Deep Learning for Pediatric Posterior Fossa Tumor Detection and Classification: A Multi-Institutional Study

J.L. Quon, W. Bala, L.C. Chen, J. Wright, L.H. Kim, M. Han, K. Shpanskaya, E.H. Lee, E. Tong, M. Iv, J. Seekins, M.P. Lungren, K.R.M. Braun, T.Y. Poussaint, S. Laughlin, M.D. Taylor, R.M. Lober, H. Vogel, P.G. Fisher, G.A. Grant, V. Ramaswamy, N.A. Vitanza, C.Y. Ho, M.S.B. Edwards, S.H. Cheshier and K.W. Yeom

AJNR Am J Neuroradiol 2020, 41 (9) 1718-1725
doi: https://doi.org/10.3174/ajnr.A6704
http://www.ajnr.org/content/41/9/1718

Deep Learning for Pediatric Posterior Fossa Tumor Detection and Classification: A Multi-Institutional Study

J.L. Quon, W. Bala, L.C. Chen, J. Wright, L.H. Kim, M. Han, K. Shpanskaya, E.H. Lee, E. Tong, M. Iv, J. Seekins, M.P. Lungren, K.R.M. Braun, T.Y. Poussaint, S. Laughlin, M.D. Taylor, R.M. Lober, H. Vogel, P.G. Fisher, G.A. Grant, V. Ramaswamy, N.A. Vitanza, C.Y. Ho, M.S.B. Edwards, S.H. Cheshier, and K.W. Yeom

ABSTRACT

BACKGROUND AND PURPOSE: Posterior fossa tumors are the most common pediatric brain tumors. MR imaging is key to tumor detection, diagnosis, and therapy guidance. We sought to develop an MR imaging–based deep learning model for posterior fossa tumor detection and tumor pathology classification.

MATERIALS AND METHODS: The study cohort comprised 617 children (median age, 92 months; 56% males) from 5 pediatric institutions with posterior fossa tumors: diffuse midline glioma of the pons ($n = 122$), medulloblastoma ($n = 272$), pilocytic astrocytoma ($n = 135$), and ependymoma ($n = 88$). There were 199 controls. Tumor histology served as ground truth except for diffuse midline glioma of the pons, which was primarily diagnosed by MR imaging. A modified ResNeXt-50-32x4d architecture served as the backbone for a multitask classifier model, using T2-weighted MRIs as input to detect the presence of tumor and predict tumor class. Deep learning model performance was compared against that of 4 radiologists.

RESULTS: Model tumor detection accuracy exceeded an AUROC of 0.99 and was similar to that of 4 radiologists. Model tumor classification accuracy was 92% with an F_1 score of 0.80. The model was most accurate at predicting diffuse midline glioma of the pons, followed by pilocytic astrocytoma and medulloblastoma. Ependymoma prediction was the least accurate. Tumor type classification accuracy and F_1 score were higher than those of 2 of the 4 radiologists.

CONCLUSIONS: We present a multi-institutional deep learning model for pediatric posterior fossa tumor detection and classification with the potential to augment and improve the accuracy of radiologic diagnosis.

ABBREVIATIONS: PF = posterior fossa; EVD = external ventricular drain; CAMs = class activation maps; DMG = diffuse midline glioma of the pons; EP = ependymoma; MB = medulloblastoma; PA = pilocytic astrocytoma; PF = posterior fossa; ROC = receiver operating characteristic; t-SNE = t-distributed stochastic neighbor embedding

Pediatric brain tumors are the most common solid cancer and the leading cause of cancer-related deaths in children, with approximately 4600 new diagnoses per year in the United States alone.[1,2] MR imaging plays a key role in tumor detection, and preliminary imaging diagnosis[3] helps guide initial management.

While the final diagnosis and treatment depend on surgical specimens, accurate classification before surgery can help optimize the surgical approach and the extent of tumor resection. MR imaging contributes to presurgical planning by defining the spatial relationship of the tumor within the brain. In addition, it allows high-dimensional image-feature analysis[4] that

Received January 12, 2020; accepted after revision May 27.

From the Departments of Neurosurgery (J.L.Q., G.A.G., M.S.B.E.), Electrical Engineering (E.H.L.), Radiology (E.T., M.I.), and Pathology (H.V.), Stanford University, Stanford, California; Department of Radiology (W.B., J.S., M.P.L., K.W.Y.), and Division of Child Neurology (P.G.F.), Lucile Packard Children's Hospital, Stanford University, Palo Alto, California; Department of Urology (L.C.C.), Stanford University School of Medicine (L.H.K., M.H., K.S.), Stanford, California; Department of Radiology (J.W.), Seattle Children's Hospital, University of Washington School of Medicine, Seattle, Washington; Departments of Clinical Radiology & Imaging Sciences (K.R.M.B., C.Y.H.), Riley Children's Hospital, Indiana University, Indianapolis, Indiana; Departments of Radiology (T.Y.P.), Boston Children's Hospital, Boston, Massachusetts; Departments of diagnostic Imaging (S.L.), and Neurosurgery (M.D.T.), and Haematology/Oncology (V.R.), The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada; Department of Neurosurgery (R.M.L.), Dayton Children's Hospital, Wright State University Boonshoft School of Medicine, Dayton, Ohio; Division of Pediatric Hematology/Oncology (N.A.V.), Department of Pediatrics, University of Washington, Seattle Children's Hospital, Seattle Washington; Fred Hutchinson Cancer Research Center (N.A.V.), Seattle, Washington; and Departments of Neurosurgery (S.H.C.), University of Utah School of Medicine, Salt Lake City, Utah.

