Title: Noninvasive molecular subtyping of pediatric low-grade glioma with self-supervised transfer learning

Authors: Divyanshu Tak[1,2+], Zezhong Ye[1,2+], Anna Zapaischykova[1,2], Aidan Boyd[1,2], Sridhar Vajapeyam[3], Rishi Chopra[1,2], Yining Zha[1,2], Hasaan Hayat[1,2], Sanjay Prabhu[3], Kevin X. Liu[2], Hesham Elhalawani[2], Ali Nabavidazeh[4,5], Ariana Familiar[4,6], Adam Resnick[6], Sabine Mueller[7,8,9], Hugo J.W.L. Aerts[1,2,10,11], Pratiti Bandopadhayay[12], Keith Ligon[13], Daphne Haas-Kogan[2], Tina Poussaint[3], and Benjamin H. Kann[1,2*]

Affiliations:

1. Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA
2. Department of Radiation Oncology, Dana-Farber Cancer Institute | Brigham and Women's Hospital | Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
3. Department of Radiology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
4. Center for Data-Driven Discovery in Biomedicine (D3b), Children's Hospital of Philadelphia, Philadelphia, PA, USA
5. Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
6. Department of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, USA
7. Department of Neurology, University of California San Francisco, San Francisco, CA. USA
8. Department of Pediatrics, University of California San Francisco, San Francisco, CA, USA
9. Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA
10. Department of Radiology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA
11. Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands
12. Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
13. Department of Pathology, Dana-Farber Cancer Institute, Boston Children's Hospital, Harvard Medical School, Boston, A, USA

* Corresponding author

+ These authors contributed equally to this manuscript.

Correspondence address to:

Benjamin H. Kann, M.D.

Department of Radiation Oncology,

Dana-Farber Cancer Institute and Brigham and Women's Hospital,

Harvard Medical School, 75 Francis Street, Boston, MA 02115, MA, USA

Tel: +1 617-732-6310

Email: Benjamin_Kann@dfci.harvard.edu

Summary Statement

The authors developed and externally validated an automated, scan-to-prediction deep learning pipeline that classifies BRAF Mutational status in pediatric low-grade gliomas directly from T2-Weighted MRI scans.

Key Results

- An innovative training approach combining self-supervision and transfer learning ("TransferX") is developed to boost model performance in low data settings;
- TransferX enables the development of a scan-to-prediction pipeline for pediatric LGG mutational status (BRAF V600E, fusion, or wildtype) with high accuracy and mild performance degradation on external validation;
- An evaluation metric "COMDist" is proposed to increase interpretability and quantify the accuracy of the model's attention around the tumor.

Keywords

Deep Learning, BRAF Mutational Status, Segmentation, Classification, Pediatric Low-Grade Gliomas

List of Abbreviations

pLGG = pediatric low grade glioma; T2W = T2 Weighted; CNN = Convolutional neural network; SD = Standard Deviation; CI = Confidence Interval; AUC = Area under the curve; CBTN = Child brain tumor network.

ABSTRACT

Purpose

To develop and externally validate a scan-to-prediction deep-learning pipeline for noninvasive, MRI-based BRAF mutational status classification for pLGG.

Materials and Methods

We conducted a retrospective study of two pLGG datasets with linked genomic and diagnostic T2-weighted MRI of patients: Boston Children's Hospital (development dataset, N=214), and Child Brain Tumor Network (CBTN) (external validation, N=112). We developed a deep learning pipeline to classify BRAF mutational status (V600E vs. fusion vs. wild-type) from whole-scan input via a two-stage process: 1) 3D tumor segmentation and extraction of axial tumor images, and 2) slice-wise, deep learning-based classification of mutational status. We investigated knowledge-transfer approaches to prevent model overfitting with a primary endpoint of the area under the receiver operating characteristic curve (AUC). To enhance model interpretability, we developed a novel metric, COMDist that quantifies the accuracy of model attention with respect to the tumor.

Results

A combination of transfer learning from a pretrained medical imaging-specific network and self-supervised label cross-training (TransferX) coupled with consensus logic yielded the highest AUC, taken as a weighted average across the three mutational classes, (0.82 [95% CI: 0.70-0.90], Accuracy: 77%) on internal validation and (0.73 [95% CI 0.68-0.88], Accuracy: 75%) on external validation. Training with TransferX also led to an AUC improvement of 17.7% and a COMDist Improvement of 6.42% over training from scratch on the development dataset.

Conclusion

Transfer learning and self-supervised cross-training improved classification performance and generalizability for noninvasive pLGG mutational status prediction in a limited data scenario.

INTRODUCTION

Pediatric low-grade gliomas (pLGGs) are the most common pediatric brain tumors, comprising up to 40% of tumors in this population[1]. These tumors exhibit diverse clinical outcomes and molecular characteristics, often driven by an activating BRAF mutation, either the BRAF V600E point mutation or fusion events. Molecular classification is vital for accurate treatment selection and risk stratification in pLGGs, particularly given the emergence of novel BRAF-directed therapies[2]. The presence of the BRAF V600E mutation, found in 15-20% of cases, was historically associated with poor survival, particularly when combined with CDKN2A deletion[3], though with targeted BRAF pathway-directed therapies this may be changing. BRAF V600E-mutated pLGGs also exhibit an increased risk of malignant transformation[4] while patients with BRAF fusion and neurofibromatosis type 1 have a favorable outcome. Accurate distinction between BRAF V600E, BRAF fusion, and wildtype tumors, plays a crucial role in determining prognosis and optimal treatment strategy.

Surgical resection for pLGG allows for assessment of mutational status. However, in over one-third of cases, resection, or even biopsy, may not be feasible nor recommended[5]. In these situations, children may require alternative therapies to control a symptomatic tumor or undergo periodic magnetic resonance imaging (MRI) surveillance. In these situations, non-invasive, imaging-based, tumor molecular subtyping, if accurate and reliable, could enable proper selection of patients for BRAF-targeted therapies and clinical trials.

In recent years, deep learning, which extracts features from large quantities of raw data passed through multiple layers in a network[6], has become the state-of-the-art for medical imaging analysis[7] [8], including imaging-based molecular classification [9,10], and may have utility for pediatric brain tumor classification. However, DL performance degrades dramatically in limited data scenarios, due to instability, overfitting, and shortcut learning,[11] and a key barrier to applying deep learning to pediatric brain tumor imaging, is the lack of training data available for these rare tumor cases. For these reasons, there has been limited success in using deep learning for pediatric glioma mutational classification. Another barrier to clinical usability is that most algorithms have required manual tumor segmentation as input, which is resource-intensive and requires specialized expertise. To our knowledge, only one study has been published in an extended abstract preprint attempting to differentiate between BRAF V600E and fusion mutations in LGG with deep learning[12]. This study trained a convolutional neural network (CNN)

on a small number of manually segmented tumors, and was confined to a single institution, limiting its generalizability.

