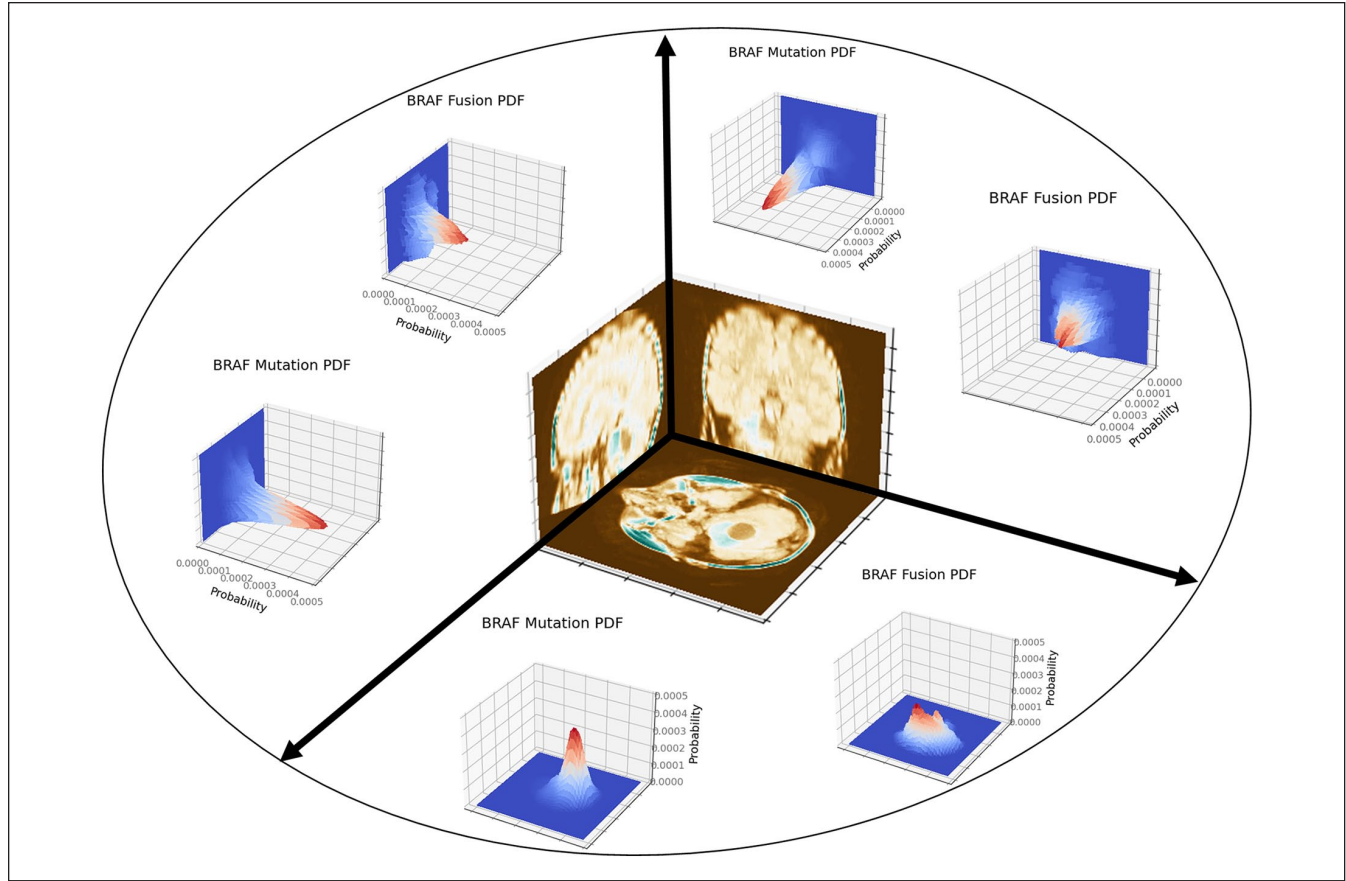


Figure 1. Projections of the tumour location PDFs in axial, coronal, and sagittal planes.

images. In contrast to previous work where tumour location was used as a binary variable,^{13,19} we used tumour location probability density functions (PDF) to achieve voxel-level granularity. Tumour location PDFs of BRAF p.V600E mutation and BRAF fusion were defined through summing and normalizing the 3D binary (manual) segmentation masks in FLAIR images for each class in the development dataset (ie, the union of training and validation datasets). In the test cohort, the probabilities of belonging to each class (p_c) were calculated by summation of a voxel-wise multiplication of the binary (manual) segmentation mask of the test case (*segmentation*) and the 3D PDF of the corresponding class (pdf_c) in the development dataset (equation (1)).

$$p_c = \sum (segmentation \odot pdf_c) \mid C \in \{fusion, mutation\} \quad (1)$$

To facilitate Receiver Operating Characteristic Curve (ROC) analysis and address the issue where the sum of probabilities for fusion and mutation classes does not equal one, equation (2) was utilized to calculate predicted probabilities for each patient.

$$p = \begin{cases} \min(0.5 + p_{mutation}, 1) & \text{if } p_{mutation} > p_{fusion} \\ \max(0.5 - p_{fusion}, 0) & \text{otherwise} \end{cases} \quad (2)$$

Figure 1 illustrates the projections of the PDFs in axial, coronal, and sagittal planes. We repeated the data split (80/20 for

validation and test) 100 times and calculated the Area Under the ROC (AUROC) for each run.

CNN-Based Analysis

We used off-the-shelf CNN models such as 3D ResNet,²⁰ as well as an in-house developed shallow CNN architecture described in Appendix A. The codes for defining the models, deriving PDFs and conducting location-based classification are publicly available (<https://github.com/IMICSLab/TumorLocationPDF>). The input to the CNNs were VOIs formed through element-wise multiplication of the manual segmentations and images. To train the models, we chose a batch size of 8, maximum number of epochs of 10, learning rate of 0.1, Cross Entropy (CE) as the loss function, and stochastic gradient descent (SGD)²¹ as the optimizer. The models were implemented using PyTorch 1.10.2, in a Python 3.9.7 environment with cuda 11.3. We utilized 2 GeForce RTX 3090 Ti GPUs on a Lambda Vector GPU workstation.