Paper previously presented, in part, at: Annual Meeting of the American Academy of Neurological Surgery/Congress of Neurological Surgery, Section on Pediatric Neurological Surgery.

J.L. Quon and W. Bala contributed equally to this work.

Please address correspondence to Kristen W. Yeom, Department of Radiology, Lucile Packard Children's Hospital, Stanford University School of Medicine, 725 Welch Rd, MC 5654, Palo Alto, CA 94304; e-mail: kyeom@stanford.edu

Indicates article with supplemental on-line tables.

Indicates article with supplemental on-line photos.

http://dx.doi.org/10.3174/ajnr.A6704

can potentially be correlated to the molecular profiling[5-8] included in recent updates to the World Health Organization brain tumor classification system.[9]

Modern advances in computing power and machine learning tools such as deep learning can augment real-time clinical diagnosis.[10,11] Deep learning is an improvement over radiomics and other traditional machine learning approaches that use labor- and time-intensive handcrafted feature extraction.[3,4,11] In this study, we aimed to develop an MR imaging–based deep learning model for predicting pediatric posterior fossa (PF) tumor pathology and to compare its performance against that of board-certified radiologists. We targeted PF tumors, given their high incidence in the pediatric population and leveraged a large, multi-institutional image dataset for deep learning.

MATERIALS AND METHODS

Study Cohort

Data-use agreements were developed between the host institution (Stanford Lucile Packard Children's Hospital) and 4 academic pediatric hospitals across North America (The Hospital for Sick Children, Seattle Children's, Indiana Riley Children's, Boston Children's) for this retrospective, multicenter study, after institutional review board approval at each institution. The following served as the inclusion criteria for 803 patients with tumors: brain MR imaging of treatment-naïve PF brain tumors: medulloblastoma (MB), ependymoma (EP), pilocytic astrocytoma (PA), and diffuse midline glioma of the pons (DMG, formerly DIPG); and tissue specimens that served as ground truth pathology except for DMG. A subset of patients were included who required emergent ventricular drain placement before tumor resection or other therapies. Brain MR imaging from 199 children without brain tumors were randomly sampled from the normal database of the host institution to serve as controls. A board-certified pediatric neuroradiologist (K.W.Y. with >10 years' experience), with a Certificate of Added Qualification, visually inspected all scans for quality control to confirm that they met the inclusion criteria.

The study cohort was subdivided into development (training and validation) and held-out test sets using stratified random sampling by tumor subtype. For tumor MRIs, the breakdown was 70% and 10% for the training/validation sets and 20% for the test set. For patients with >1 preintervention scan, all scans of that patient were included in either the development or test set, with no crossover. For control MRIs without tumor, data distribution was 10% and 90% for the validation and held-out test sets, respectively, as normal MRIs were not used to train the model.

MR Imaging Protocols

MR imaging scans were obtained at 1.5 or 3T at multiple centers with equipment from the following vendors: GE Healthcare, Siemens, Philips, and Toshiba Medical Systems (Canon Medical Systems). The T2 scans were the following: T2 TSE clear/sense, T2 FSE, T2 PROPELLER, T2 BLADE (Siemens), T2 drive sense (TR/TE = 2475.6–9622.24/80–146.048; slice thickness = 1–5 mm with 0.5-or 1-mm skip; matrix ranges = 224–1024 × 256–1024. T1 postgadolinium scans included T1 MPRAGE, T1 BRAVO (GE Healthcare), T1 fast-spoiled gradient recalled, T1 spoiled gradient-echo, and T1 spin-echo. ADC maps were created using a mono exponential algorithm with b-values from 0 to 1000 s/mm^2, varying by institution.

Image Processing and Data Augmentation

Axial DICOM images were processed using the Python language with the pydicom (https://pypi.org/project/pydicom/) and SimpleITK (https://anaconda.org/SimpleITK/simpleitk) packages. Images were resampled to 256 × 256 pixels in the axial dimension. Slice thickness was not modified. Data augmentation was performed by incorporating random flips, rotations, translations, and crops to 224 × 224 pixels to improve model generalizability. Gray-scale images were fed as RGB color images into an adapted ResNeXt model (https://github.com/titu1994/Keras-ResNeXt).

Ground Truth Labels

Pathology from surgical specimens served as ground truth (MB, EP, PA) except for most patients with DMG who were diagnosed primarily by MR imaging. An attending pediatric neuroradiologist (K.W.Y.) manually classified each axial slice as having tumor versus no tumor: A slice was considered positive if any tumor was visible.