In this study, we address these gaps by developing and externally validating the first automated, scan-to-prediction deep learning pipeline capable of non-invasive BRAF mutational status prediction for pLGG. To achieve this, we propose several innovations, including a multistage pipeline with built-in pLGG segmentation, BRAF mutation classifiers, and a consensus decision block to predict BRAF mutation status, including BRAF V600E and BRAF fusion. We leverage the pLGG dataset as our developmental dataset and a novel combination of in-domain transfer learning and self-supervision approach, called "TransferX" to maximize performance and generalizability in a limited data scenario. Additionally, to improve interpretability of our pipeline, we introduce a way to quantify the model attention via spatial maps, called Center of Mass Distance **(**COMDist**)** analysis**.** COMDist estimates the distance (in mm) between the center of mass of the GradCAM heatmap and the tumor's center of mass. Together, these methods enable practical, accurate noninvasive mutational classification for pLGG.

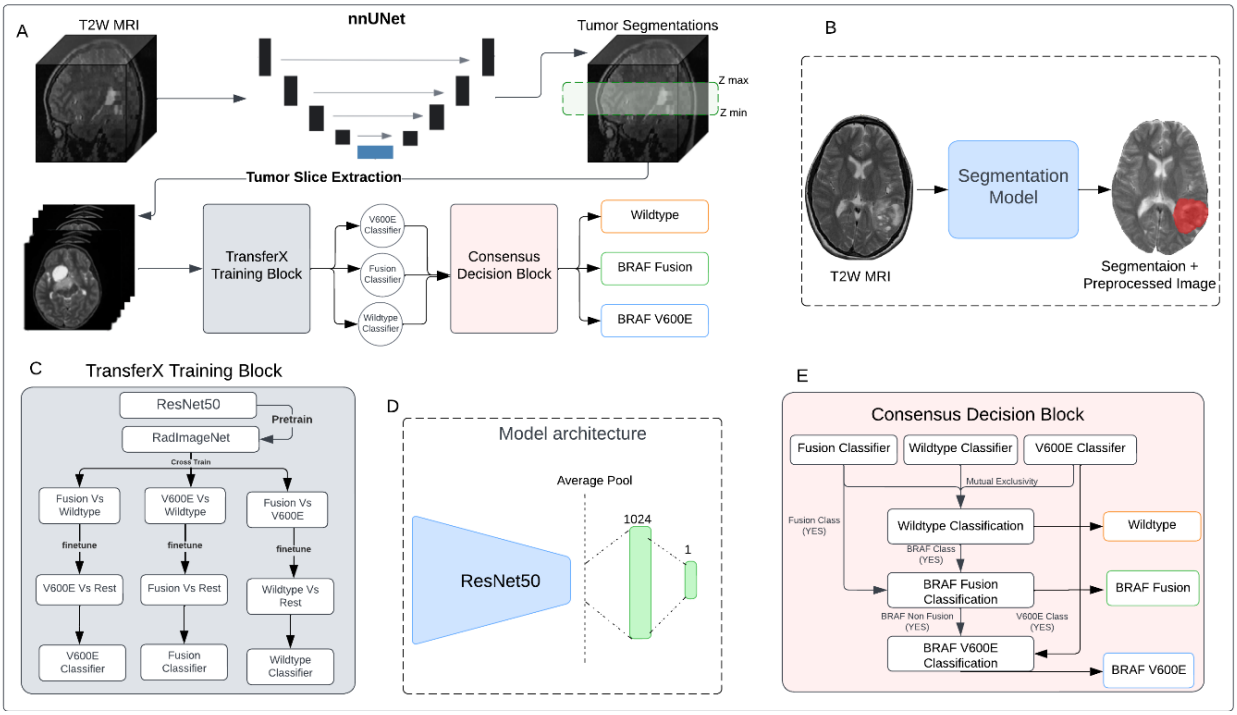


Figure 1*. (A) Schematic of the scan-to-prediction pipeline for molecular subtype classification. The pipeline inputs the raw T2W MRI scan and outputs the mutation class prediction. (B) Input and output depiction of the segmentation model from stage 1 of the pipeline. The segmentation*

block also involves registration and preprocessing of the input scan. The output consists of the preprocessed input MRI scan along with the co-registered segmentation mask. (C) Flow diagram of the TransferX training block and approach. The TransferX algorithm is employed to train three individual subtype classifiers (BRAF V600E, Wild-type, BRAF Fusion). (D) The model architecture of individual binary molecular subtype classifier. (E) Schematic of consensus decision block. The block inputs the classification outputs and corresponding scores from the three individual subtype classifiers and fits them into a consensus logic, and outputs the final predictions. The mutational class predictions are output sequentially where the input is first checked for wild-type or non-BRAF class first. If the input doesn't belong to wildtype or non-BRAF class, then the logic progresses to check the BRAF mutation class with BRAF Fusion checked first then followed by BRAF V600E. T2W: T2-weighted.

METHODS

Study Design and Datasets

This study was conducted in accordance with the Declaration of Helsinki guidelines and following the approval of the Dana-Farber/Boston Children's/Harvard Cancer Center Institutional Review Board (IRB). Waiver of consent was obtained from IRB prior to research initiation due to public datasets or retrospective study. This study involved two patient datasets: a developmental dataset from Boston Children's Hospital (BCH; n=214), for training, internal validation, and hypothesis testing, and a dataset from the Children's Brain Tumor Network (CBTN; n=112)[13] for external validation. Both datasets contained linked, pretreatment diagnostic T2W MRI and genomic information for children aged 1-25 years with a diagnosis of WHO grade I-II glioma. Each patient scan in the CBTN dataset was coupled with a ground truth segmentation and corresponding BRAF Mutation class label. These subsets represent all scans that passed initial quality control of DICOM metadata. Patient inclusion criteria were the following: 1) 1-25 years of age, 2) histopathologically confirmed pLGG, and 3) availability of preoperative brain MR imaging with a T2W imaging sequence. BRAF status was determined by OncoPanel, which performs targeted exome-sequencing of 227 to 477 cancer-causing genes (depending on panel versions 1-3). BRAF mutational status may also have been captured by genomic sequencing via in-house PCR on tissue specimens. In cases where neither could not be performed, immunohistochemistry (IHC) was used to determine V600E status. BRAF fusion status was determined by a gene fusion sequencing panel. DNA copy-number profiling via whole-genome microarray analysis was also performed in some cases. We report our results in

accordance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines[14] (Supplemental Methods: SM 1).