Tumour-Location-Guided CNN Analysis

The motivation behind the proposed CNN algorithm, guided by tumour location, was to enhance classification accuracy by utilizing areas beyond the manually segmented

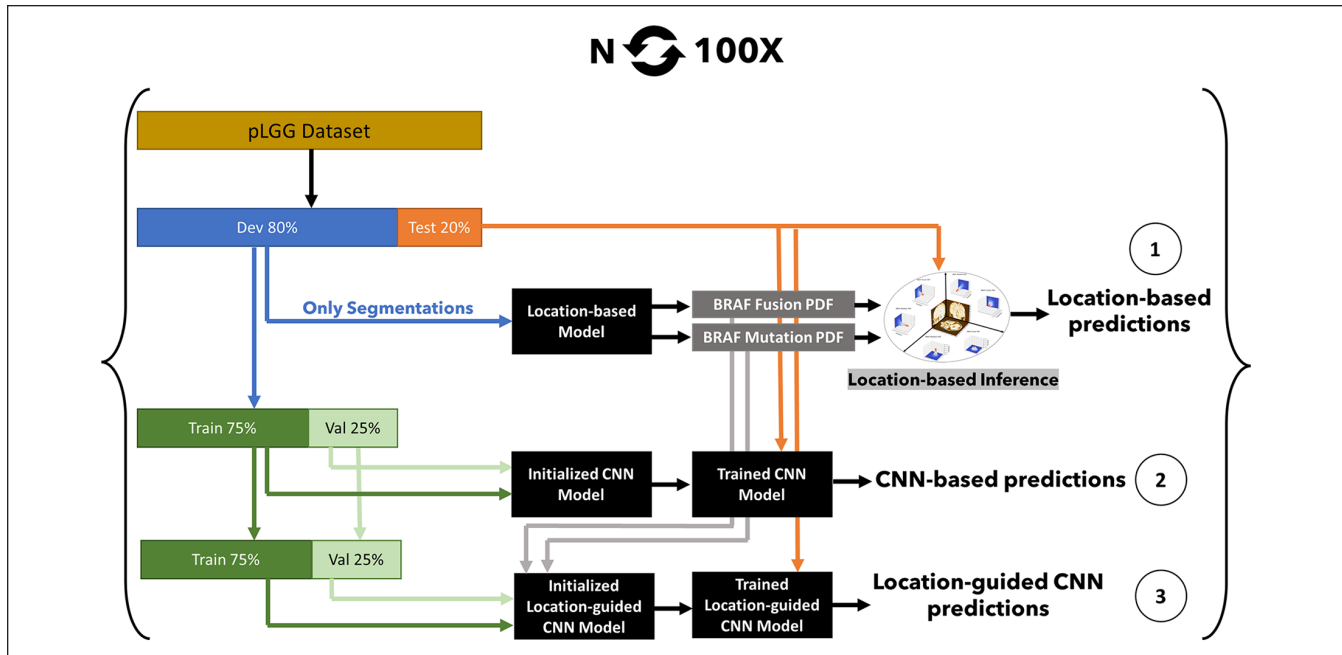


Figure 2. Location-based pLGG molecular biomarker identification pipelines: (1) location-only, (2) CNN-only, (3) location-augmented CNN.

regions. In contrast to the CNN-based pipeline where VOIs were created through element-wise multiplication of manual segmentations and images, 2 revisions were applied to design the tumour-location-guided CNN: (a) when multiplied by the image, an offset scalar was added to the binary segmentation mask to avoid eliminating the image areas where the mask elements are zero, (b) the 2 PDFs were weighted based on their probability for a given image (p_c , according to equation (1)) and summed up, and the result was used as a mask to dim image areas where tumour presence was unlikely.

In the tumour-location-guided CNN algorithm, the location PDFs are applied to each *image* according to equation (3), where *offset* is a scalar to help retain regions outside of the *segmentation*.

$$\text{input} = (\text{offset} + \text{segmentation}) \times \text{image} \times \sum_C p_C \times \text{pdf}_C$$

$$| C \in \{\text{fusion}, \text{mutation}\} \quad (3)$$

The setting for the tumour-location-guided CNN analysis was identical to how the CNN pipeline was developed except for the maximum number of epochs which was increased to 20 based on observations on the first 5 experiments. Figure 2 illustrates the 3 approaches we used to identify pLGG molecular subtype. A Monte Carlo random data splitting approach was used to evaluate the pipeline over 100 experiments, similar to the OpenRadiomics protocol.²² In each experiment, the dataset was randomly split into development/test sets with an 80/20 ratio. The development set was further split into train/validation sets using a 75/25 ratio, randomly.

Statistical Analysis

The Monte Carlo method for data splitting and model initialization (repetitive train/validation/test splits with different model initializations) enabled the acquisition of 100 test AUROC results across 3 models: location-only, CNN-only, and location-guided CNN. To assess the significance of performance enhancements among these models, we applied the Student's *t*-test. This rigorous statistical approach provided a framework for evaluating the efficacy of incorporating location information into CNN models, thereby allowing for a detailed comparison of their predictive capabilities.

Results

After initial screening, 397 patients were identified for the study. Absence of FLAIR and non-NRAF fusion or p.V600E mutation subtype resulted in exclusion of 168 patients. Additionally, 15 patients were excluded due to motion-degraded FLAIR images, as illustrated in Figure 3. The internal dataset, described in Table 1, included MR images from 214 patients with the 2 predominant molecular subtypes of pLGG: 143 with BRAF fusion (mean age 7.64 years, 70 male patients) and 71 with BRAF p.V600E mutation (mean age 10.36 years, 40 male patients).

