Radiology: Artificial Intelligence

TECHNICAL DEVELOPMENT

MRI-based Identification and Classification of Major Intracranial Tumor Types by Using a 3D Convolutional Neural Network: A Retrospective Multi-institutional Analysis

Satrajit Chakrabarty, MS • Aristeidis Sotiras, PhD • Mikhail Milchenko, PhD • Pamela LaMontagne, PhD • Michael Hileman, BS • Daniel Marcus, PhD

From the Department of Electrical and Systems Engineering, Washington University in St Louis, 1 Brookings Dr, St Louis, MO 63130 (S.C.); Department of Radiology and Institute for Informatics, Washington University School of Medicine, St Louis, Mo (A.S.); and Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (M.M., P.L., M.H., D.M.). Received December 20, 2020; revision requested February 5, 2021; revision received June 23; accepted July 14. **Address correspondence to** S.C. (e-mail: *satrajit.chakrabarty@wustl.edu*).

D.M. is supported by the National Institutes of Health (grants P30 NS098577, U24 CA204854, and R01 EB009352).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2021; 3(5):e200301 • https://doi.org/10.1148/ryai.2021200301 • Content codes: AI NR

Purpose: To develop an algorithm to classify postcontrast T1-weighted MRI scans by tumor classes (high-grade glioma, low-grade glioma [LGG], brain metastasis, meningioma, pituitary adenoma, and acoustic neuroma) and a healthy tissue (HLTH) class.

Materials and Methods: In this retrospective study, preoperative postcontrast T1-weighted MR scans from four publicly available datasets—the Brain Tumor Image Segmentation dataset (n = 378), the LGG-1p19q dataset (n = 145), The Cancer Genome Atlas Glioblastoma Multiforme dataset (n = 141), and The Cancer Genome Atlas Low Grade Glioma dataset (n = 68)—and an internal clinical dataset (n = 1373) were used. In all, a total of 2105 images were split into a training dataset (n = 1396), an internal test set (n = 361), and an external test dataset (n = 348). A convolutional neural network was trained to classify the tumor type and to discriminate between images depicting HLTH and images depicting tumors. The performance of the model was evaluated by using cross-validation, internal testing, and external testing. Feature maps were plotted to visualize network attention. The accuracy, positive predictive value (PPV), negative predictive value, sensitivity, specificity, F1 score, area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUPRC) were calculated.

Results: On the internal test dataset, across the seven different classes, the sensitivities, PPVs, AUCs, and AUPRCs ranged from 87% to 100%, 85% to 100%, 0.98 to 1.00, and 0.91 to 1.00, respectively. On the external data, they ranged from 91% to 97%, 73% to 99%, 0.97 to 0.98, and 0.9 to 1.0, respectively.

Conclusion: The developed model was capable of classifying postcontrast T1-weighted MRI scans of different intracranial tumor types and discriminating images depicting pathologic conditions from images depicting HLTH.

Supplemental material is available for this article.

More than 150 types of brain tumor have been documented on the basis of histopathologic characteristics (1). Although histopathologic assessment is the diagnostic standard for brain tumor classification, it requires an invasive surgical procedure and is complicated by intratumoral spatial heterogeneity (2). MRI may be used as a complement or, in some cases, as an alternative to histopathologic examination because of its noninvasive nature and high soft-tissue contrast (3). Machine and deep learning approaches can potentially automate the detection and classification of brain tumors with MRI.

Several studies have investigated machine and deep learning techniques for brain tumor type classification (4). Most recent work using deep learning methods (4) has been facilitated by the availability of open-access labeled tumor MRI data published by Cheng et al (3). This dataset comprises manually curated two-dimensional (2D) sections depicting an easily discernible tumor area. However, use of this dataset limits the development of algorithms in two ways. First, these methods depend on manually detecting a similar section for each unseen image before classification. Second, 2D sections for location-specific tumor types, such as pituitary adenoma (PA), depict similar brain anatomy, which is distinct from that of other tumor types and can confound algorithms, leading to overfitting problems. Furthermore, several methods require additional manual input, such as tumor location or zoomed tumor areas (5), tumor segmentations (6), or bounding boxes (7). Most of the recent studies (4) have focused on classifying PA, glioma, and meningioma (MEN). However, many studies have not included other common tumor types, such as brain metastasis (METS) (8) and acoustic neuroma (AN). Last, few works have attempted to differentiate images depicting healthy tissue (HLTH) from images depicting brain tumors (8,9).

To address these limitations, we adopted a three-dimensional (3D) convolutional neural network (CNN) architecture for classifying MR images into a HLTH class and

Abbreviations

AN = acoustic neuroma, AUC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve, BraTS = Brain Tumor Image Segmentation, CNN = convolutional neural network, HGG = high-grade glioma, HLTH = healthy tissue, LGG = low-grade glioma, MEN = meningioma, METS = brain metastasis, NPV = negative predictive value, PA = pituitary adenoma, PPV = positive predictive value, 3D = three dimensional, 2D = two dimensional, WUSM = Washington University School of Medicine

Summary

A convolutional neural network model was developed to classify postcontrast T1-weighted MRI scans into a healthy tissue class and six tumor classes: high- and low-grade glioma, brain metastasis, meningioma, pituitary adenoma, and acoustic neuroma.

Key Points

- Without the assistance of any manual tumor segmentations or bounding boxes, the convolutional neural network model, developed with a large heterogeneous multi-institutional dataset (n = 2105) acquired from four different sources, could classify six brain tumor types and discriminate images depicting healthy tissue from images depicting pathologic conditions by using a single postcontrast T1-weighted volume from each patient.
- On an internal testing dataset comprising all seven image classes, the model achieved areas under the receiver operating characteristic curve (AUCs) in the range of 0.98 to 1.00 and areas under the precision-recall curve (AUPRCs) of 0.91 to 1.00.
- On an external test set comprising high-grade glioma and low-grade glioma classes, the model achieved AUCs and AUPRCs ranging from 0.97 to 0.98 and 0.9 to 1.0, respectively, demonstrating good generalization on external data.

Keywords

MR-Imaging, CNS, Brain/Brain Stem, Diagnosis/Classification/Application Domain, Supervised Learning, Convolutional Neural Network, Deep Learning Algorithms, Machine Learning Algorithms

six tumor classes—high-grade glioma (HGG), low-grade glioma (LGG), METS, MEN, PA, and AN—by using only a single 3D postcontrast T1-weighted MRI volume for each patient and without the requirement of any additional manual interaction. By using cross-validation and testing on a large heterogeneous multisite dataset, we demonstrated that the model was accurate across different sites.