Table 1. *Patient cohort characteristics.*

	Development (BCH, n = 214)	**External Validation (CBTN, n = 112)**	***p*-values**
Age (years)			0.19*
median (range)	5 (1 – 20)	6 (1 - 21)	
Sex n (%)			0.82+
Female	95 (44.4%)	51 (45.5%)	
Male	113 (52.8%)	55 (49.1%)	
Unknown	6 (2.8%)	4 (3.6%)	
Race/Ethnicity n (%)			1.076e-06+
Non-Hispanic Caucasian/white	145 (67.8%)	71 (64.5%)	
African American/Black	6 (2.8%)	14 (12.7%)	
Hispanic/Latinx	3 (1.4%)	10 (9.1%)	
Asian American/Asian	9 (4.2%)	3 (2.7%)	
American Indian/Alaska Native	0	1 (0.9%)	
More than once race	0	1 (0.9%)	
Other/Unknown	51 (23.8%)	10 (9.1%)	
Histologic diagnosis n (%)			0.0005+
Pilocytic Astrocytoma	52 (24.2%)	68 (61.8%)	
Fibrillary Astrocytoma	0	8 (7.3%)	
Pilomyxoid Astrocytoma	8 (3.7%)	17 (15.5%)	
Ganglioglioma	13 (6.1%)	0	
Dysembryoplastic neuroepithelial tumor	7 (3.3%)	0	
Diffuse Astrocytoma	1 (0.5%)	7 (6.4%)	
Angiocentric Glioma	1 (0.5%)	1 (0.9%)	
Other Low-Grade Glioma/Astrocytoma	132 (61.7%)	9 (8.2%)	
BRAF Mutation Status n (%)			0.0005+
V600E	50 (23.4%)	17 (15.2%)	
Fusion	60 (28.0%)	60 (53.6%)	
Wildtype	104 (48.6%)	35 (31.3%)	
Tumor Locations n (%)			0.0005+
Cerebellum/Posterior fossa	40 (18.7%)	33 (29.4%)	
Temporal lobe	43 (20.1%)	12 (10.7%)	
Frontal Lobe	22 (10.3%)	4 (3.6%)	
Suprasellar	6 (2.8%)	32 (28.6%)	
Optic Pathway	8 (3.7%)	17 (14.9%)	
Brainstem	7 (3.3%)	9 (7.9%)	

Thalamus	15 (7.0%)	2 (1.8%)	
Ventricles	14 (6.5%)	2 (11.4%)	
Others	59 (27.6%)	1 (0.9%)	

CBTN: Children Brain Tumor Network; BCH: Boston Children's Hospital. The Kruskal-Wallis rank sum test (*) was performed for numerical data age to test the statistical significance between age medians. The Fisher's Exact test (+) was performed for categorical data to test the statistical significance differences between CBTN and BCH datasets. A p-value less than 0.05 is statistically significant.

Deep Learning Pipeline

The proposed pipeline for mutation class prediction operates in two stages (Fig. 1A). The initial stage involves T2W MRI preprocessing (Supplemental Methods: SM 2). and input to the tumor segmentation model, which is a pretrained nnUNet developed by our group in prior work. Briefly, we developed a pLGG auto-segmentation algorithm that demonstrated performance indistinguishable from human experts[17]. This first stage outputs a preprocessed, skull-stripped image along with a corresponding segmentation tumor mask (Fig. 1B) (Supplemental Methods: SM 3).

The second stage of the pipeline encompasses three binary subtype classifiers (BRAF Fusion vs. all; BRAF V600E vs. all; Wild-type vs. all), each specifically trained to identify one of the following classes: BRAF V600E, BRAF Fusion, and Wild-type. For each subtype classifier a ResNet50 model[18] was chosen as the fundamental encoder for extracting feature embeddings from 2D images, given its high performance on medical imaging classification problems [19 20] and the availability of pretrained network weights [21]. The fully connected layers succeeding the average pooling layer of the ResNet50 were replaced by a layer of 1024 neurons, and a final layer of single neurons for binary classification (Fig. 1D, Supplemental Methods: SM 3). Following binary classification from each binary subtype classifiers, a consensus decision block collates the predictions from the classifiers, yielding the overall mutational status prediction.

Subtype Classifier

Three individual binary subtype classifiers were trained, wild-type classifier, BRAF Fusion classifier, and BRAF V600E classifier. For the training, the multi-class development dataset, with instances of wild-type, BRAF fusion and BRAF V600E, was divided into three binary datasets in a "one Vs rest" format. Each subtype classifier was trained and inferred on the

corresponding One Vs rest dataset. For example, wildtype classifier was trained and inferred on the wild-type Vs rest binary dataset, similarly for the other subtype classifiers. The external validation dataset was split into three one Vs rest dataset for external validation of the individual subtype classifiers and the entire pipeline.

Model Training and Evaluation

Three different strategies were investigated for training individual binary classifiers. The initial approach, training from scratch, involved initializing the binary classifier model with random weights. For the second approach, called RadImageNet Finetune, the classifier model was initiated with pretrained weights from the RadImageNet [22] for the ResNet50 model. This prior initialization was intended to yield superior feature embeddings compared to random weight initialization and training from scratch or out-of-domain transfer learning [23].

The third approach, called TransferX, starts with pretrained weights from RadImageNet, but then adds two sequential stages of finetuning on separate, but related, classification tasks which act as pretext tasks for self-supervision, followed by a final finetuning on the target class (Fig. 1C). As an illustrative example, the training of a BRAF fusion classifier began with initialization via pretrained RadImageNet weights and sequential finetuning for BRAF V600E prediction, followed by Wild-type prediction, and finally finetuning for BRAF Fusion prediction. We hypothesized that combining transfer learning and self-supervised cross-training would enable the model to learn stronger, more generalizable features for mutational status prediction by exposure to different, though similar, classification problems. The Models were trained to minimize loss at the axial slice-level on the development dataset and internally tested on an internal validation set (25% of data randomly selected; Supplemental Methods: SM 3) and externally tested on external validation dataset.