In our initial experiments, the shallow CNN was faster (58.5 vs 12.2 minutes run time), while marginally outperforming 3D ResNet²⁰ (mean AUROC of 86.0 vs 85.3 on 5 runs), and thus we used the shallow architecture throughout the remainder of the study (Table 2). Additionally, we used 3D ResNet models pre-trained on Kinetics 400 dataset,²³ which did not surpass the shallow model (mean AUROC 85.4 on 5 runs).

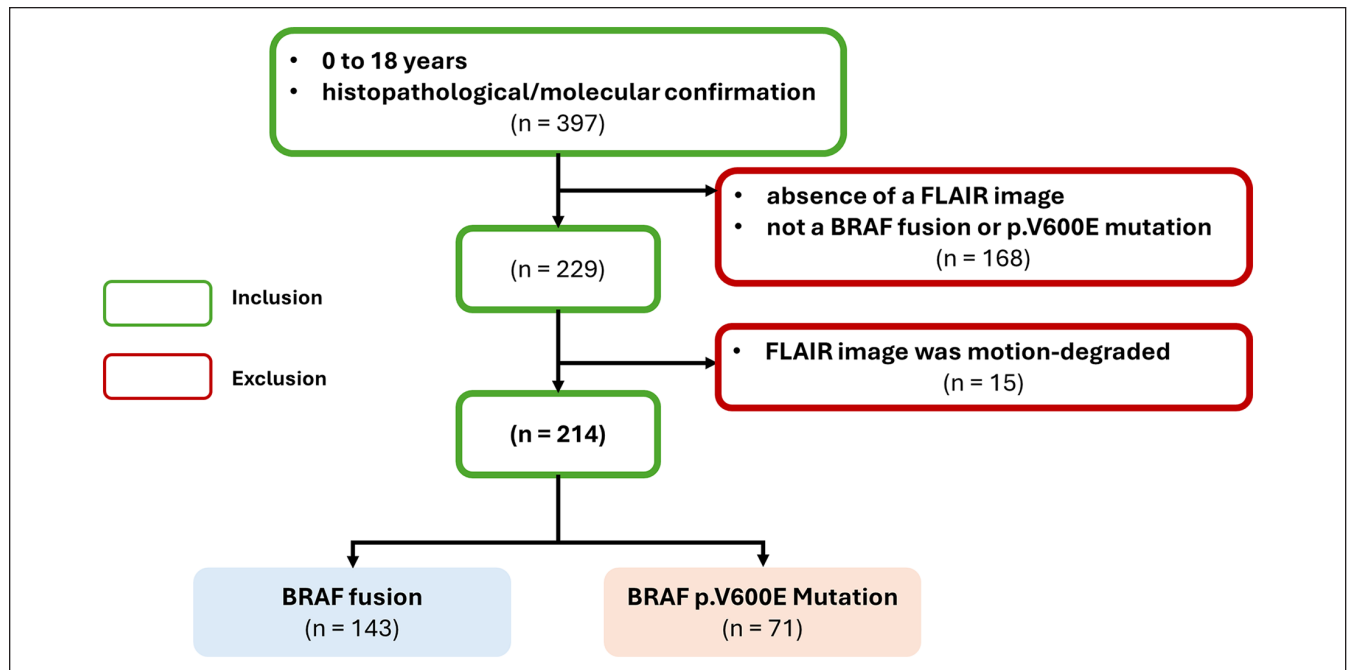


Figure 3. Inclusion/exclusion flowchart for the dataset.

We repeated the experiments 100 times and unified the data splits for the 3 pipelines (ie, the same patients were dedicated to training, validation, and test cohorts across the pipelines). The location-based classifiers achieved an AUROC of 77.9, 95% CI (76.7, 79.0). CNN-based classifiers resulted in AUROC of 86.1, 95% CI (85.0, 87.3), and the tumour-location-guided CNNs surpassed the other 2 models with an average AUROC of 88.6, 95% CI (87.6, 89.7). Figure 4 shows the results for the 3 pLGG subtype identification methods.

We varied the offset parameter as delineated in equation (3), monitoring its impact on the validation performance. Although the performance was not sensitive to the offset value, we determined an optimal value to be 0.2.

Figure 5 highlights 2 examples of how the 3D tumour-location PDFs allow the CNN to investigate regions beyond the manual segmentations. The top and bottom rows depict MR images in the axial plane for a patient with a BRAF fusion and a BRAF mutation, respectively. The BRAF fusion slice is at $z=36$ (infratentorial) and the BRAF mutation slice is at $z=56$ (supratentorial). The preprocessed MRI scans in the axial plane have 155 slices. In each row, the first column displays the MR image, the second column shows the manual segmentation of the tumour, and the third column illustrates the corresponding tumour location PDF. It is important to note that the third column does not represent the model's attention map, and rather, the probability of tumour location based on the training dataset. Thus, the PDF and the manual segmentation are not expected to align perfectly. The tumour location PDF modifies the input, enabling the CNN to consider brain regions that extend beyond but are adjacent to the boundaries of the segmentation mask.

In terms of utilizing tumour location in the CNN pipeline, we tried different approaches for incorporating tumour location information into the CNNs. The architectures that were tried are included in the code repository (<https://github.com/IMICSLab/TumorLocationPDF>). Injecting location-based probabilities into different layers of the CNN architecture, and ensemble of the CNN and location-based models were among the methods that were tested. However, we achieve marginal or no improvement in terms of average AUROC compared with the proposed pipeline. Using a binary variable (supra-/infratentorial tumour location) instead of the location-based probabilities did not improve the average AUROC.

Using Yuden's J point,²⁴ we evaluated the models in terms of sensitivity, specificity, and accuracy. As shown in Table 3, the proposed tumour-location-guided model achieved a sensitivity, specificity, and accuracy of 0.851 (95% CI [0.817, 0.885]), 0.847 (95% CI [0.812, 0.883]), 0.850 (95% CI [0.826, 0.873]), respectively.