Materials and Methods

Retrospective de-identified data were obtained from Washington University School of Medicine (WUSM), with a waiver of consent being provided, in accordance with the Health Insurance Portability and Accountability Act, as approved by the institutional review board. Additional data were obtained from public datasets after completion of necessary data usage agreements.

Datasets

Dataset overview.— In the development and testing of the model, images depicting a total of six brain tumor types (HGG, LGG, METS, MEN, AN, and PA) and images depicting HLTH were derived from either an internal dataset (WUSM) or five different public datasets (the Brain Tumor Image Segmentation [BraTS] 2018 and 2019 datasets, the LGG 1p19q dataset [from The Cancer Imaging Archive], The Cancer Genome Atlas Glioblastoma Multiforme dataset, and The Cancer Genome Atlas Low Grade Glioma dataset). Two main datasets were developed (Fig 1C): *(a)* an internal dataset for training, cross-validation, and testing (n = 1757) and *(b)* an external test dataset consisting of HGG and LGG scans (n = 348). For each individual, a single preoperative postcontrast T1-weighted image (Fig 1A) was used (see Fig E1 [supplement] for the location information for each tumor and Appendix E1 [supplement] for details on preprocessing [10]). The resulting dataset was heterogeneous, with a high variability in acquisition protocol (Tables E1–E4; glioma histologic findings are shown in Table E5 [supplement]).

Internal training and test dataset.— MR images of HGG and LGG were obtained from the 2019 BraTS (11–13) training dataset (n = 259 HGG, n = 76 LGG), the 2018 BraTS test dataset (13–15) (n = 43 LGG), and the LGG-1p19q database (16,17) (n = 159). The BraTS datasets comprise routine, clinically acquired preoperative scans associated with histopathologically confirmed diagnoses from 19 institutions (12). The LGG-1p19q dataset comprises consecutive (October 1, 2002, to August 01, 2011) preoperative LGG scans with stereotactic MRI from the Mayo Clinic. Fourteen patients were excluded from the LGG-1p19q database because of misregistration.

Postcontrast T1-weighted images of METS (n = 710), MEN (n = 143), AN (n = 158), PA (n = 82), and HLTH (n = 141) were obtained from patients at WUSM between February 2001 and October 2019. All images except those of HLTH were from patients undergoing gamma knife protocol and were obtained from a retrospective data repository maintained by the radiation oncology department. The tumor types were confirmed through radiology reports provided by board-certified radiologists.

The HLTH group (n = 141) included individuals who initially underwent the clinical tumor MRI protocol because of tumor suspicion but were later confirmed to be healthy (ie, negative for the presence of tumor as well as central nervous system diseases or any other kind of abnormality) by a neuroradiologist. The patient list was compiled by using MONTAGE (Nuance mPower, version 3.2.4) to search the radiologic database for "normal brain MRI" and filtering the scans by using postcontrast T1-weighted scan. The radiology reports of these individuals were subsequently manually reviewed to verify the absence of tumor.

In all, a total of 1757 patient scans were included in the internal dataset; these were split into 1396 within the training set and 361 within the internal test dataset.

External test dataset of LGG and HGG.— In addition, an external test dataset was curated, which consisted of histopathologically confirmed HGG and LGG on images from The Cancer Genome Atlas Glioblastoma Multiforme (n = 262 HGG), The Cancer Genome Atlas Low Grade Glioma (n =