Consensus Decision Block

Following binary classification, a consensus decision block collates the predictions from multiple binary classifiers, yielding the overall mutational status prediction of the pipeline (Supplemental Methods: SM 4). The consensus decision block was designed to emulate logic that would optimize the signal-to-noise ratio for mutational status prediction, particularly given the limited data scenario (Fig. 1E). We hypothesized that morphologic differences (and signal-to-noise ratio) between wild-type and any mutations are greater than between BRAF mutation subtypes (BRAF V600E and BRAF Fusion), thus wild-type mutation check is performed first. In instances

where the patient exhibits a wild-type mutation, signifying the absence of a BRAF mutation, the diagnostic process culminates. Conversely, if the patient possesses any BRAF mutation, further classification between BRAF Fusion and BRAF V600E mutations is performed. In this way, the use of sequential logic and binary classification form a rationale path to overall mutational status prediction and avoids the need for multi-class algorithms that would increase the risk of overfitting on a limited dataset. The final output of the consensus decision block and the pipeline consequently is a classification decision and its corresponding probability.

Center of Mass Distance Analysis (COMDist) to evaluate model attention

Gradient-weighted Class Activation Maps **(**GradCAM)[24] images are a common visualization tool for a model's focus within images (Fig. 4A), yet currently they are used for qualitative insights on where a model's attention is strongest for image classification. To enable the use of GradCAM as a quantitative performance evaluation tool, we developed "Center of Mass Distance" (COMDist), a quantifiable metric for comparing GradCAM images across different methodologies (Fig. 4C). COMDist calculates and averages the distance (in mm) between the tumor's center of mass (from the segmentation mask) and the center of mass of the GradCAM heatmap over the entire dataset, with smaller values indicating that the model is more accurately focusing on the tumor region (Fig. 4B). A COMDist score provides the clinical user with a metric to gauge whether the model is basing its prediction on intra-tumoral information (as one would expect) or extemporaneous information far from the tumor (indicating an implausible model "shortcut" that should not be trusted).

Performance Evaluation and Statistical Analysis

Since each of the MRI scan of each patient was factored into multiple tumor slice images (Supplemental Methods: SM 2), to generate aggregated patient-level prediction, the output probability scores of the individual 2D axial images were averaged to calculate the patient level probability score. The patient-level classification was then done by applying the threshold on the patient level probability score [Eq 1].

$$\text{Patient probability score} = \frac{\text{å image probability scores}}{\text{number of image slices for a given patient}} \quad \text{[Eq 1]}$$

The primary performance endpoint was the area under the curve (AUC) of receiver operating characteristic (ROC) at the patient-level. We calculated composite AUC and accuracy based on

a weighted average of the output of the three mutational subtype classifiers. The three DL approaches were initially evaluated on the internal test set (BCH), and the highest performing model was locked for external validation (CBTN). Secondary endpoints included sensitivity and specificity, precision, and accuracy, and were calculated using the model output, with threshold to optimize the Youden Index on the internal test set (Sensitivity + Specificity – 1). Post-hoc calibration was applied on the internal validation set and model calibration was assessed graphically pre- and post-calibration (Supplemental Methods: SM 5, Fig. S5). We compared AUC's for different models and calculated 95% Confidence Intervals (CIs) using the DeLong method [25]. The standard error of the AUC was calculated considering the numbers of positive and negative cases in the sample, and the derived variance of AUC. A two-sided *p*-value of <0.05 was considered statistically significant. Statistical metrics and curves were calculated using Scikit-learn packages [26] in Python v3.8.

RESULTS

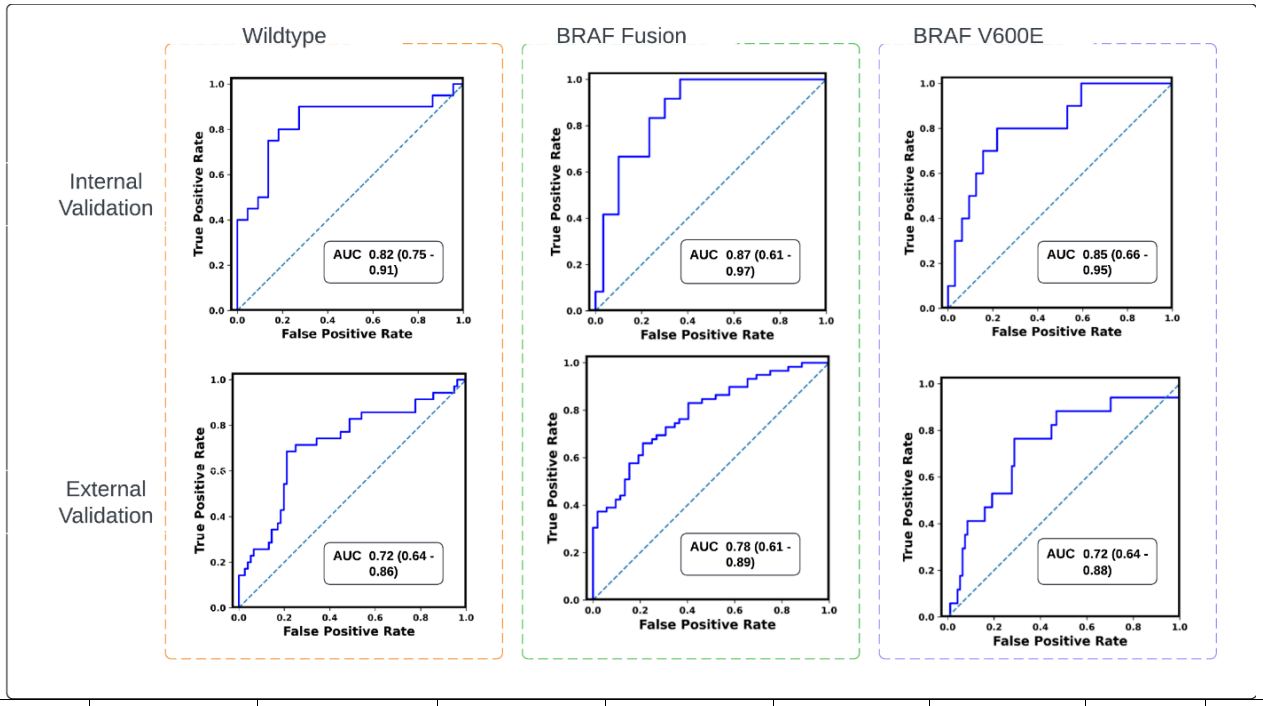
Patient Characteristics

The total pLGG patient cohort consisted of 326 pLGG patients from two cohorts, with 214 patients in the development set from BCH cohort and 112 patients in the external test set from CBTN (Table 1). Median age was 6 (range: 1-21) in the CBTN cohort and 8 (range: 1-25) in the BCH cohort. All patients had pathologically or clinically diagnosed grade I/II low-grade glioma, with a mixture of histologic subtypes and intracranial locations (Table 1). The developmental dataset contained 50 (23%), 60 (28%), and 104 (49%) patients with BRAF V600E, BRAF Fusion, and Wild-type, respectively, and the external validation dataset contained 17 (15%), 60 (53%), and 35 (32%) patients with BRAF V600E, BRAF Fusion, Wild-type, respectively (Table 1). Slice thickness, T2-repetition time, and T2-echo time were significantly different between BCH and CBTN datasets (Supplement Fig. S1 and S2). Univariate analysis showed that patient age (p=0.14) and sex (p=0.71) do not possess a strong association with BRAF Mutation classification decision boundary (Fig. S5, Fig. S6).