Discussion

In this study, we developed multiple ML-based non-invasive pipelines to identify molecular subtypes of pLGG tumours using MR images. First, a tumour-location-only pipeline based on the 3D PDFs of BRAF fusion and BRAF p.V600E tumours was implemented and achieved a mean AUROC of 77.9. Traditionally, tumour location has been used as a binary variable to improve the molecular subtype classification of pLGG.^{13,19} To test the hypothesis that a tumour-location-only pipeline based on the 3D PDFs is superior to the traditional approach, we used a Random Forest (RF)-based pipeline with

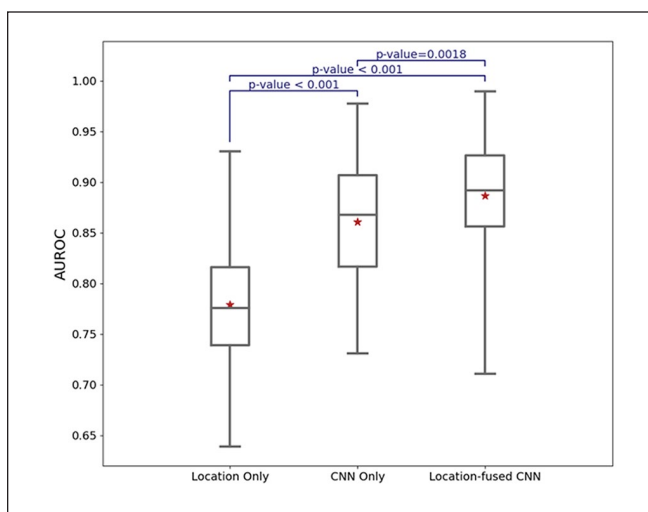
Table 1. Patient Demographics.

Demographics and Clinical Features	Whole dataset (n=214)	BRAF fusion (n=143)	BRAF p.V600E (n=71)
Mean age (SD)	8.54 (4.97)	7.64 (4.77)	10.36 (4.91)
Gender			
Male (%)	51.40 ($\frac{110}{214}$)	48.95 ($\frac{70}{143}$)	56.34 ($\frac{40}{71}$)
Female (%)	48.60 ($\frac{104}{214}$)	51.05 ($\frac{73}{143}$)	43.66 ($\frac{31}{71}$)
Tumour location			
Infratentorial	50.93% ($\frac{109}{214}$)	70.63% ($\frac{101}{143}$)	11.27% ($\frac{8}{71}$)
Supratentorial	49.07% ($\frac{105}{214}$)	29.37% ($\frac{42}{143}$)	88.73% ($\frac{63}{71}$)
Pathology			
Pilocytic astrocytoma	54.67% ($\frac{117}{214}$)	76.92% ($\frac{110}{143}$)	9.86% ($\frac{7}{71}$)
Low grade astrocytoma	14.95% ($\frac{32}{214}$)	9.09% ($\frac{13}{143}$)	26.76% ($\frac{19}{71}$)
Ganglioglioma	13.08% ($\frac{28}{214}$)	4.20% ($\frac{6}{143}$)	30.99% ($\frac{22}{71}$)
Diffuse astrocytoma	5.61% ($\frac{12}{214}$)	2.10% ($\frac{3}{143}$)	12.68% ($\frac{9}{71}$)
Pilomyxoid astrocytoma	4.21% ($\frac{9}{214}$)	5.59% ($\frac{8}{143}$)	1.41% ($\frac{1}{71}$)
Pleomorphic xanthoastrocytoma	2.80% ($\frac{6}{214}$)		8.45% ($\frac{6}{71}$)
Dysembryoplastic neuroepithelial tumour	0.93% ($\frac{2}{214}$)		2.82% ($\frac{2}{71}$)
Neurocytoma	0.93% ($\frac{2}{214}$)	1.40% ($\frac{2}{143}$)	
Oligodendroglioma	0.93% ($\frac{2}{214}$)		2.82% ($\frac{2}{71}$)
Mixed tumour components	0.93% ($\frac{2}{214}$)	0.70% ($\frac{1}{143}$)	1.41% ($\frac{1}{71}$)
Gangliocytoma	0.47% ($\frac{1}{214}$)		1.41% ($\frac{1}{71}$)
Glioneuronal tumour	0.47% ($\frac{1}{214}$)		1.41% ($\frac{1}{71}$)

Table 2. Comparing 3D ResNet and the In-House Shallow 3D CNN.

Architecture	Average AUROC (on 5 random experiments)	Average run time per experiment (min)
3D ResNet	85.3	58.5
Pretrained 3D ResNet	85.4	57.8
Shallow 3D CNN	86.0	12.2

the same settings (ie, random training/test splits with 80/20 ratio and 100 repeats) to compare tumour location as a binary variable (supra- vs infratentorial) with the 3D PDFs that resulted in a mean AUROC of 75.8, 95% CI (74.8, 76.7), which was significantly lower than the proposed 3D PDFs (P -value=.0089). The hyperparameters and grid search setting for the experiment was similar to that proposed in a previous work.²⁵ Second, a 3D CNN-based pipeline was trained and evaluated to classify FLAIR VOIs and significantly improved

**Figure 4.** AUROC performance of the classification algorithms on test cohorts.

Note. The plot illustrates ranges of AUROCs, not CIs.