Deep Learning Model Architecture

We chose a 2D ResNeXt-50-32x4d deep learning architecture (https://github.com/titu1994/Keras-ResNeXt) rather than a 3D architecture, given the wide variation in slice thickness across scans. Transfer learning was implemented using weights from a model pretrained on ImageNet (http://image-net.org/),[12] a consortium of >1.2 million images with 1000 categories (On-line Fig 1*A*), for all layers except the final fully connected layer, which was modified to predict 1 of 5 categories: no tumor, DMG, EP, MB, or PA. The model was trained to minimize cross-entropy loss, or error, between the predicted and actual tumor type. The architecture was modified to predict the relative slice position of tumor tissue within the entire scan, calculated by interpolating the most inferior axial slice as zero and the most superior as 1 (On-line Fig 1*B*). Relative slice position was included to account for differences in slice thickness in the z-plane across different scans. Thus, position was normalized to each individual patient. With normalization, the zero position referred to the foramen magnum; the 1 position, to the vertex; and 0.5 varied slightly between the upper midbrain and the midbrain-thalamic junction, depending on head size and image acquisition. This component was trained to minimize mean-squared loss between the predicted-versus-actual slice location. Setting the slice position contribution to 10% of the total loss had the most improvement (On-line Table 1). A final ensemble of 5 individual models was used to generate a confidence-weighted vote for the predicted class for each slice (On-line Fig 1*C*). To generate the model prediction for the entire scan, we aggregated all slice-level predictions. Scans with a proportion of tumor slices that exceeded a certain threshold were considered to have tumor (On-line Fig 1*D*). Based on the results from our training and validation sets, the minimal detection threshold was set to 0.05. For scans predicted to have tumor, the model then predicted the tumor subtype using a confidence-weighted voting system (On-line Fig 1*E*).

Table 1: Complete dataset of 803 patients from 5 institutions with 4 tumor types[a]

	Institution 1	Institution 2	Institution 3	Institution 4	Institution 5
MB	90	117	20	30	41
DMG	85	0	0	45	21
EP	42	41	41	22	8
PA	129	0	0	45	26

[a] One hundred eighty-six patients with no T2 sequences or only postintervention imaging were excluded.

Table 2: A total of 739 scans were distributed into a training set, a validation set, and a held-out test set

	Training	Validation	Test	Total
MB	242	34	55	331
DMG	88	10	24	122
EP	83	13	15	111
PA	114	20	41	175
Total	527	77	135	739

Model Training

An Ubuntu computer (https://ubuntu.com/download) with 4 TitanXp Graphic Processing Units (NVIDA) with 12 GB of memory was used for model development. Batch size was 160 slices per iteration. Training was performed using Adam optimization with an initial learning rate of 0.003 for 50 epochs and a cosine annealing learning rate decay to zero. Drop-out was set to 10% in the final fully connected layer to reduce overfitting. All model layers were fined-tuned throughout training. Models were saved if they improved validation set performance following a 10-epoch patience period. The top 5 models with the best validation results were selected for the final slice-level ensemble model.

Model Evaluation

Tumor-detection accuracy was evaluated based on whether the model correctly predicted the presence or absence of a tumor for the entire scan. Receiver operating characteristic (ROC) curves were generated by varying the set threshold for the proportion of tumors slices. For tumor classification, the F_1 score was calculated as the harmonic mean of precision (positive predictive value) and recall (sensitivity). Sensitivity and specificity for each tumor type were calculated by grouping all of the nontarget tumors together as negative examples.

Radiologist Interpretation

Board-certified attending radiologists with Certificates of Added Qualification in either Pediatric Radiology (J.S. with >10 years' experience; M.P.L. with >5 years' experience) or Neuroradiology (M.I. with >5 years' experience; E.T. with >2 years' experience) were given all T2 scans from the held-out test set and asked to detect tumors and select pathology among the 4 subtypes (MB, EP, PA, DMG). Radiologists were blinded to the ground truth labels and other clinical information and allowed to interpret at their own pace. They were permitted to window the scans and view in all orientations (axial, sagittal, or coronal).

Comparative Performance and Statistical Analysis

Subgroup analysis of model classification accuracy was performed using a Fisher's exact test. Radiologists' tumor detection sensitivity and specificity were plotted against the tumor-detection ROC curve of the model. Model and radiologists' tumor-detection and classification accuracy were compared using McNemar's test, with a P value threshold of .05.

RESULTS

PF Tumor Dataset

Of 803 patients with the 4 tumor types from 5 pediatric hospitals (Table 1), we excluded 186 patients due to lack of T2 scans, resulting in a total of 617 patients with tumors. Ages ranged from 2.5 months to 34-years old (median, 81 months); 56% were boys. Some patients had multiple preintervention scans from different dates, resulting in a total of 739 T2 scans. The training, validation, and test sets included 527, 77, and 135 scans, respectively (Table 2).

Deep Learning Model

Given that radiologists benefit from using multiple image sequences, we isolated a subset of the tumor cohort ($n = 260$ scans) with all 3 MR imaging sequences (T2-weighted, T1-weighted post-gadolinium, and ADC). To identify the MR imaging sequences most likely to allow successful model development, we compared the use of these 3 sequences versus a single T2-weighted scan (T2-scan) as model input. Surprisingly, we found superior initial model performance with T2-scans alone (On-line Table 2) and thus focused on T2-scans. Given that T2-based MRIs are also commonplace among clinical protocols for the initial evaluation of clinical symptoms, a deep learning model using T2 alone would also be more broadly applicable.