Table 2. *The pipeline's performance on classification on BRAF status for internal validation set and external validation set.*

	BRAF Status	AUC (95%CI)	Sensitivity	Specificity	Accuracy	Precision	Recall	F1-Score
Internal Validation	Wild-type	0.82 (0.75 - 0.91)	0.73	0.80	0.77	0.76	0.77	0.77

(n=59)	BRAF Fusion	0.87 (0.61 - 0.97)	0.87	0.70	0.81	0.81	0.80	0.80
	BRAF V600E	0.85 (0.66 - 0.95)	0.75	0.80	0.76	0.82	0.77	0.77
	Composite	0.84 (70 - 90)	0.77	0.76	0.77	0.78	0.77	0.77
External Validation (n=112)	Wild-type	0.72 (0.64 - 0.86)	0.72	0.71	0.72	0.75	0.72	0.73
	BRAF Fusion	0.78 (0.61 - 0.89)	0.60	0.90	0.75	0.77	0.74	0.74
	BRAF V600E	0.72 (0.64 - 0.88)	0.78	0.60	0.75	0.82	0.74	0.77



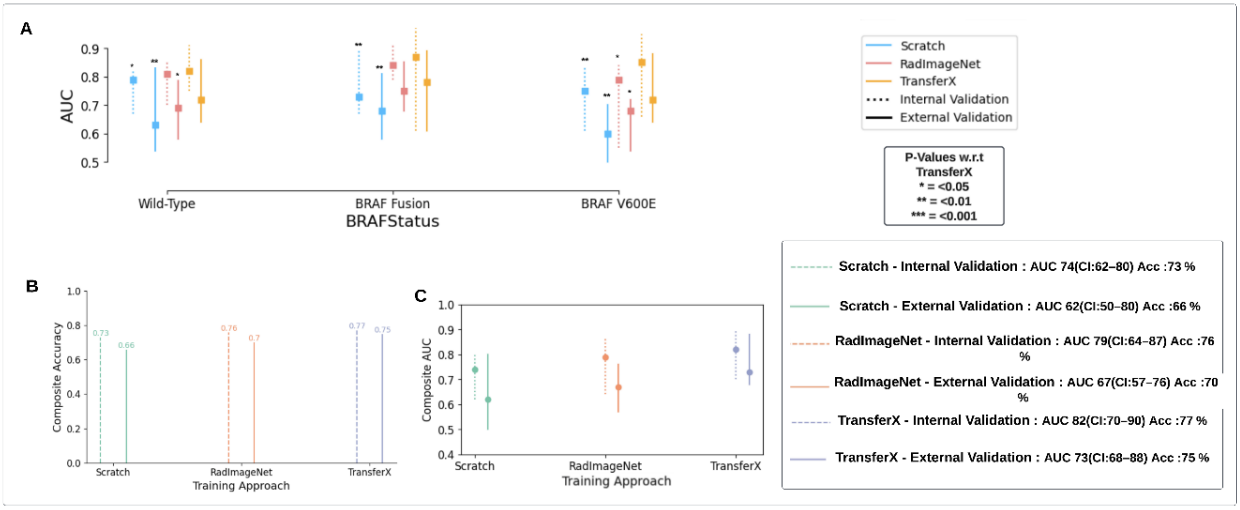
	Composite	0.73 (0.68 –0.88)	0.66	0.79	0.75	0.77	0.73	0.74

Figure 2. *Receiver operating characteristics (ROC) curves of the scan-to-prediction pipeline's predictions for all the three molecular subtype classes for internal validation (n=59) and external validation (n=112). The models, trained with TransferX, form the individual subtype classifiers. The outputs of the subtype classifiers are pooled using consensus logic, to result the pipeline predictions for each mutation class.*

TransferX improves deep learning model performance and generalizability

TransferX outperformed the pipeline with classifiers trained by RadImageNet FineTune and training from scratch for BRAF mutational status subtype prediction with composite classification AUC: 0.83 (95% CI 0.71-0.88) and 77% accuracy on internal validation, compared to AUC: 0.74 (95% CI 0.62-0.80) and 73% for training from scratch (Fig. 3B, Fog. 3C). All training approaches, including TransferX, were most accurate at identifying BRAF fusion, followed by wild-type and V600E, though TransferX was the only approach to maintain AUC > 0.80 for all individual subtype classifications (Fig. 3A).

On external validation, there was a mild degradation in performance across all approaches, with TransferX still demonstrating the highest performance with composite AUC 0.73 (95% CI: 0.68 – 0.88) and 75% accuracy (Fig. 3C). TransferX also demonstrated best performance for classification of wildtype vs any BRAF mutational class with AUC 0.82 (95% CI: 0.75 – 0.91) and 77% accuracy (Table 2, Fig. 3A). TransferX showed adequate calibration on the external validation set, which was further improved after calibrating the model on the internal validation set (Fig. S7).



TransferX also resulted in superior performance compared to other training approaches when subtype classifiers (without consensus logic) were tested on the internal and external validation set for each subtype class (Fig. S8)

Figure 3. *(A) AUC is plotted and compared for the pipeline results with individual subtype classifiers trained using different training approaches (Scratch, RadImageNet FineTune,*

TransferX) for respective mutation class (BRAF wild-type versus fusion versus V600E). P-values are generated from model comparisons with respect to TransferX. (B) Accuracy and (C) AUC comparison of the pipeline with individual subtype classifiers trained with three different training approaches. The composite Accuracy and AUC for the entire dataset is calculated by the weighted average of the AUCs and Accuracy across the three mutational classes. AUC: area under the curve.