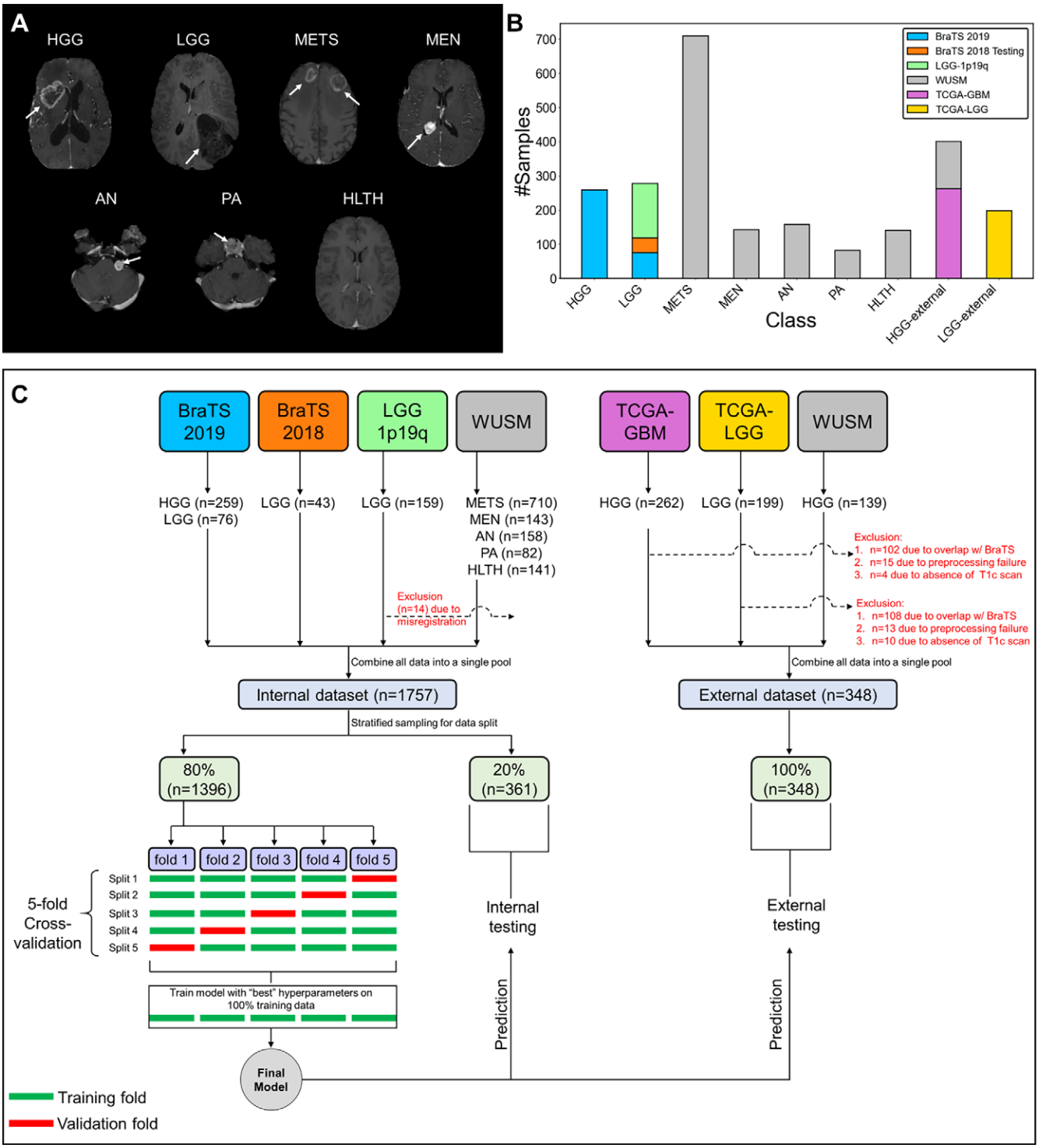

Figure 1: **(A)** Sample postcontrast T1-weighted images (axial section, right, anterior, and superior orientation) for the high-grade glioma (HGG), low-grade glioma (LGG), brain metastases (METS), acoustic neuroma (AN), pituitary adenoma (PA), meningioma (MEN), and healthy tissue (HLTH) classes included in the study (white arrows). **(B)** Class-wise distribution of data and **(C)** flow of images and data split for cross-validation, internal, and external testing. BraTS = Brain Tumor Image Segmentation, T1c = postcontrast T1-weighted image, TCGA = The Cancer Genome Atlas, TCGA-GBM = TCGA Glioblastoma Multiforme, WUSM = Washington University School of Medicine.

199 LGG), and the WUSM (n = 139 HGG) datasets. After exclusion of patient scans due to overlap with the BraTS datasets (n = 210), misregistration (n = 28), and the unavailability of postcontrast T1-weighted sequence (n = 14), 348 scans were included in the external test set (Fig 1C).

Neural Network Architecture

The developed network was inspired by successful encoder–decoder brain tumor segmentation architectures (18) (Appendix E2 [supplement]). It incorporates the entire 3D volume and predicts a class assignment only on the basis of the context-encoding path

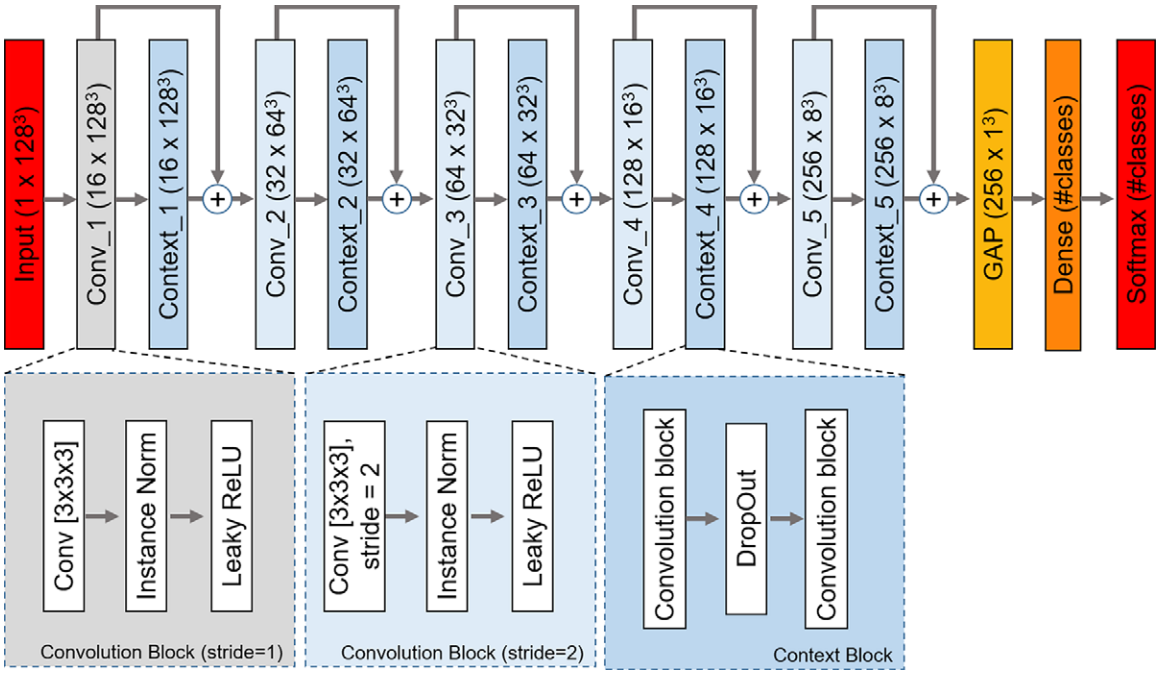

Figure 2: The proposed network architecture. Each rectangle contains the name of the block or layer with the dimension of the feature map (*n* channels × height × width × depth) in parentheses. Context_i = *i*th context block, Conv = convolutional layer, Conv_i = *i*th convolutional block, GAP = global average pooling layer, Instance Norm = instance normalization layer, ReLU = rectified linear unit.

(Fig 2). This path was connected to a global average pooling layer that was followed by a dropout layer and a dense softmax layer with seven output nodes producing the class probabilities (further details on hyperparameter optimization and the network training strategy are detailed in Appendixes E2–E3 [supplement]).

Cross-Validation

The performance of the model was evaluated by using cross-validation, internal testing, and external testing. Stratified sampling was used to split the internal dataset into training and testing sets (Fig 1C) to maintain the original ratio of different image classes in every split (Table 1). For hyperparameter tuning, we performed a random search (19) by using fivefold cross-validation on the training data and by using 80% of the data for training (n = 1116) and 20% of the data for validation (n = 280). The hyperparameters that yielded the best cross-validation results were then selected as the "best" hyperparameters, and the corresponding cross-validation results are reported. Next, the model was trained on 100% of the training data (n = 1396) by using the best set of hyperparameters, and these data were subsequently used for prediction on the internal and external test datasets.

Statistical Analyses

We used χ^2 and Mann-Whitney tests to evaluate differences in patient demographics between data splits. Model performance was quantified by using the accuracy, positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, F1 score, receiver operating characteristics, and precision-recall curves. Bootstrap 95% CIs were calculated when appropriate (Appendix E4 [supplement]). We used the gradient-weighted class activation mapping method (20,21) (Appendix E5 [supplement]) to visualize network attention. The model performance was compared with a state-of-the-art 2D ResNet50 baseline model (22) (Appendix E6 [supplement]) on the internal test data. Statistical comparisons were performed by using the McNemar test (23); the generalized score statistic proposed by Leisenring et al (24); and the DeLong test (25) (Appendix E4 [supplement]). Additionally, we analyzed potential overfitting of the model to location-specific tumors (Appendix E7 [supplement]), as well as model performance under section thickness variability.