Several convolutional deep learning approaches, including the ResNet, ResNeXt, and DenseNet (https://towardsdatascience.com/densenet-2810936aeebb) architectures with varying numbers of layers as well as the InceptionV3 architecture (https://blog.paperspace.com/popular-deep-learning-architectures-resnet-inceptionv3-squeezenet/), were evaluated on a subset of the training data. Preliminary experiments demonstrated that the ResNeXt-50-32x4d architecture best balanced accuracy with computational cost. Our final model architecture consisted of modified 2D ResNeXt-50-32x4d residual neural networks to generate a prediction for each axial slice in the scan (On-line Fig 1*A*). The baseline ResNeXt-50-32x4d, which classified each T2 axial slice as no tumor, MB, EP, PA, or DMG, achieved an F_1 score of 0.60 per axial slice. Given that radiologists and clinical experts often use tumor location to assess brain tumors, we modified the architecture for multitask learning to also predict the relative position of each slice, which improved performance by 4% (On-line Fig 1*A*-, *B*). Because prior studies have shown that combining multiple individual models improves overall performance by reducing variance between predictions,[13] we created an ensemble model comprising the 5 best-performing individual models (On-line Fig 1*C*), as this further improved accuracy while maintaining reasonable computational requirements (On-line Table 3).

To generate scan-level predictions, we then tallied all individual slice predictions (tumor versus no tumor) using a confidence-based voting algorithm (On-line Fig 1*D*). This schema resulted in accurate scan-level prediction of tumor versus no tumor with an area under the ROC curve of 0.99. Setting the

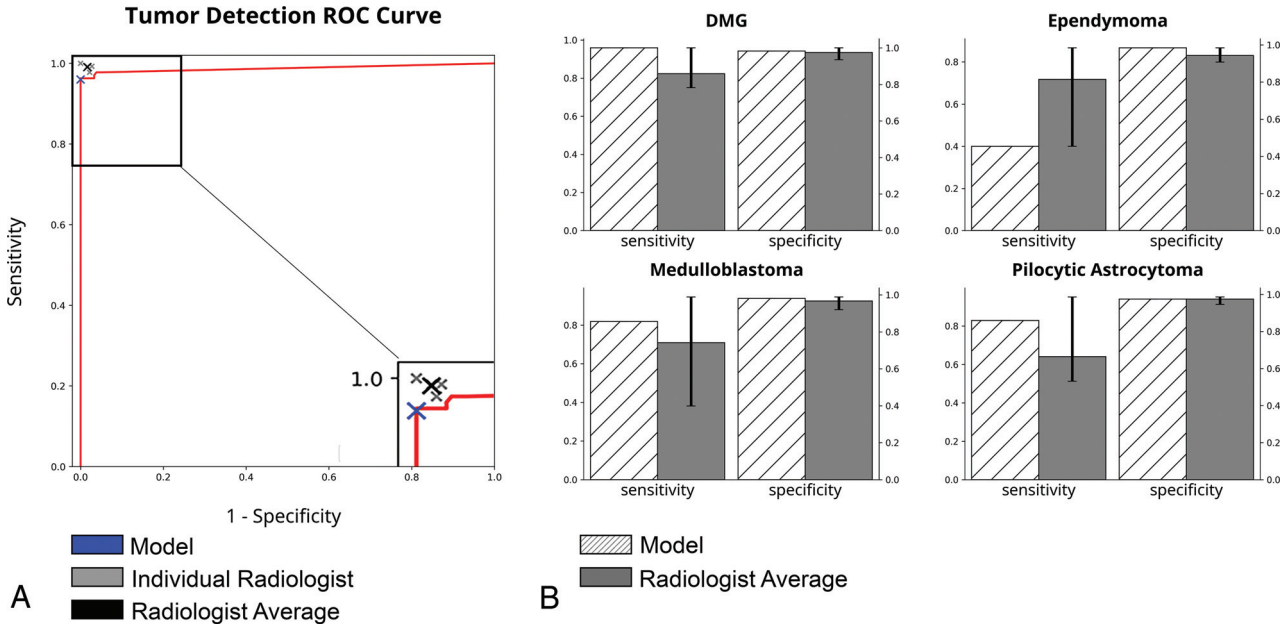

FIG 1. Comparison of model-with-radiologist performance. *A*, ROC curve for scan-level tumor detection. Model, individual radiologist, and average radiologist performance are indicated with *crosshairs*. *B*, Model and average radiologist performance for tumor subtype classification results. *Error bars* represent standard error among radiologists.

threshold at 5% (at least 1 tumor slice per every 20 slices) allowed maximal specificity and a sensitivity of at least 95% in the validation set. A 5% threshold achieved a sensitivity of 96% and a specificity of 100% on the held-out test set (Fig 1). Final scan-level tumor-type classification accuracy was 92% with an F_1 score of 0.80. Subgroup analysis demonstrated no difference in classification accuracy between patients younger and older than 2 years of age ($P = .22$) and no difference between patients with tumor with and without external ventricular drains (EVDs) ($P = .50$).

Class Activation Maps for Discriminative Localization of Tumor Type

Internal operations of deep learning algorithms often appear opaque and have been referred to as a "black box." Post hoc approaches for interpreting results have been described, such as using class activation maps (CAMs) to improve transparency and understanding of the model.[14] CAMs can serve as a quality assurance tool such that they highlight image regions relevant to the model's prediction and denote the model's confidence in the prediction but are not intended to precisely segment tumor voxels.[15] We implemented CAMs to visualize which regions of the image were most contributory to model prediction (Fig 2).[16] Qualitatively, pixels in close vicinity to the tumor appeared to strongly influence correct predictions, whereas incorrect predictions showed scattered CAMs that prioritized pixels in non-tumor regions. Because CAMs are not intended to provide perfect segmentations of tumor boundaries, we performed additional analyses to evaluate whether CAM mismatch correlated with the softmax score. The CAM for each slice was thresholded so that only intensities beyond a certain intensity threshold were considered positive tumor regions.[16] Next, for each image slice, we calculated the Dice similarity coefficient [(2x true positives)/(2x true positives + false positives + false negatives)][17] between positive CAM regions and manual tumor segmentation by a board-certified pediatric neuroradiologist (K.W.Y.). Finally, we correlated the Dice score with model confidence (softmax score) for each slice-level prediction. We found that at a threshold of 0.25, model confidence, in fact, correlated with the Dice score ($r = 0.42$, $P < .001$).