Table 3. *Median COMDist value (mm) comparison for three training approaches, of each subtype classifier on its corresponding mutation class data.*

	BRAF Status	**TransferX**	**Scratch**	**RadImageNet**
Internal Validation (n=59)	Wild-Type	38.02	41.54 (p=0.09)	39.48 (p=0.46)
	BRAF Fusion	25.8	27.14 (p=0.49)	26.13 (p=0.86)
	BRAF V600E	33.02	36.86 (p=0.09)	34.40 (p=0.52)
External Validation (n=112)	Wild-Type	27.8	28.11 (p=0.90)	34.2 (p=0.009)
	BRAF Fusion	28.0	29.7 (p=0.47)	28.7 (p=0.76)
	BRAF V600E	23.03	25.24 (p=0.40)	25.21 (p=0.40)

TransferX yields more accurate model attention

GradCAMs were generated for the three approaches on all cases (Fig. 4A), and corresponding COMDist scores were calculated. TransferX consistently yielded the best average COMDist scores across all classification tasks, indicating improved model focus on intra- and peritumoral regions (Table 3 & Fig. 4C).

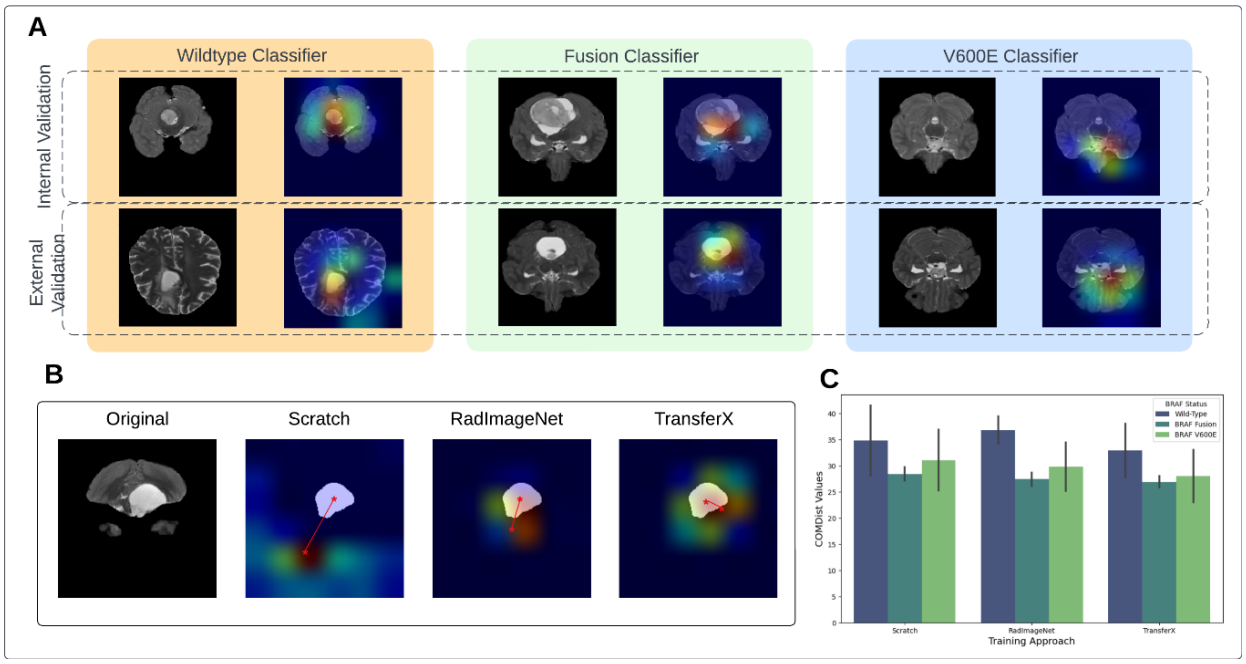


Figure 4. *(A) GradCAM image overlay for each mutational class for internal and external validation sets. (B) COMDist representation for three training approaches. (C) COMDist value comparison of the scan-to-prediction pipeline for each molecular subtype class, with corresponding individual subtype classifiers trained with three different training approaches. GradCam: Gradient-weighted Class Activation Maps; COMDist: Center of Mass Distance.*

DISCUSSION

Pediatric low-grade gliomas can arise in locations that make resection, and even biopsy, morbid and infeasible. In these situations, the ability to noninvasive detect BRAF mutational status via diagnostic imaging would be helpful to determine which patients may benefit from targeted therapeutics that act on the BRAF pathway and enrollment in clinical trials of novel targeted therapies. In this study, we developed and externally validated a scan-to-prediction algorithm to noninvasively predict BRAF mutational status that could be used in settings where tissue diagnosis is infeasible. The limited quantity of data available for pediatric brain tumor analysis has limited the impact of artificial intelligence in detection of such rare occurring tumors compared to other malignancies. Our study surmounts this obstacle by combining elements of transfer learning and self-supervision to develop high performing model that maintains good performance on testing at an external institution with heterogeneous tumor and scanner characteristics. Additionally, we introduced COMDist, a quantitative metric to evaluate

information captured in attention maps that will help make medical imaging algorithms more interpretable to end users and clinicians. Our study findings contribute to bridging the gap between artificial intelligence development and clinical translation of medical imaging classification tools in a limited data scenario. To this end, we have published the code and pretrained models to provide usable tools for the scientific community and to encourage clinical testing.