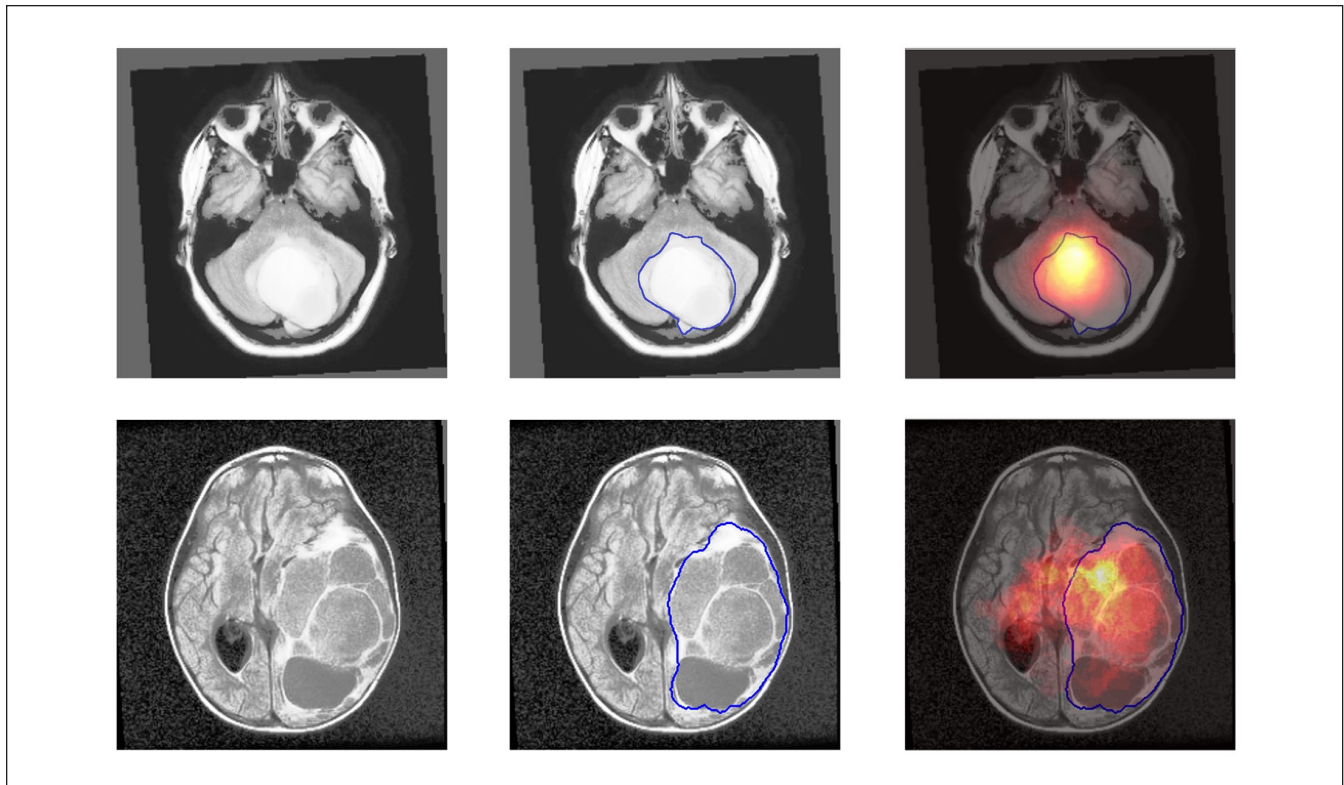


Figure 5. Axial plane slices showing the application of 3D tumour-location PDFs. Top row: BRAF fusion patient slice at $z=36$ (infratentorial). Bottom row: BRAF mutation patient slice at $z=56$ (supratentorial). Columns represent (left) MR image, (middle) manual segmentation, and (right) corresponding tumour location PDF. The PDFs, distinct from the model's attention map, modify model's input and allow considering regions beyond the manual segmentation boundaries.

Table 3. Sensitivity, Specificity, and Accuracy of the Proposed Tumour-Location-Guided Model.

Metric	Performance level	95% CI
Sensitivity	85.1	[81.7, 88.5]
Specificity	84.7	[81.2, 88.3]
Accuracy	85.0	[82.6, 87.3]
Baseline accuracy (proportion of the largest class to sample size)	66.8	—

the performance to 86.1 (P -value $< .001$). Lastly, the CNN-based pipeline was augmented by applying the 3D tumour location PDFs to the MR images and achieved a mean AUROC of 88.6, which showed a significant improvement (P -value $.0018$). We observed wider CI ranges for mean test sensitivity, specificity, and accuracy. This highlights the higher variance (ie, significantly low or high values across the experiments) of these metrics compared with AUROC. Thus, more advanced calibration of the models (ie, beyond using Yuden's J point) and ensemble methods should be investigated in future research to improve the parametric performance metrics of the models (eg, sensitivity and specificity). Our contributions include (a) using tumour location as an independent modality, as opposed to the

conventional approach of using tumour location as a categorical variable, (b) improving performance and explainability of CNN pipelines through the utilization of tumour location, and (c) utilizing Monte Carlo data splitting and a repetitive evaluation approach to measure randomness of CNN pipelines for pLGG molecular diagnosis.

In equation (3), offset introduces an additional hyperparameter to the pipeline, which can be optimized using grid search. After evaluating offset values of 0.1, 0.2, 0.3, 0.4, and 0.5, no significant difference in validation AUROC was observed, and an offset of 0.2 was selected for the experiments. The value of offset is not a decisive hyperparameter because it is added to the segmentation mask rather than multiplied. Any value of offset allows the CNN to explore the entire image according to equation (3), with the summation of the PDFs acting as the primary mechanism to restrict unrelated regions. However, larger offset values increase the voxel intensity in the model's input. Given that CNN inputs are typically normalized with a bounded norm,²⁶ smaller offset values are generally preferable to maintain consistency with standard normalization practices.

Segmentation-free pLGG subtype identification using MR images is a challenging task that has not yet been addressed in the literature. We initially conducted experiments without segmentation masks, but the CNNs failed to converge. It should be highlighted that the offset is not an indicator of reliance on

manual segmentation, and no value of offset can result in a truly segmentation-free approach since the offset is summed with the segmentation mask. Even assuming that a large offset does not significantly impact the norm of the CNN's input and can be processed by the network, zeros in the PDFs will continue to exclude irrelevant areas of the image. Given that the PDFs are derived from the segmentation masks, the process would not be considered segmentation-free. Additionally, CNNs are sensitive to patterns rather than absolute values. Thus, adding a large offset to the segmentation still preserves the underlying tumour boundary patterns. Similarly, small offset values would still reflect the segmentation patterns. A special case occurs when the offset is set to zero. Although this scenario emphasizes the segmentation mask, the results are different from the CNN-based analysis due to the influence of the PDFs. We conducted an experiment with an offset of zero and achieved a mean AUROC of 86.0, 95% CI (85.0, 87.2), which was not significantly different from the CNN-based analysis.