Visualization of Learned Features Using Principal Component Analysis and t-SNE

DMG occupied the most distinct feature space, followed by PA and MB, whereas the EP feature space overlapped with MB. The feature vectors were also analyzed using t-distributed stochastic neighbor embedding (t-SNE), which can show nonlinear relationships and potentially more distinct clustering,[18] and a similar clustering pattern was found for the 4 tumor pathologies (Fig 3*B*).

Comparison of Deep Learning Model versus Radiologist Performance

Four board-certified radiologists read the scans in the held-out test set and generated predictions for each scan. The radiologists detected the presence of tumor with an average sensitivity and specificity of 0.99 and 0.98, respectively (Fig 1 and Table 3), which was not statistically different from the detection accuracy of the model. For tumor subtype classification, the model showed higher sensitivity and specificity for PA, MB, and DMG, but lower sensitivity in predicting EP compared with the radiologists' average (Fig 1). Model classification accuracy and the F_1 score were higher than those of 2 of the 4 radiologists (C and D) and not statistically different from those of the other 2 radiologists (A and B) (Table 3 and On-line Fig 2).

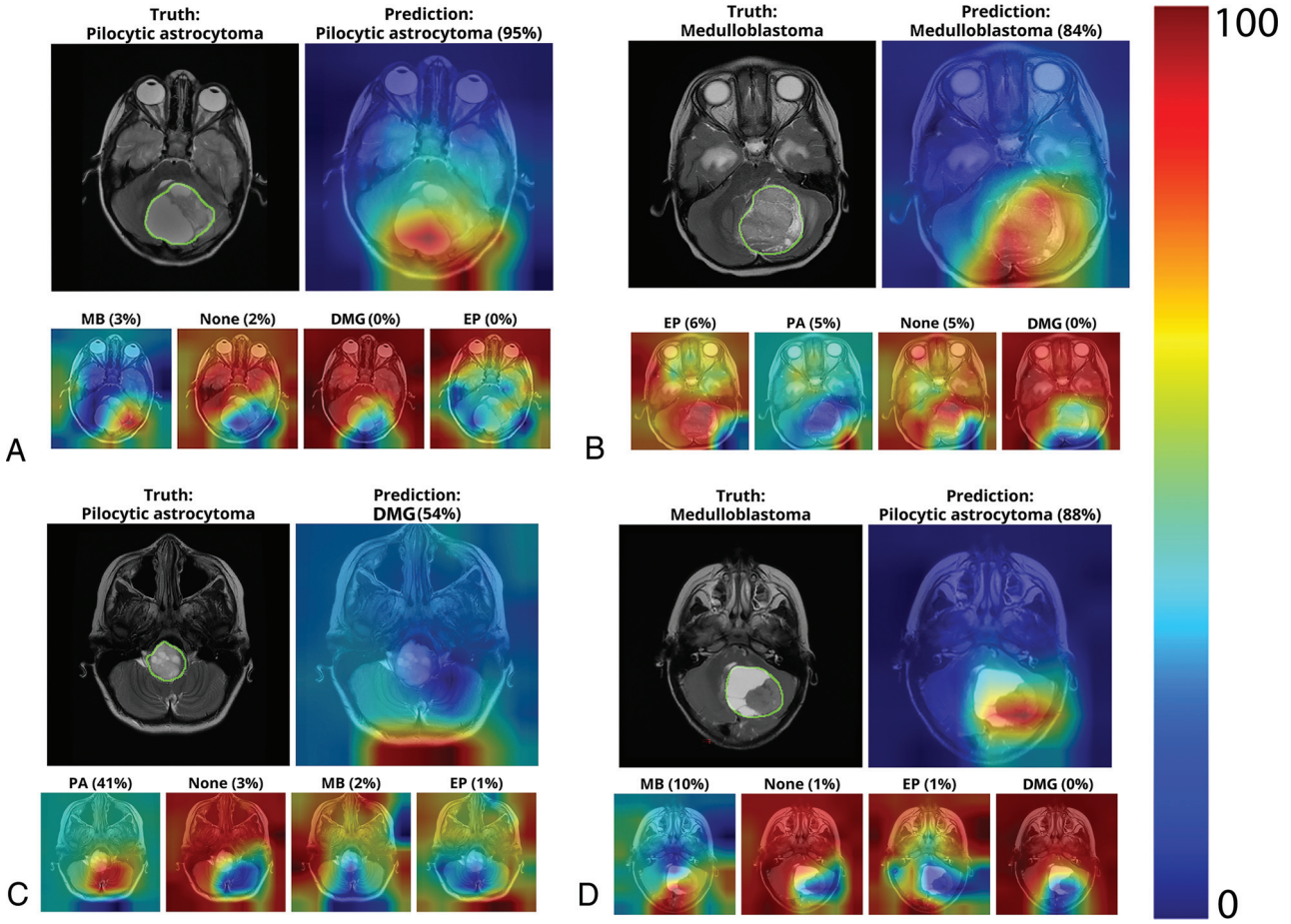

FIG 2. CAMs depicting the areas of the input slice that the model preferentially emphasizes when predicting tumor subtype on individual scan slices. The *upper row* of each subpanel shows the T2 slice with tumor areas manually denoted (*upper left*) and CAM overlay of the most confident prediction of the model (*upper right*). The *lower row* of each panel shows less confident predictions. Examples of correct predictions of PA (*A*) and MB (*B*) and incorrect predictions of PA (*C*) and MB (*D*) are shown.