Code Availability

Code for this work has been made publicly available at *https://github.com/satrajitgithub/tumor_classifier_3d.git*.

Results

Dataset Characteristics

The patient demographics and class distribution were calculated (Table 1) for the entire pool of individuals together as well as separately for the training, internal testing, and external test sets. The training and internal testing data had no significant difference with regard to age or sex. The internal and external test data had no significant difference with regard to age but differed significantly in terms of sex.

Quantitative Analyses on the Internal Test Dataset

The performance of the model was first evaluated on the internal testing data (Table 2, Fig 3A). The model achieved an accuracy of 93.35% (337 of 361) across seven classes. The

Table 1: Summary of Patient Demographics and Class Distribution for the Entire Dataset

Parameter	Total	Internal Data		External Data Testing
		Training	Testing	
Patient characteristics				
No. of patients	2105 (100.0)	1396 (66.32)	361 (17.15)	348 (16.53)
Median age (y)*	57 (47–65)	56 (47–65)	57 (46–66)†	58 (46–66)‡
Sex§‖				
Male	1045	719	188	138
Female	876	529	141	206
Unknown	184	148	32	4
Tumor category				
HGG	539 (26)	205 (15)	54 (15)	280 (80)
LGG	332 (16)	210 (15)	54 (15)	68 (20)
METS	710 (34)	567 (41)	143 (40)	NA
MEN	143 (7)	113 (8)	30 (8)	NA
AN	158 (8)	125 (9)	33 (9)	NA
PA	82 (4)	64 (5)	18 (5)	NA
HLTH	141 (7)	112 (8)	29 (8)	NA

Note.—Data are shown as counts with percentages in parentheses or as medians with interquartile ranges in parentheses. AN = acoustic neuroma, BraTS = Brain Tumor Image Segmentation, HGG = high-grade glioma, HLTH = healthy tissue, LGG = low-grade glioma, MEN = meningioma, METS = brain metastases, NA = not applicable, PA = pituitary adenoma, TCGA = The Cancer Genome Atlas, TCGA-GBM = TCGA Glioblastoma Multiforme, WUSM = Washington University School of Medicine.

* For the BraTS 2019 HGG data, age information was determined from the survival data provided with the dataset. For the BraTS 2019 LGG data, age information was determined from the Genomic Data Commons research data of the National Cancer Institute for patients who were also included in The Cancer Imaging Archive dataset. Of 2105 patients, age information could not be determined for 19 patients from the BraTS 2019 HGG data, one patient from the TCGA-GBM data, 12 patients from the BraTS 2019 LGG data, one patient from the BraTS 2018 LGG data, five patients from the WUSM AN data, 11 patients from the WUSM METS data, two patients from the WUSM PA data, and 56 patients from the WUSM HGG data.

† $P = .21$ for comparison between training and internal testing data for age.

‡ $P = .46$ for comparison between external test dataset and internal data (training and testing data combined) for age.

§For the BraTS 2019 HGG and LGG data, sex information was determined from the Genomic Data Commons research data of the National Cancer Institute. Of 2105 patients, sex information could not be determined for 158 patients from the BraTS 2019 HGG data, one patient from the TCGA-GBM data, three patients from the WUSM HGG data, one patient from the BraTS 2018 LGG data, 14 patients from the BraTS 2019 LGG data, two patients from the WUSM AN data, and five patients from the WUSM HLTH data.

‖$P = .92$ for comparison between training and internal testing data for sex and P value of less than .001 for comparison between external test dataset and internal data (training and testing data combined) for sex.

model had NPVs in the range of 98% to 100% across all classes. Among the well-represented classes, METS had an NPV of 99%. A similar trend was observed for specificity (range, 97%–100% across all classes). Sensitivities ranged from 91% to 100%, and PPVs ranged from 85% to 100%. Most errors were due to LGG being misclassified as HGG (13%, seven of 54) and vice versa (13%, seven of 54); HLTH being misclassified as METS (6.9%, two of 29); and LGG being misclassified as HLTH (3.7%, two of 54) and vice versa (3.4%, one of 29; Fig 3A).

The receiver operating characteristic and precision-recall curves (Fig 3A) indicate good classification performance for all classes. The model achieved areas under the receiver operating characteristic curve (AUCs) in the range of 0.98 to 1.00 and areas under the precision-recall curve (AUPRCs) in the range of 0.92 to 1.00 (Table 2). We visualized the network attention for six correctly predicted images (Fig 4) belonging to each of the tumor classes. Network attention overlapped with the tumor areas for all tumor types.

Quantitative Analyses on the External Test Dataset

For the external test dataset (Table 2, Fig 3B), the model had an accuracy of 91.95% (320 of 348) across seven classes. The model achieved similar sensitivity (91%, 254 of 280) but achieved a much higher PPV (99%, 254 of 256) for HGG compared with the internal test dataset. For LGG, the sensi-

tivity was 97% (66 of 68), but the PPV dropped (73%, 66 of 90) compared with the internal test set because of 24 HGGs being misclassified as LGGs. The AUC values for HGG (0.97 [95% CI: 0.96, 0.99]) and LGG (0.98 [95% CI: 0.96, 0.99]) decreased by 2% compared with those for the internal test dataset. The AUPRC value for HGG was 0.99 (95% CI: 0.99, 1.00), and the AUPRC for LGG was 0.90 (95% CI: 0.83, 0.96). The high accuracy in the internal and external testing datasets demonstrates good model generalization capability.

The above quantitative analyses were also performed for the cross-validation results (Appendix E8, Fig E2, Table E6 [supplement]).

Comparison of CNN to 2D ResNet 50

Overall, the proposed 3D model had a performance gain compared with the 2D ResNet50 baseline model (Fig 3C, Tables E7, E8 [supplement]) in terms of all performance metrics. Specifically, for HGG (gains of 2.48% in the NPV [P = .0313], 14.81% in the sensitivity [P = .0325], and 2.05% in the AUC [P = .0307]), LGG (gain of 5.22% in the AUC [P = .0095]), METS (gains of 13.16% in the PPV [$P < .001$], 4.46% in the NPV [P = .0086], 6.29% in the sensitivity [P = .0125], 9.63% in the specificity [$P < .001$], and 2.95% in the AUC [$P < .001$]) and MEN (gains of 31.16% in the PPV [P = .002], 3.27% in the NPV [P = .0018], 36.67% in the sensitivity [P = .0023], and 2.42% in the specificity [P = .0114]), the performance gain was determined to be statistically significant (all statistical comparisons are shown in Table E8 [supplement]).