Radiomics and CNNs form the 2 established branches of ML, applicable to pLGG molecular subtype identification.²⁷ Unlike radiomics, CNNs learn to extract the features and are not limited to predefined formulas for pattern recognition. Thus, CNNs have the potential to outperform radiomics-based models²⁸ which motivates why we focused on CNNs in this study. In conventional CNNs, feature extraction is done by sequential convolution layers, and fully connected (FC) layers classify the features. Deep learning (DL) uses CNNs with a high number of convolutional layers, and deep models are state-of-the-art on multiple large-scale datasets.²⁹ However, on small datasets where no pre-trained model is available, deeper models may not improve performance.³⁰ In our experiments, the shallow CNN outperformed 3D ResNet and enabled us to lower the computational load and increase the data split repetitions.

The importance of tumour location for MRI-based pLGG molecular subtype identification has been highlighted in several studies. Bag et al identified 5 genetic profiles, location, age at presentation, and histology as the decisive features for pLGG phenotype risk assessment and emphasized the differences in tumour location between BRAF fusion and BRAF V600E mutation.³¹ In their study, 75% of tumours with BRAF fusion were located in the cerebellum, while 56% of tumours with BRAF V600E mutation were found in the cerebral hemispheres. Wagner et al used radiomics to differentiate BRAF-mutated and BRAF-fused tumours based on pre-therapeutic FLAIR images.¹³ In their bi-institutional retrospective study of 115 pediatric patients, they achieved an average AUROC of 0.75 on the internal cohort and 0.85 on the external cohort. Location (supratentorial vs infratentorial) and age were significant clinical predictors of BRAF status and the average AUROC increased to 0.77 on their internal cohort when age and location were added to the radiomics features. Halder et al retrospectively studied a dataset of 157 patients with pLGG from the Children's Hospital of Philadelphia using a conventional unsupervised ML algorithm.³² They employed Principal Component Analysis (PCA) followed by a K-means algorithm to assign patient images into 3 subgroups; a subsequent Kruskal-Wallis test demonstrated the distribution of tumour

histology and location to be different between the 3 imaging clusters. Xu et al analyzed a dataset of 113 patients with pLGG (43 with BRAF V600E mutations vs 70 with other subtypes) using radiomics features to identify BRAF mutations.³³ Tumour location (supratentorial vs infratentorial) was a significant predictor of BRAF mutation and when combined with radiomics, improved the average training AUROC to 0.754 and the test AUROC to 0.934. Laterality of tumour location (left vs right) was not a significant predictor of BRAF mutation.

Dataset size is another crucial factor for reliability of any ML model. In a study on a larger bi-institutional dataset of 251 patients with pLGGs, Wagner et al conducted a dataset size sensitivity analysis.²⁵ They showed that data splits and model initialization impose randomness on the performance of the ML classifiers which impacts the results in 2 aspects: average AUROC and variance of the AUROCs. The study confirmed that the dataset size was sufficient to train robust pipelines, as evidenced by acceptable mean performance with bounded variance. However, it was observed that an individual global model could exhibit bias, indicated by a performance drop on the external dataset and an increase in variance. With only 60% ($\frac{132}{220}$) of the training data, they achieved comparable results to those of models that used the entire training set (average AUROC of 0.83 compared to 0.85). We therefore consider the size of the dataset of our study to be reliable and the Monte Carlo method to be appropriate for data splitting.

There are multiple sources of variability impacting generalizability and reproducibility of ML pipelines.³⁴ Classifiers are sensitive to any form of change in input, including differences in imaging protocols, sequences parameters, models and manufacturers of MRI scanners, tumour segmentations, and image normalization. We employ the repetitive approach proposed in OpenRadiomics²² for reproducible ML research on relatively small datasets. With a limited sample size, outliers might impact the fairness of data splits.³⁵ Hence, creating a single reproducible model becomes infeasible and the focus should be on training repeatable pipelines.

A limitation of our study is that the predictive significance of tumour location PDFs may diminish as additional pLGG subtypes are incorporated into the dataset. However, recent work by Tak et al has demonstrated that a sequential classification approach among different classes can be effective.³⁶ In such a scenario, location PDFs can enhance the differentiation between BRAF Fusion and p.V600E mutation subtypes. While the utility of location PDFs may vary with the complexity of the dataset, they can still provide valuable insights in specific classification tasks.

While the experiments showed that a rough supra-/infratentorial tumour segmentation was not helpful, using rough tumour segmentation needs to be tested in future work to eliminate the need for tedious manual segmentations. Bounding boxes of the tumours can be used as alternatives to the manual segmentation.

In our experiments, the shallow architecture marginally outperformed the randomly initialized and pre-trained 3D ResNet. We believe the higher number of learnable parameters results in overfitting with the 3D ResNet. However, transfer learning is

shown to be promising in medical image classification³⁵ and needs to be investigated in more detail. Pre-training on a brain MRI dataset instead of Kinetics 400, which includes human videos, would have a higher potential for improving the results and should be explored in future research. Nevertheless, shallower architectures remain computationally favourable.

Heterogeneity of the dataset remains a major concern that might induce bias to the model. We employed a comprehensive preprocessing pipeline including resampling, bias correction, and registration. However, the model and the dataset should be evaluated to ensure resolution, scanner vendor, and settings are not confounding. This requires additional pieces of information from the MR images which should be curated for future studies.

Conclusion

Incorporating tumour location probability maps into CNN models led to statistically significant improvements for molecular subtype identification of pLGG. These results suggest that conventional CNNs using manual segmentations may not be optimal as location information is lost.

Appendix A

The CNN architecture is visualized in Figure A1, which is generated using the PyTorchViz Python library (<https://github.com/szagoruyko/pytorchviz>). The model has three 3D CNN blocks, followed by 2 FC layers to map the 486 extracted features into binary labels. Each 3D CNN block includes a 3D CNN layer with kernel size of 3, ReLU activation function, and MaxPooling with kernel size of 2.