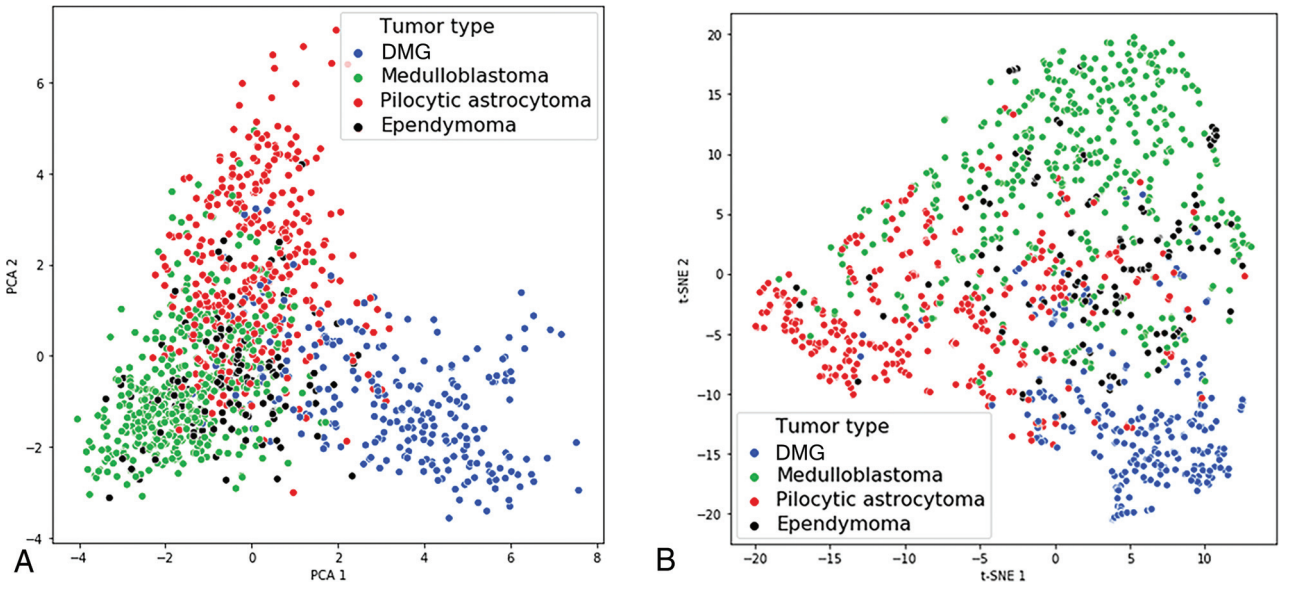

FIG 3. Learned feature vectors were reduced to 2D and visualized using principal component analysis (PCA) (*A*) and t-SNE (*B*). DMG has the most distinctive feature space, followed by PA and MB. EP has the least distinctive feature space and overlaps with MB.

Table 3: Comparison of tumor detection and classification results between the deep learning model and radiologists[a]

	Tumor Detection			Tumor Classification		
	Sensitivity	Specificity	*P*	Accuracy	F_1 Score	*P*
Model	0.96	1.00	–	0.92	0.80	–
Radiologist average	0.99	0.98	–	0.87	0.75	–
Radiologist A	1.00	1.00	.06	0.95	0.89	.09
Radiologist B	0.99	0.97	1.00	0.89	0.79	.24
Radiologist C	0.99	0.98	1.00	0.79	0.61	<.01
Radiologist D	0.98	0.98	.73	0.84	0.70	<.01

Note:— –indicates n/a.

[a] *P* value calculated using the McNemar test comparing the model with individual radiologists.

DISCUSSION

In this study, we present a deep learning model to detect and classify the 4 most common pediatric PF tumor pathologies using T2-weighted MRIs. We modified a state-of-the-art deep learning architecture and trained our model using MRIs from >600 patients with PF tumors at 5 independent pediatric institutions, representing the largest pediatric PF tumor imaging study to date. The model achieved an overall tumor-detection and classification accuracy that was comparable with the performance of 4 board-certified radiologists.

While prior machine learning approaches for PF tumor classification have applied feature engineering or a priori hand-crafted feature extraction, no prior study has used deep learning. Deep learning offers the advantages of automated high-dimensional feature learning through billions of parameters that pass through nonlinear functions within the deep layers of neural networks to tackle complex pattern-recognition tasks.[19,20] Unlike feature-engineering methods such as radiomics that require manual tumor segmentation and hand-crafted computational feature extraction for statistical modeling, data labeling for our deep-learning model was relatively simple: The model required only axial slices from T2-scans with labels of "no tumor" or the specific tumor subtype present on the slice. Notably, the present detection and classification model is not dependent on the precise segmentation of the tumor region of the model. Rather, the model uses the entire slice to make a prediction. Because deep learning models are task-oriented and tailored to the task at hand, the model is essentially free to extract any relevant imaging features to assist with the task. Therefore, we implemented several techniques to better understand the performance of the model. While CAMs do not provide precise tumor segmentations, they can help identify areas of focus. Our finding that the CAM Dice score correlated with the softmax score suggests that when the focus areas of the model had higher overlap with the precise tumor boundary, the model was more confident in the tumor-type prediction.