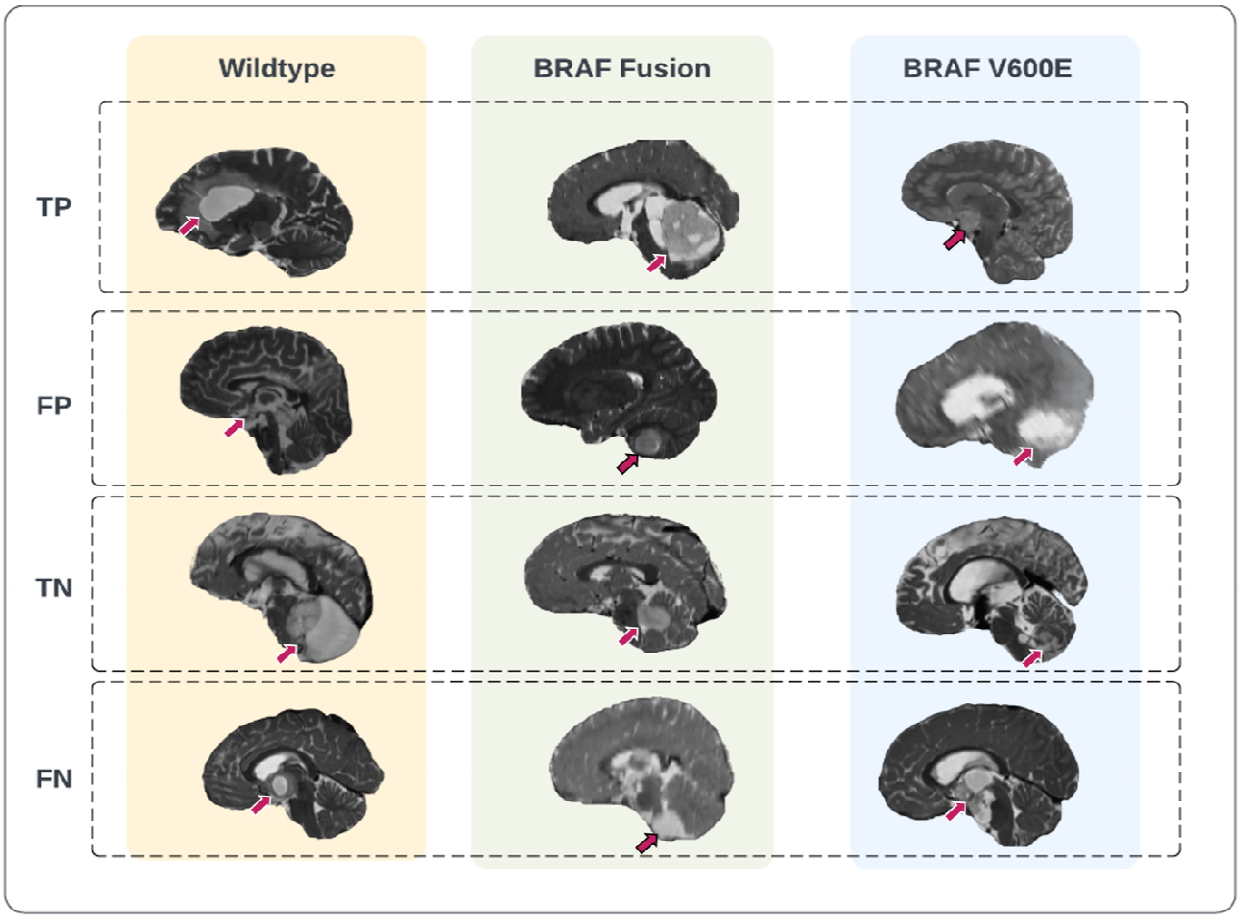


Figure 5. *Representative prediction cases of the scan-to-prediction pipeline on the external dataset. The final scan-to-prediction pipeline consists of three subtype classifiers, trained using TransferX, further pooled together in consensus logic by the consensus decision block. Tumor lesions in the T2-weighted images were highlighted with arrows. TP: true positive; FP: false positive; TN: true negative; FN: false negative.*

With the emergence of the novel BRAF-directed therapies, the segregation of wild-type tumor cases from BRAF subtypes in pLGG has become critical. With an External AUC > 0.71 for classifying wild-type tumor cases vs BRAF cases, the pipeline can be used as an assistive tool

by clinicians to provide key information in settings where tissue biopsy is infeasible or low-resource settings that preclude genomic analysis. Beyond BRAF classification, the pipeline's ability to identify BRAF V600E, specifically, position its use as a means to select patients for specific V600E inhibitors such as of dabrafenib and trametinib which has shown better progression free survival than chemotherapy[27,28]. While there is a mild performance degradation on external dataset, the differences in MR parameters between these datasets are notable (Fig. S1, Fig. S2). Similarity in these parameters would result in comparable performance on external validation. Importantly, the scan-to-prediction pipeline is practical and not reliant on manual segmentation which is resource-intensive and requires specialized expertise. The pipeline also exhibits robust performance, with external AUC of 0.74 (n=28), in the classification of BRAF mutation status, particularly with tumor cases originating from traditionally challenging regions for biopsy such as the Optic Pathway, Thalamus, and Brainstem. This allows for the diagnosis, followed by directed therapy, of these challenging tumor locations in a much safer manner.

Classification for mutations in pLGGs have been previously attempted by a few studies, with manual segmentation-derived, pre-engineered radiomics being the more common approach. Radiomic features have been extracted from MRI images and fitted to classifiers models like XGboost, SVMs [1,29]. The sensitivity of the dataset size on BRAF mutation classification performance was studied by Wagner et al. [30] in a radiomics based study. They showed that Neural networks outperform XGBoost for classification AUC and that the performance was affected by the size of the data used in training. In general, imaging-based methods have not seen much success in BRAF mutation prediction, given the limited data availability and likely low signal-to-noise, in terms of geometrical features of tumors. We demonstrate here that inter-class cross training can lead to more meaningful training rounds with limited data and improve performance. This idea was been explored more generally by Muhamedrahimov et al. [31] by relaxing the assumption of independence between multiple categories. TransferX expands on this work by completely dropping the assumptions of independence between different categories of a multiclass dataset with stepwise inter-class training as a pretext task to learn robust feature representations. Furthermore, incorporating consensus decision logic to combine multiple binary classifiers also helped mitigate overfitting from the limited dataset. For BRAF Mutations in LGGs tumor location has a significant correlation with the categorization of gliomas (Fig. S9), this positional information is picked by TransferX as the sequential fine-tuning process allows the models to learn the spatial dependence of the different mutations, hence leading to robust classification performance.

Interpretability is a well-recognized important factor for deep learning models for clinical translation. A variety of metrics like saliency maps, guided backpropagation have been developed to depict the pixels that are contributing for the maximum activation in the network and hence being more significant for classification[32 33]. Another approach which has been very popular recently is GradCAM[24]. Although adding a degree of qualitative interpretability, the GradCAM approach has only allowed for case-by-case visualizations for the end-user, which are not very useful when trying to establish trust of a model overall. We expand the utility of GradCAM in this work with COMDist. By incorporating spatial knowledge of the tumor from auto-segmentation, COMDist can quantify, in terms of distance, the model's attention with respect to the correct, biologically rational region of interest in the image. We expect this methodology will be valuable for the AI research community as well as clinical end-users evaluating and implementing medical imaging AI applications in clinic.

Limitations

There are several limitations to this work. Firstly, this work is retrospective in nature and subject to the biases of our patient samples. We attempted to mitigate this effect of bias by using a blinded, external validation set. Thus, we would encourage further independent validation of our results, including prospective testing. Additionally, the pipeline is exclusively based on T2W MRI scans. While T2W images are the most common and available diagnostic sequence for pLGG, T1c, T1, and FLAIR may contain complementary information that enhances performance, which we aim to explore in future work. In this work, we decided to leverage a 2D approach with slice-averaging to minimize overfitting on our limited data set. It is possible that with further data collection a 3D approach may work better, however this would significantly increase the model parameter size and thus make the model even more prone to overfitting.