Abbreviations

AUROC	area under receiver operating characteristic curve
BRAF	B-Raf proto-oncogene, serine/threonine kinase
CE	cross entropy
CNN	convolutional neural network
DL	deep learning
EHR	electronic health record
FC	fully connected
ML	machine learning

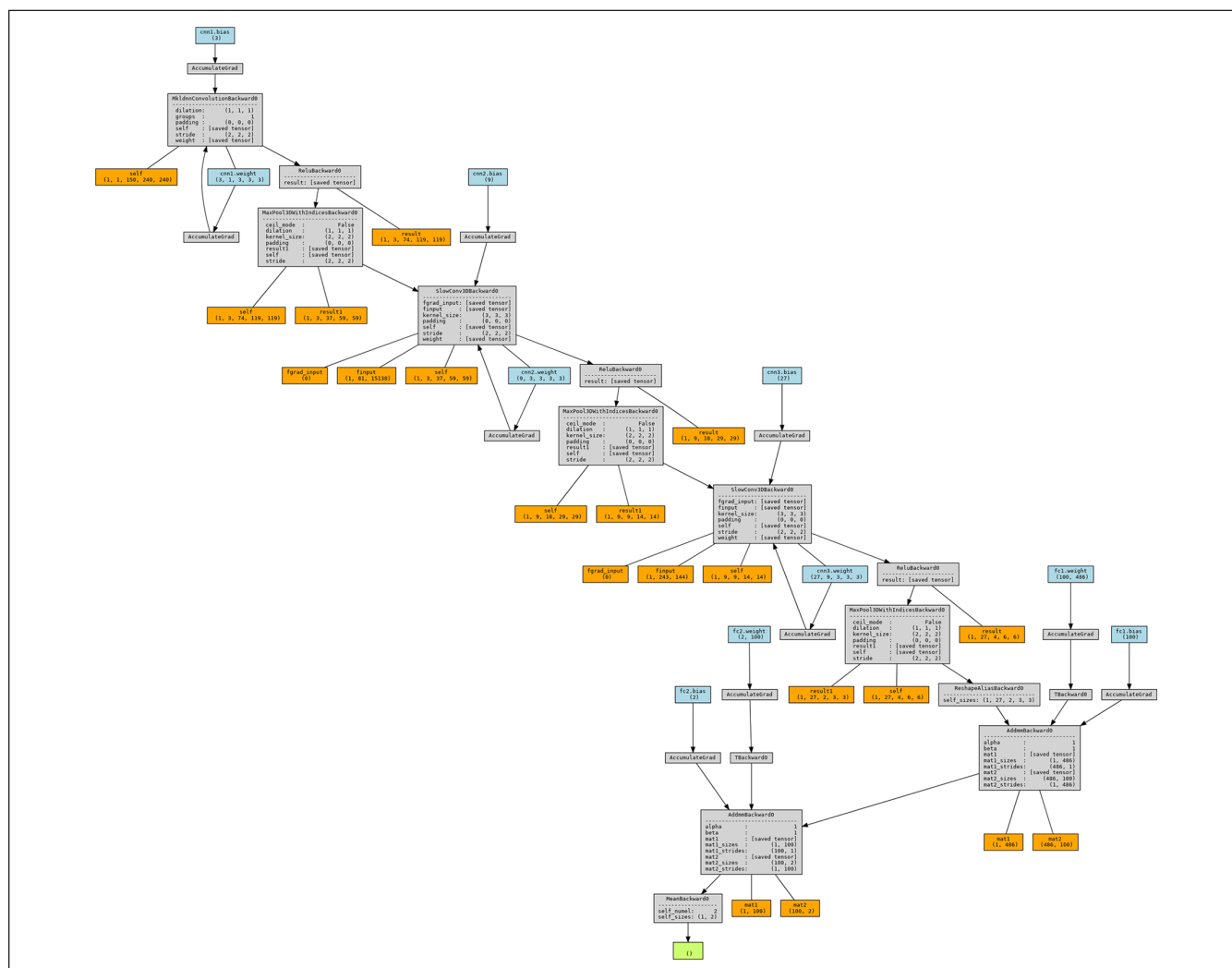


Figure A1. Visualization of the CNN architecture and parameters.

PCA	principal component analysis
PDF	probability density functions
pLGG	pediatric low-grade glioma
RF	random forest
ROC	receiver operating characteristic
SGD	stochastic gradient descent

Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has been made possible with the financial support of the Canadian Institutes of Health Research (CIHR; Funding Reference Number: 184015).