Additionally, our large, heterogeneous dataset from geographically distinct institutions consisted of scans from multiple vendors and magnet strengths, thus allowing increased generalizability of our model as previous simulation studies have suggested.[21] By evaluating our model on a previously unseen held-out test, our model accuracy is likely to be reflective of real-world accuracy, unlike prior studies with much smaller datasets that used leave-one-out or k-fold cross-validation approaches, which are more prone to overfitting.[22-24]

Prior studies have shown variation in radiologists' interpretations.[25] In this study, we also observed differences among the performance of individual radiologists (Table 3). As the discussion on artificial intelligence in medicine continues to evolve, the radiology community has suggested a potential role for artificial intelligence in augmenting care by bridging knowledge gaps among clinical experts.[26] In this context, we propose that our model could serve to augment the radiologist's performance, particularly among those less experienced in pediatric neuro-oncology.

While our deep learning model exhibited an overall high accuracy for tumor classification, its performance varied with tumor pathology, with the highest accuracy for DMG, followed by PA and MB. Compared with the average performance of human experts, the model more accurately predicted all tumor types except for EP. This outcome might be attributed to the smaller proportion of EP in the training set. It is also possible that learned features for EP overlapped with those of MB, as shown by the principal component analysis and t-SNE plots (Fig 3), which contributed to a more difficult decision boundary for EP and, to a lesser degree, MB. Future studies with even more EP scans could help address these possibilities.

There are several limitations of this study. We restricted model input to T2 scans because our initial experiments showed that training on T2 scans alone outperformed training on a combination of T2, T1-postcontrast, and ADC sequences. We attribute these findings to model overfitting when using all 3 sequences. With the T1-postcontrast/ADC/T2 model, there was a greater difference in performance accuracy between the training and validation sets, indicating that there was more model overfitting. This is likely due to the increased number of input parameters when using all 3 sequences compared to only 1 sequence. In addition, the T2 parameters had greater consistency compared to the T1 parameters (such as image-contrast dynamic range) between institutions: Most used fast spin-echo or turbo spin-echo. T1-postcontrast images, on the other hand, were acquired at a wide variety of parameters and included spin-echo, spoiled gradient recalled echo (SPGR)/Magnetization Prepared - Rapid Gradient Echo (MPRAGE) /Bravo and fluid attenuated inversion recovery (FLAIR). Although we compared the performance using different sequences for the exact same subset of patients, the parameter variation between scans essentially limited the number of T1-postcontrast images within each parameter subtype.

Finally, there was lower scan resolution and greater noise with ADC sequences compared to the anatomic scans (T1- and T2-

sequences). The combination of these three factors likely contributed to our finding that a T2-only model outperformed a T1-postcontrast/ADC/T2 model within our subset of 260 scans. It is possible that with more training data, performance of the T1-postcontrast/ADC/T2 model could improve. Given our dataset and preliminary findings as well as our clinical motivations, we decided to focus our study on optimizing a T2 only model. Thus, our radiologists' performances may have been limited by the restriction to T2-only and may have been improved if they had access to T1-postcontrast and ADC sequences. However, T2 scans are the most universally acquired MR imaging sequences because they are relatively fast, easy to implement, and ubiquitous across the vendors. Our decision to use T2-scans also allowed maximal use of our dataset without incurring the computational cost of sequence coregistration, additional image preprocessing, and potentially larger neural networks that would be required for incorporation of other MR imaging sequences. Nevertheless, our model showed high predictive performance with wide generalizability. Its flexibility in accepting T2-derivative scans across multiple vendors and magnet strengths, with variable slice thicknesses, could also facilitate direct clinical translation.

We also did not evaluate model performance for classifying other pediatric or PF tumors. Because our model was trained on only the 4 most common tumor pathologies, it is not generalizable to other PF tumors, such as choroid plexus tumors or atypical teratoid/rhabdoid tumors. Furthermore, our model was not trained to distinguish between molecular subtypes for each tumor type. Given the growing importance of molecular subtyping for understanding tumor behavior, treatment response, and patient outcomes, we hope to incorporate such information in future iterations of our model.

Finally, our model was not trained to segment precise tumor regions but rather make slice- and scan-level predictions of tumor presence and type. However, tumor segmentation plays a valuable role in monitoring tumor growth and treatment response and is the focus of future work.

CONCLUSIONS

We present a multi-institutional deep learning model for pediatric PF tumor detection and classification with the potential to augment clinical diagnosis. Our work represents applied artificial intelligence in medicine and encourages future research in this domain.

Disclosures: Jennifer Quon—*RELATED: Support for Travel to Meetings for the Study or Other Purposes*: Stanford University, *Comments*: I received institutional reimbursement from the Stanford Neurosurgery Department for travel to the 2019 Pediatric Section Meeting of American Association of Neurological Surgeons to present the preliminary findings of this work. Jayne Seekins—*UNRELATED*: *Consultancy*: Genentech, *Comments*: This is consultancy related to adult malignancies. Matthew P. Lungren—*UNRELATED*: *Consultancy*: Nine-AI, Segmed; *Stock/Stock Options*: Nine-AI, Segmed, Bunker Hill. Tina Y. Poussaint—*UNRELATED*: *Grants/Grants Pending*: Pediatric Brain Tumor Consortium Neuroimaging Center, National Institutes of Health*; *Royalties:* Springer Verlag, book royalties. Hannes Vogel—*UNRELATED*: *Employment*: Stanford University; *Expert Testimony*: miscellaneous; *Grants/Grants Pending*: miscellaneous.* *Money paid to the institution.