Conclusions

In summary, we developed and externally validated a scan-to-prediction pipeline to analyze T2W MRI as input and output BRAF mutational subtype for pediatric low-grade glioma. We leveraged a novel combination of transfer learning and self-supervision to mitigate overfitting and develop a high-performing and generalizable model. We also proposed a novel evaluation metric, COMDist, that can be used to further assess performance and interpretability of AI imaging models. Our resulting pipeline warrants prospective validation to determine if it could be clinically used in settings where tissue and/or genomic testing is unavailable.

Funding

This study was supported in part by the National Institutes of Health (NIH) (U24CA194354, U01CA190234, U01CA209414, R35CA22052, and K08DE030216), the National Cancer Institute (NCI) Spore grant (2P50CA165962), the European Union – European Research Council (866504), the Radiological Society of North America (RSCH2017), the Pediatric Low-Grade Astrocytoma Program at Pediatric Brain Tumor Foundation, and the William M. Wood Foundation.

Competing Interests

All the authors declare no competing interests.

Author Contributions

Study design: D.T., Z.Y. and B.H.K.; code design, implementation and execution: D.T. and Z.Y.; acquisition, analysis or interpretation of data: D.T., Z.Y., A.Z., and B.H.K.; writing of the manuscript: D.T., Z.Y., B.H.K.; critical revision of the manuscript for important intellectual content: all authors; statistical analysis: Y.Z. and D.T.; study supervision: B.H.K., H.J.W.L.A., T.P., and D.H.K.

Code availability

The code of the deep learning system, as well as the trained model and statistical analysis are publicly available at the GitHub webpage: https://github.com/DivyanshuTak/BRAF_Classification.

References

1. Radiomics of Pediatric Low-Grade Gliomas: Toward a Pretherapeutic Differentiation of BRAF- Mutated and BRAF-Fused Tumors. *AJNR Am. J. Neuroradiol.* vol. 42 (2021).
2. Talloa, D. *et al.* BRAF and MEK Targeted Therapies in Pediatric Central Nervous System Tumors. *Cancers* **14**, (2022).
3. KIAA1549: BRAF Gene Fusion and FGFR1 Hotspot Mutations Are Prognostic Factors in Pilocytic Astrocytomas. *J. Neuropathol. Exp. Neurol.* vol. 74 (2015).

4. Marker, D. F. & Pearce, T. M. Homozygous deletion of CDKN2A by fluorescence in situ hybridization is prognostic in grade 4, but not grade 2 or 3, IDH-mutant astrocytomas. *Acta Neuropathol Commun* **8**, 1–12 (2020).

5. Sievert, A. J. & Fisher, M. J. Pediatric Low-Grade Gliomas. *J. Child Neurol.* **24**, 1397–1408 (2009).

6. Lundervold, A. S. & Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z. Für Med. Phys.* **29**, 102–127 (2019).

7. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

8. Razzak, M. I., Naz, S. & Zaib, A. Deep Learning for Medical Image Processing: Overview, Challenges and the Future. in *Classification in BioApps: Automation of Decision Making* (eds. Dey, N., Ashour, A. S. & Borra, S.) 323–350 (Springer International Publishing, 2018). doi:10.1007/978-3-319-65981-7_12.

9. DeepGlioma: AI-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. (2023).

10. Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L. & Liotta, A. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Comput. Methods Programs Biomed.* **223**, 106951 (2022).

11. Brigato, L. & Iocchi, L. A Close Look at Deep Learning with Small Data. Preprint at http://arxiv.org/abs/2003.12843 (2020).

12. Namdar, K. *et al.* Tumor-location-guided CNNs for Pediatric Low-grade Glioma Molecular Biomarker Classification Using MRI. Preprint at http://arxiv.org/abs/2210.07287 (2022).

13. Children's Brain Tumor Network. https://cbtn.org/.

14. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* **13**, 1 (2015).

15. SimpleITK - Home. https://simpleitk.org/.

16. HD-BET. (2023).

17. Boyd, A. *et al.* Expert-level pediatric brain tumor segmentation in a limited data scenario with stepwise transfer learning. *medRxiv* 2023.06.29.23292048 (2023) doi:10.1101/2023.06.29.23292048.

18. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016). doi:10.1109/CVPR.2016.90.

19. Wang, W. *et al.* Medical Image Classification Using Deep Learning. in *Deep Learning in Healthcare: Paradigms and Applications* (eds. Chen, Y.-W. & Jain, L. C.) 33–51 (Springer International Publishing, 2020). doi:10.1007/978-3-030-32606-7_3.

20. Sarwinda, D., Paradisa, R. H., Bustamam, A. & Anggia, P. Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer. *Procedia Comput. Sci.* **179**, 423–431 (2021).

21. RadImageNet. (2023).

22. RadImageNet | Medical Image Artificial Intelligence. *My Site* https://www.radimagenet.com.

23. Ravishankar, H. *et al.* Understanding the Mechanisms of Deep Transfer Learning for Medical Images. in *Deep Learning and Data Labeling for Medical Applications* (eds. Carneiro, G. et al.) 188–196 (Springer International Publishing, 2016). doi:10.1007/978-3-319-46976-8_20.

24. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).

25. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**, 837–845 (1988).

26. sklearn.metrics.plot_roc_curve. *scikit-learn* https://scikit-learn/stable/modules/generated/sklearn.metrics.plot_roc_curve.html.

27. Dual Targeted Treatment Improves Response Over Chemotherapy in BRAF V600+ Pediatric Low-Grade Glioma. *Cancer Network* https://www.cancernetwork.com/view/dual-targeted-treatment-improves-response-over-chemotherapy-in-braf-v600-pediatric-low-grade-glioma (2022).

28. Nobre, L. *et al.* Outcomes of BRAF V600E Pediatric Gliomas Treated With Targeted BRAF Inhibition. *JCO Precis. Oncol.* **4**, PO.19.00298 (2020).

29. Xu, J. *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.* **222**, 107478 (2022).

30. Wagner, M. W. *et al.* Dataset size sensitivity analysis of machine learning classifiers to differentiate molecular markers of paediatric low-grade gliomas based on MRI.

31. Raouf, M., Amir, B. & Ayelet, A.-B. Learning Interclass Relations for Image Classification. Preprint at http://arxiv.org/abs/2006.13491 (2020).

32. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Preprint at https://doi.org/10.48550/arXiv.1312.6034 (2014).

33. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for Simplicity: The All Convolutional Net. Preprint at https://doi.org/10.48550/arXiv.1412.6806 (2015).