Performance of CNN under Varying Section Thicknesses and Location-specific Tumors

The model performed well for section thicknesses of 1–2.5 mm, whereas for section thicknesses of 3 mm and 5 mm, the performance was slightly lower, possibly because of the misclassification of LGG cases (two of four for thickness = 3 mm, three of eight for thickness = 5 mm) (Appendix E9, Fig E3 [supplement]). Additionally, the model performance was not affected by potential overfitting caused by location-specific tumors like ANs and MENs (Appendix E9, Fig E5 [supplement]).

Analysis of Model Misclassifications and Model Probability Scores

On analyzing the top-two predicted class data for the misclassified testing cases, we observed that the model often predicted the correct class with the second highest probability (Fig 5A, Appendix E10 [supplement]). In general, for all classes, the model made correct predictions with high probability scores (ie, the probability of the correct class averaged 0.97 ± 0.06 across seven classes; Fig 5B). On the contrary, for the misclassifications, the probability score of the model was lower (ie, the probability of predicted erroneous class averaged 0.81 ± 0.17; Fig 5C, Appendix E11 [supplement]).

Discussion

In this work, we developed a deep learning architecture for classification of six intracranial tumor types and a healthy type and validated its performance. The model achieved high accu-

Table 2: Quantitative Results for Internal and External Test Datasets

	Internal Test (n = 361)							External Test (n = 348)	
Metric	HGG (n = 54)	LGG (n = 54)	METS (n = 143)	PA (n = 18)	AN (n = 33)	HLTH (n = 29)	MEN (n = 30)	HGG (n = 280)	LGG (n = 68)
PPV	87 (47/54)	85 (44/52)	97 (141/145)	95 (18/19)	100 (33/33)	90 (26/29)	97 (28/29)	99 (254/256)	73 (66/90)
NPV	98 (300/307)	97 (299/309)	99 (214/216)	100 (342/342)	100 (328/328)	99 (329/332)	99 (330/332)	72 (66/92)	99 (256/258)
Sensitivity	87 (47/54)	81 (44/54)	99 (141/143)	100 (18/18)	100 (33/33)	90 (26/29)	93 (28/30)	91 (254/280)	97 (66/68)
Specificity	98 (300/307)	97 (299/307)	98 (214/218)	100 (342/343)	100 (328/328)	99 (329/332)	100 (330/331)	97 (66/68)	91 (256/280)
F1 score	0.87	0.83	0.98	0.97	1.00	0.90	0.95	0.95	0.84
AUC	0.99 (0.98,0.99)	0.98 (0.97,0.99)	1.00 (1.00,1.00)	1.00 (1.00,1.00)	1.00 (1.00,1.00)	1.00 (0.99,1.00)	1.00 (1.00, 1.00)	0.97 (0.96, 0.99)	0.98 (0.96, 0.99)
AUPRC	0.92 (0.84,0.98)	0.92 (0.87,0.96)	0.99 (0.99,1.00)	1.00 (1.00,1.00)	1.00 (1.00,1.00)	0.97 (0.92,0.99)	1.00 (1.00,1.00)	0.99 (0.99,1.00)	0.90 (0.83, 0.96)

Note.—For the PPV, NPV, sensitivity, and specificity metrics, data are shown as percentages, with numbers in parentheses providing the numerators and denominators used to calculate the percentages. For the AUC and AUPRC data, numbers in parentheses are 95% CIs. AN = acoustic neuroma, AUC = area under the receiver operating characteristic curve, AUPRC = area under precision-recall curve, HGG = high-grade glioma, HLTH = healthy tissue, LGG = low-grade glioma, MEN = meningioma, METS = brain metastases, NPV = negative predictive value, PA = pituitary adenoma, PPV = positive predictive value.