REFERENCES

1. Pollack IF, Agnihotri S, Broniscer A. **Childhood brain tumors: current management, biological insights, and future directions.** *J Neurosurg Pediatr* 2019;23:261–73 CrossRef Medline
2. Segal D, Karajannis MA. **Pediatric brain tumors: an update.** *Curr Probl Pediatr Adolesc Health Care* 2016;46:242–50 CrossRef Medline
3. Medina LS, Kuntz KM, Pomeroy S. **Children with headache suspected of having a brain tumor: a cost-effectiveness analysis of diagnostic strategies.** *Pediatrics* 2001;108:255–63 CrossRef Medline
4. Zhou M, Scott J, Chaudhury B, et al. **Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches.** *AJNR Am J Neuroradiol* 2018;39:208–16 CrossRef Medline
5. Northcott PA, Korshunov A, Witt H, et al. **Medulloblastoma comprises four distinct molecular variants.** *J Clin Oncol* 2011;29:1408–14 CrossRef Medline
6. Ramaswamy V, Remke M, Bouffet E, et al. **Risk stratification of childhood medulloblastoma in the molecular era: the current consensus.** *Acta Neuropathol* 2016;131:821–31 CrossRef
7. Nejat F, El Khashab M, Rutka JT. **Initial management of childhood brain tumors: neurosurgical considerations.** *J Child Neurol* 2008;23:1136–48 CrossRef Medline
8. Capper D, Jones DTW, Sill M, et al. **DNA methylation-based classification of central nervous system tumours.** *Nature* 2018;555:469–74 CrossRef Medline
9. Louis DN. *WHO Classification of Tumours of the Central Nervous System*. International Agency for Research on Cancer; 2016
10. Park A, Chute C, Rajpurkar P, et al. **Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model.** *JAMA Netw Open* 2019;2:e195600 CrossRef Medline
11. El-Dahshan ES, Mohsen HM, Revett K, et al. **Computer-aided diagnosis of human brain tumor through MRI: a survey and a new algorithm.** *Expert Systems with Applications* 2014;41:5526–45 CrossRef
12. Deng J, Dong W, Socher R, et al. **ImageNet: a large-scale hierarchical image database.** In: *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, Florida. June 20–25, 2009:248–55
13. Krizhevsky A, Sutskever I, Hinton GE. **Imagenet classification with deep convolutional neural networks.** In: *Proccedings of the Advances in Neural Information Processing Systems Conference*, Lake Tahoe, California. December 3–8, 2012;1097–1105
14. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. **Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI**. *Information Fusion* 2019;58:82–115 CrossRef
15. Simonyan K, Vedaldi A, Zisserman A. **Deep inside convolutional networks: visualising image classification models and saliency maps.** *arXiv* December 2013. https://arxiv.org/pdf/1312.6034.pdf. Accessed April 5, 2020
16. Zhou B, Khosla A, Lapedriza A, et al. **Learning deep features for discriminative localization.** *In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada. June 27–30, 2016:2921–29
17. Dice LR. **Measures of the amount of ecologic association between species.** *Ecology* 1945;26:297–302 CrossRef
18. van der Maaten L, Hinton G. **Visualizing data using t-SNE.** *Journal of Machine Learning Research* 2008;9:2579–2605
19. Hosny A, Parmar C, Quackenbush J, et al. **Artificial intelligence in radiology.** *Nat Rev Cancer* 2018;18:500–10 CrossRef Medline
20. Savadjiev P, Chong J, Dohan A, et al. **Demystification of AI-driven medical image interpretation: past, present and future.** *Eur Radiol* 2019;29:1616–24 CrossRef Medline
21. Chang K, Balachandar N, Lam C, et al. **Distributed deep learning networks among institutions for medical imaging.** *J Am Med Inform Assoc* 2018;25:945–54 CrossRef Medline
22. Rodriguez Gutierrez D, Awwad A, Meijer L, et al. **Metrics and textural features of MRI diffusion to improve classification of pediatric posterior fossa tumors.** *AJNR Am J Neuroradiol* 2014;35:1009–15 CrossRef Medline
23. Arle JE, Morriss C, Wang ZJ, et al. **Prediction of posterior fossa tumor type in children by means of magnetic resonance image**

properties, spectroscopy, and neural networks. *J Neurosurg* 1997;86:755–61 CrossRef Medline

24. Bidiwala S, Pittman T. **Neural network classification of pediatric posterior fossa tumors using clinical and imaging data.** *Pediatr Neurosurg* 2004;40:8–15 CrossRef Medline
25. Abujudeh HH, Boland GW, Kaewlai R, et al. **Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists.** *Eur Radiol* 2010;20:1952–57 CrossRef Medline
26. Allen B, Jr., Seltzer SE, Langlotz CP, et al. **A Road Map for Translational Research on Artificial Intelligence in Medical Imaging: From the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop.** *J Am Coll Radiology* 2019;16:1179–89 CrossRef Medline