ORCID iDs

Matthias W. Wagner  <https://orcid.org/0000-0001-6501-839X>

Farzad Khalvati  <https://orcid.org/0000-0001-5616-8660>

References

- Goebel AM, Gnekow AK, Kandels D, Witt O, Schmidt R, Hernáiz Driever P. Natural history of pediatric low-grade glioma disease - first multi-state model analysis. *J Cancer*. 2019;10(25):6314-6326. doi:10.7150/jca.33463
- Lassaletta A, Scheinemann K, Zelcer SM, et al. Phase II weekly vinblastine for chemotherapy-naïve children with progressive low-grade glioma: a Canadian pediatric brain tumor consortium study. *J Clin Oncol*. 2016;34(29):3537-3543. doi:10.1200/JCO.2016.68.1585
- Armstrong GT, Conklin HM, Huang S, et al. Survival and long-term health and cognitive outcomes after low-grade glioma. *Neuro Oncol*. 2011;13(2):223-234. doi:10.1093/neuonc/noq178
- Naderi-Azad S, Sullivan R. The potential of BRAF-targeted therapy combined with immunotherapy in melanoma. *Expert Rev Anticancer Ther*. 2020;20(2):131-136. doi:10.1080/14737140.2020.1724097
- Manoharan N, Liu KX, Mueller S, Haas-Kogan DA, Bandopadhyay P. Pediatric low-grade glioma: targeted therapeutics and clinical trials in the molecular era. *Neoplasia*. 2023;36:100857. doi:10.1016/j.neo.2022.100857
- Sturm D, Pfister SM, Jones DTW. Pediatric gliomas: current concepts on diagnosis, biology, and clinical management. *J Clin Oncol*. 2017;35(21):2370-2377. doi:10.1200/JCO.2017.73.0242
- Pollack IF, Agnihotri S, Broniscer A. Childhood brain tumors: current management, biological insights, and future directions. *J Neurosurg Pediatr*. 2019;23(3):261-273. doi:10.3171/2018.10.PEDS18377
- AlRayahi J, Zapotocky M, Ramaswamy V, et al. Pediatric brain tumor genetics: what radiologists need to know. *Radiographics*. 2018;38(7):2102-2122. doi:10.1148/rg.2018180109
- Krishnatre R, Zhukova N, Guerreiro Stucklin AS, et al. Clinical and treatment factors determining long-term outcomes for adult survivors of childhood low-grade glioma: a population-based study. *Cancer*. 2016;122(8):1261-1269. doi:10.1002/cncr.29907
- Woodworth GF, McGirt MJ, Samdani A, Garonzik I, Olivi A, Weingart JD. Frameless image-guided stereotactic brain biopsy procedure: diagnostic yield, surgical morbidity, and comparison with the frame-based technique. *J Neurosurg*. 2006;104(2):233-237. doi:10.3171/jns.2006.104.2.233
- Gao H, Jiang X. Progress on the diagnosis and evaluation of brain tumors. *Cancer Imaging*. 2013;13(4):466-481. doi:10.1102/1470-7330.2013.0039
- Bell D, Grant R, Collie D, Walker M, Whittle IR. How well do radiologists diagnose intracerebral tumour histology on CT? Findings from a prospective multicentre study. *Br J Neurosurg*. 2002;16(6):573-577. doi:10.1080/02688690209168363
- Wagner MW, Hainc N, Khalvati F, et al. Radiomics of pediatric low grade gliomas: toward a pretherapeutic differentiation of BRAF-mutated and BRAF-fused tumors. *AJNR Am J Neuroradiol*. 2021;42(4):759-765.
- Mistry M, Zhukova N, Merico D, et al. BRAF mutation and CDKN2A deletion define a clinically distinct subgroup of childhood secondary high-grade glioma. *J Clin Oncol*. 2015;33(9):1015-1022. doi:10.1200/JCO.2014.58.3922
- Lassaletta A, Zapotocky M, Mistry M, et al. Therapeutic and prognostic implications of BRAF V600E in pediatric low-grade gliomas. *J Clin Oncol*. 2017;35(25):2934-2941. doi:10.1200/JCO.2016.71.8726
- Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One*. 2014;9(7):e102107. <https://doi.org/10.1371/journal.pone.0102107>
- Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. 2010;31(5):798-819. doi:10.1002/hbm.20906
- Bakas S. 2018 International MICCAI BraTS challenge. In: Proceedings of the 7th MICCAI BraTS Challenge (2018). Published online 2018.
- Kudus K, Wagner MW, Namdar K, et al. Increased confidence of radiomics facilitating pretherapeutic differentiation of BRAF-altered pediatric low-grade glioma. *Eur Radiol*. 2024;34(4):2772-2781. doi:10.1007/s00330-023-10267-1
- Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In: BT - 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018; June 18-22, 2018; Salt Lake City, UT. IEEE; 2018:6450-6459.
- Ruder S. An overview of gradient descent optimization algorithms. Published online 2016.
- Namdar K, Wagner MW, Ertl-Wagner BB, Khalvati F. Open-radiomics: a research protocol to make radiomics-based machine learning pipelines reproducible. Published online 2022.
- Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18-23, 2018; Salt Lake City, UT. IEEE; 2018:6546-6555. doi:10.1109/CVPR.2018.00685
- Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J*. 2008;50(3):419-430. doi:10.1002/bimj.200710415
- Wagner M, Namdar K, Alqabbani A, et al. Dataset size sensitivity analysis of machine learning classifiers to differentiate molecular markers of pediatric low-grade gliomas based on MRI. Published online September 17, 2021. doi:10.21203/rs.3.rs-883606/v1
- Sola J, Sevilla J. Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Trans Nucl Sci*. 1997;44(3):1464-1468. doi:10.1109/23.589532

27. Wagner MW, Namdar K, Biswas A, Monah S, Khalvati F, Ertl-Wagner BB. Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. *Neuroradiology*. 2021;63(12):1957-1967. doi:10.1007/s00234-021-02813-9
28. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500-510. doi:10.1038/s41568-018-0016-5
29. 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;18(2):203-211. doi:10.1038/s41592-020-01008-z
30. Namdar K, Gujrathi I, Haider MA, Khalvati F. Evolution-based fine-tuning of CNNs for prostate cancer detection. In: International Conference on Neural Information Systems (NeurIPS); 2019.
31. Bag AK, Chiang J, Patay Z. Radiohistogenomics of pediatric low-grade neuroepithelial tumors. *Neuroradiology*. 2021;63(8):1185-1213. doi:10.1007/s00234-021-02691-1
32. Halder D, Kazerooni AF, Arif S, et al. Unsupervised machine learning using K-means identifies radiomic subgroups of pediatric low-grade gliomas that correlate with key molecular markers. *Neoplasia*. 2023;36:100869. doi:10.1016/j.neo.2022.100869
33. Xu J, Lai M, Li S, et al. Radiomics features based on MRI predict BRAF V600E mutation in pediatric low-grade gliomas: a non-invasive method for molecular diagnosis. *Clin Neurol Neurosurg*. 2022;222:107478. doi:10.1016/j.clineuro.2022.107478
34. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol*. 2019;20(7):1124-1137. doi:10.3348/kjr.2018.0070
35. Hao R, Namdar K, Liu L, Khalvati F. A transfer learning-based active learning framework for brain tumor classification. *Front Artif Intell*. 2021;4:61. doi:10.3389/frai.2021.635766
36. Tak D, Ye Z, Zapaischykova A, et al. Noninvasive molecular subtyping of pediatric low-grade glioma with self-supervised transfer learning. *Radiol Artif Intell*. 2024;6(3):e230333. doi:10.1148/ryai.230333