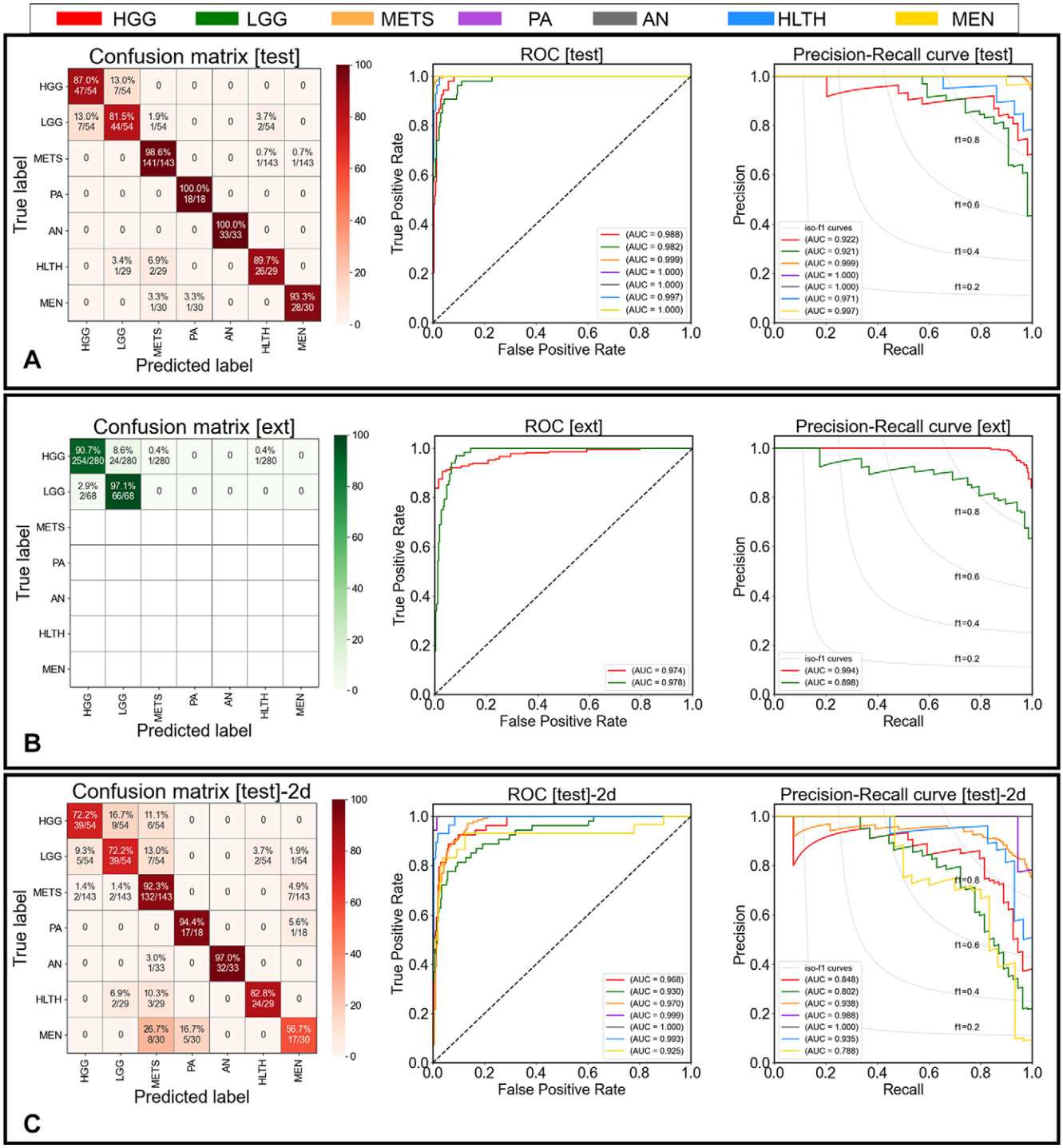

Figure 3: Confusion matrix, receiver operating characteristic (ROC) curve, and precision-recall curve for **(A)** the testing performance demonstrated by the proposed model, **(B)** the external testing performance (ext) demonstrated by the proposed model, and **(C)** the testing performance by the two-dimensional (2d) ResNet50 baseline model. In confusion matrices, diagonal elements show the sensitivity per class, and the off-diagonal elements show the error distribution among different classes. AN = acoustic neuroma, AUC = area under the curve, HGG = high-grade glioma, HLTH = healthy tissue, LGG = low-grade glioma, MEN = meningioma, METS = brain metastases, PA = pituitary adenoma.

racy on a heterogeneous dataset and showed excellent generalization capabilities on unseen testing data. These results suggest that deep learning is a promising approach for automated classification and evaluation of brain tumors.

To the best of our knowledge, this is the first study to address the most common intracranial tumor types while directly determining the tumor class as well as detecting the absence of tumor from a 3D MRI volume. Moreover, the proposed 3D method improves on the existing 2D approaches in two ways. First, the capability of classifying an entire MRI volume obviates the requirement of prior section selection or tumor segmentation. This makes the clinical translatability of this model

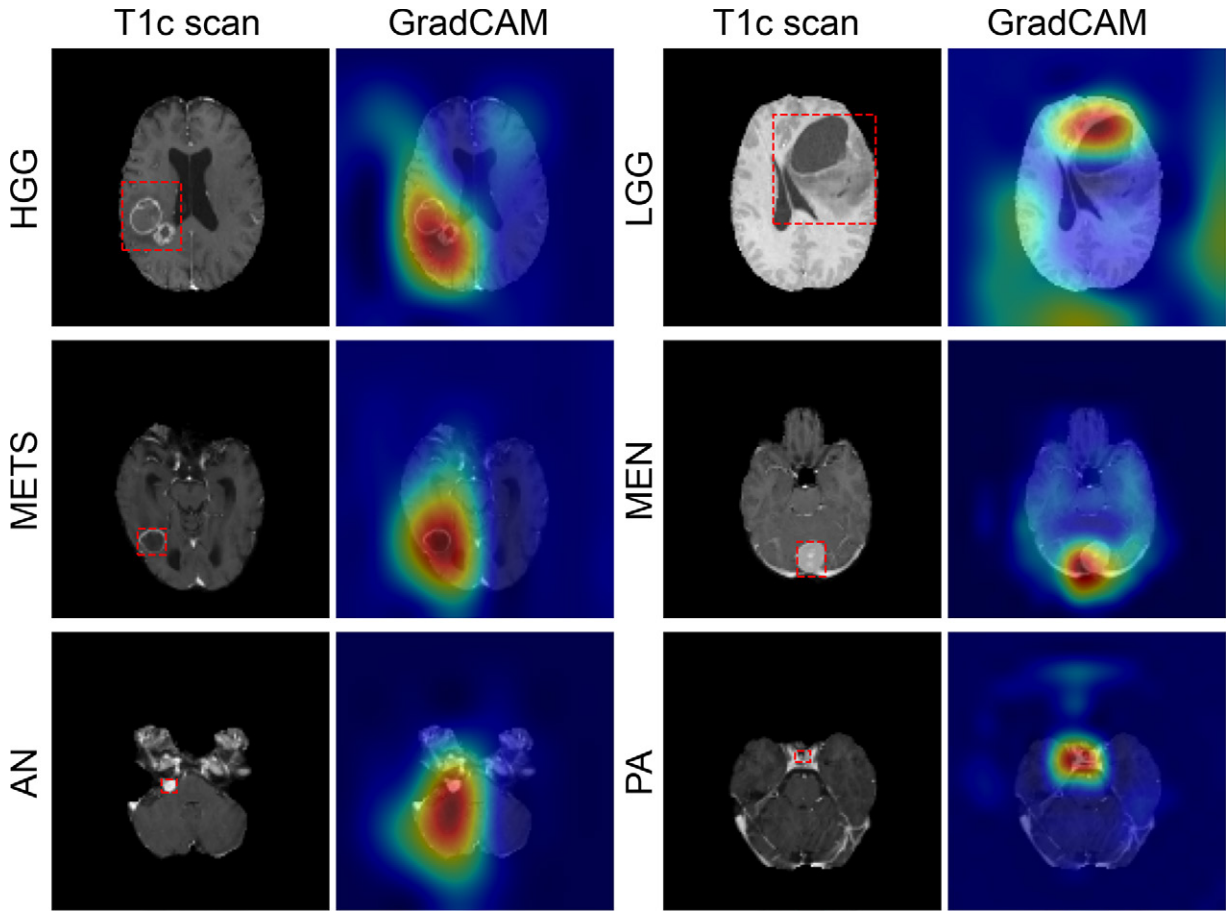

Figure 4: Coarse attention maps generated by using gradient-weighted class activation mapping (GradCAM) for correctly classified high-grade glioma (HGG), low-grade glioma (LGG), brain metastases (METS), meningioma (MEN), acoustic neuroma (AN), and pituitary adenoma (PA). For each pair, the postcontrast T1-weighted image (T1c), and the GradCAM attention map (overlaid on image) have been shown. In GradCAM maps, warmer and colder colors represent high and low contributions of pixels toward a correct prediction, respectively. Dashed red lines = tumor area.

straightforward and facilitates integration in informatics pipelines. Second, using an entire volume prevents overfitting issues caused by location-specific tumors, such as PAs or ANs.

Of the seven classes addressed in this work, more errors were observed for HGG, LGG, and HLTH. The reason for the misclassification between LGG and HLTH could be due to less contrast enhancement of LGGs on postcontrast T1-weighted images as a result of less disruption of the blood–brain barrier (12). Classification of LGG images is further challenged by the inclusion of two different grades of glioma (World Health Organization grades II and III) in a single class, which results in high variability in the contrast enhancement signal. Moreover, in the analysis of 22 LGG testing cases that were misclassified as HGG, we found that 21 were from the BraTS dataset, whereas only one case was from the LGG-1p19q dataset. This can be attributed to the differences between the BraTS dataset (multi-institution, variable imaging protocol) and the LGG-1p19q dataset (single institution, consistent imaging protocol [16]). To conclude, regarding the HGG and LGG misclassifications, the model could classify most of the cases as glioma but occasionally failed in determining their grades.

Our results indicate that a prediction probability score less than 0.9 and a difference between the top two most probable classes below 0.9 are strong indicators of erroneous predictions. These prediction score cutoffs allow for automated flagging of images as suspicious or discordant, which can be resolved by sending them to a radiologist for false-positive interpretation.

The proposed classification model can be cascaded with segmentation models, together with downstream radiomic feature extraction and quantitative radiology report generation tools, tailored for specific tumor types. This can enable artificial intelligence–augmented tools for automated tumor characterization and segmentation that can streamline clinical workflows and support clinical decision-making. In the medical imaging literature, there has been a substantial amount of work for the two independent tasks of classification and segmentation. However, because of the high sensitivity of deep learning models to the distribution of data, segmentation models are specific to the tumor type. For example, infiltrative tumors such as gliomas are segmented into three different classes (enhancing, nonenhancing, and edema) because of their intratumoral heterogeneity, whereas other noninfiltrative tumors are segmented into a single tumor class. Therefore, to render complete automation to an end-to-end neuro-oncologic workflow, it is imperative to build a classification model that can first determine the tumor type of an incoming scan and subsequently trigger the execution of tumor class–specific segmentation models.

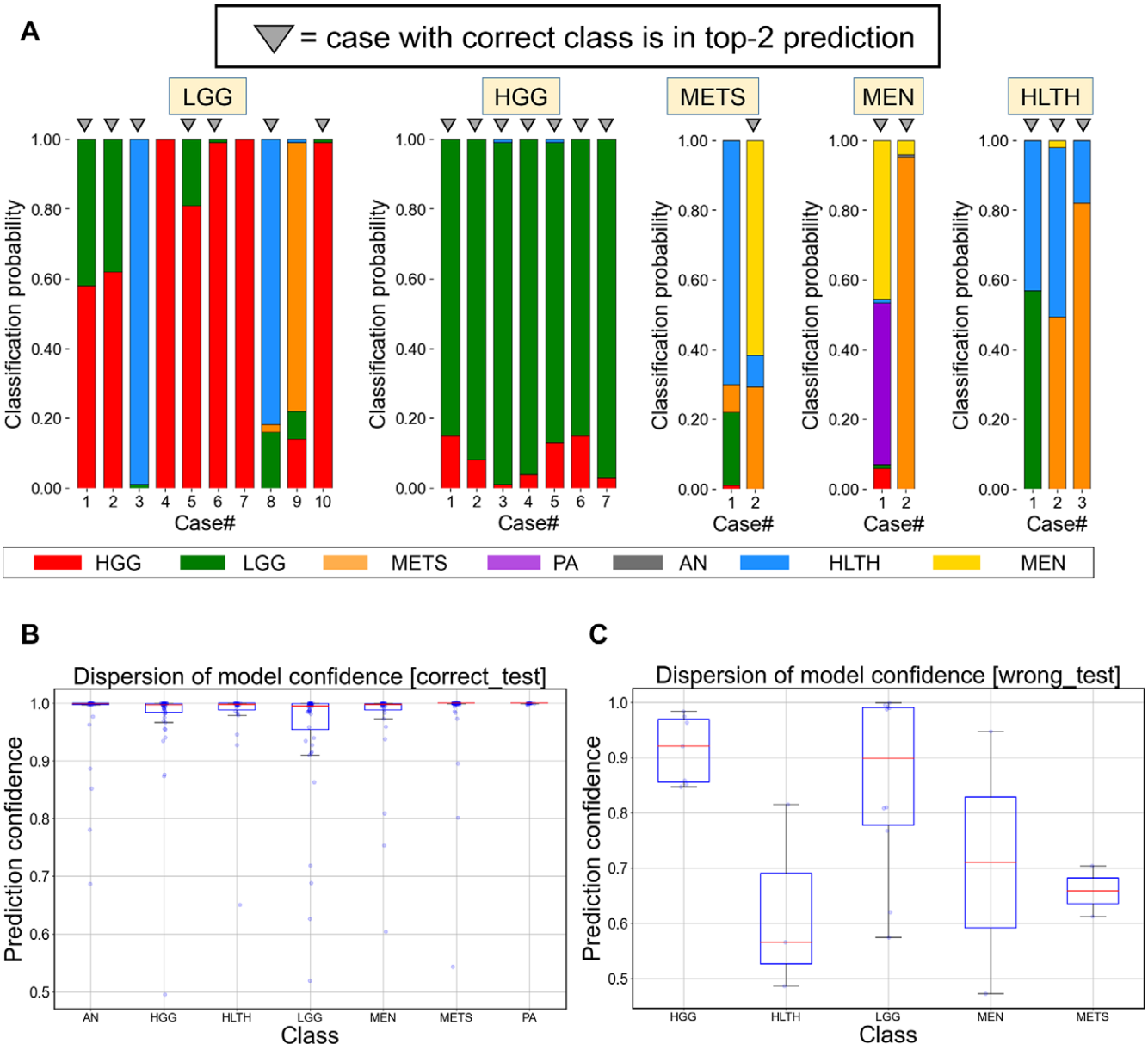

Figure 5: Graphs show **(A)** predicted class probabilities (y-axis) as stacked bar plots for all misclassified testing images (n = 24) (x-axis), **(B)** class-wise dispersion of the prediction probability score for correctly classified testing images, and **(C)** class-wise dispersion of prediction probability score for incorrectly classified testing images. AN = acoustic neuroma, HLTH = healthy tissue, HGG = high-grade glioma, LGG = low-grade glioma, MEN = meningioma, METS = brain metastases, PA = pituitary adenoma.

The main limitation of this study was the use of a single imaging modality. Despite the availability of precontrast T1-weighted, T2-weighted, and T2-weighted fluid-attenuated inversion recovery scans for most of the HGG and LGG cases, these could not be used because postcontrast T1-weighted scan constituted the only scan type present for all individuals across the entire dataset. Multiple modalities offer complementary information about the growth potential and aggressiveness of gliomas, which can be vital for determining their grade (12). Thus, leveraging all available sequences for every patient would likely yield a more accurate classification of the glioma grade. However, usage of a single postcontrast T1-weighted sequence increases the applicability of our model in a clinical scenario for a wide variety of different tumor types, as this sequence is routinely performed in tumor imaging protocols. Nevertheless, in our future work, we plan to explore strategies so that the model can take data from any available subset of modalities as input and be able to produce reliable classification results. Second, the class labels for METS, MEN, AN, PA, and HLTH acquired from the WUSM were based only on radiology reports. Third, with the exception of HGG and LGG data, data for all other classes were acquired from the WUSM. Although the model performance was stable under fairly diverse acquisition parameters for the WUSM data, the performance on those classes needs to be validated on an independent dataset. Last, an initial set of the most common intracranial tumor types that had overlapping but distinct attributes to the facilitate development of the network was chosen for this study. Future work will include additional tumor types.

In conclusion, we developed a CNN model that can accurately classify six different types of brain tumor and discriminate pathologic images from images depicting HLTH. The model can be extended to other brain tumor types or neurologic disorders that exhibit anomalous intensity profiles on MR images. The network is an initial step toward developing an artificial intelligence–augmented radiology workflow that can support image interpretation by providing quantitative information and statistics to the clinician to help improve diagnosis and prognosis.

Author contributions:: Guarantor of integrity of entire study, D.M.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, S.C., A.S., D.M.; clinical studies, P.L., M.H., D.M.; experimental studies, S.C., A.S., M.M., P.L., D.M.; statistical analysis, S.C., A.S.; and manuscript editing, S.C., A.S., P.L., D.M.

Disclosures of Conflicts of Interest: S.C. institution received grant from National Institutes of Health. **A.S.** author has stocks in TheraPanacea, which is a startup in France founded by PhD supervisor. TheraPanacea develops a medical analysis platform intended to improve cancer treatment. This platform leverages artificial intelligence technology to enable medical practitioners to treat patients with cancer with radiation therapy with improved success and lesser risks. **M.M.** institution received grant from NIH, P30 NS048056, NINDS Center Core for Brain Imaging (NCCBI). The NCCBI provides informatics, analysis, and imaging methodology service and consultation to the Washington University neuroimaging community; author employed by Washington University, St Louis as a research instructor in the department of radiology. **M.L.** employed by Washington University Medical School. **M.H.** institution received grant from National Institutes of Health. **D.M.** institution received grant from National Institutes of Health.

References

1. Brain tumors. American Association of Neurological Surgeons Web site. https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Brain-Tumors. Accessed August 27, 2021.
2. Zacharaki EI, Wang S, Chawla S, et al. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magn Reson Med 2009;62(6):1609–1618.
3. Cheng J, Huang W, Cao S, et al. Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLoS One 2015;10(10):e0140381 [Published correction appears in PLoS One 2015;10(12):e0144479.].
4. Muhammad K, Khan S, Ser J Del, de Albuquerque VHC. Deep learning for multigrade brain tumor classification in smart healthcare systems: a prospective survey. IEEE Trans Neural Netw Learn Syst 2021;32(2):507–522.
5. Paul JS, Plassard AJ, Landman BA, Fabbri D. Deep learning for brain tumor classification. In: Krol A, Gimi B, eds. Proceedings of SPIE: medical imaging 2017—biomedical applications in molecular, structural, and functional imaging. Vol 10137. Bellingham, Wash: International Society for Optics and Photonics, 2017; 1013710.
6. Sajjad M, Khan S, Muhammad K, Wu W, Ullah A, Baik SW. Multi-grade brain tumor classification using deep CNN with extensive data augmentation. J Comput Sci 2019;30:174–182.
7. Afshar P, Plataniotis KN, Mohammadi A. Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2019; 1368–1372.
8. Mohsen H, El-Dahshan ESA, El-Horbaty ESM, Salem ABM. Classification using deep learning neural networks for brain tumors. Future Comput Inform J 2018;3(1):68–71.
9. Seetha J, Selvakumar Raja S. Brain tumor classification using convolutional neural networks. Biomed Pharmacol J 2018;11(3):1457–1461.
10. Chakrabarty S, LaMontagne P, Marcus DS, Milchenko M. Preprocessing of clinical neuro-oncology MRI studies for big data applications. In: Editor A, Editor B, eds. Proceedings of SPIE: Medical Imaging 2020—imaging informatics for healthcare, research, and applications. Vol 11318. Bellingham, Wash: International Society for Optics and Photonics, 2020; 1131809.
11. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Trans Med Imaging 2015;34(10):1993–2024.
12. Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge. ArXiv 1811.02629 [preprint] https://arxiv.org/abs/1811.02629. Posted November 5, 2018. Accessed August 27, 2021.
13. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 2017;4:170117.
14. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26(6):1045–1057.
15. Bakas S, Akbari H, Sotiras A, et al. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Cancer Imaging Archive Web site. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=24282668. Published 2017. Updated May 25, 2021. Accessed August 27, 2021.
16. Akkus Z, Ali I, Sedlář J, et al. Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. J Digit Imaging 2017;30(4):469–476.
17. Erickson B, Akkus Z, Sedlar J, Kofiatis P. Data from LGG-1p19qDeletion. Cancer Imaging Archive Web site. https://wiki.cancerimagingarchive.net/display/Public/LGG-1p19qDeletion. Published 2017. Updated February 4, 2021. Accessed August 27, 2021.
18. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. Brain tumor segmentation and radiomics survival prediction: contribution to the BraTS 2017 challenge. In: Crimi A, Bakas S, Kuijf H, Menze B, Reyes M, eds. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. BrainLes 2017. Vol 10670, Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2017; 287–297.
19. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13(10):281–305.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017. 618–626.
21. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018. 9525–9536.
22. Abdelaziz Ismael SA, Mohammed A, Hefny H. An enhanced deep learning approach for brain cancer MRI images classification using residual networks. Artif Intell Med 2020;102:101779.
23. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153–157.
24. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics 2000;56(2):345–351.
25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.