Title

Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging

Permalink
https://escholarship.org/uc/item/2186j7pf

Journal
Neuro-Oncology, 24(4)

ISSN
1522-8517

Authors
Cluceru, Julia
Interian, Yannet
Phillips, Joanna J
et al.

Publication Date
2022-04-01

DOI
10.1093/neuonc/noab238

Peer reviewed

Neuro-Oncology

24(4), 639–652, 2022 | https://doi.org/10.1093/neuonc/noab238 | Advance Access date 15 October 2021

Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging

Julia Cluceru, Yannet Interian, Joanna J. Phillips, Annette M. Molinaro, Tracy L. Luks, Paula Alcaide-Leon, Marram P. Olson, Devika Nair, Marisa LaFontaine, Anny Shai, Pranathi Chunduru, Valentina Pedoia, Javier E. Villanueva-Meyer, Susan M. Chang, and Janine M. Lupo

Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, California, USA (J.C., T.L.L., P.A.L., M.P.O., D.N., M.L., V.P., J.E.V.M., J.M.L.); MS in Analytics Program, University of San Francisco, San Francisco, California, USA (Y.I.); Department of Neurological Surgery, University of California, San Francisco, San Francisco, California, USA (J.J.P., A.M.M., A.S., P.C., S.M.C.); Department of Medical Imaging, University of Toronto, Toronto, Ontario, Canada (P.A.L.); Department of Pathology, University of California, San Francisco, San Francisco, California, USA (J.J.P.)

Corresponding Author: Janine M. Lupo, PhD, Department of Radiology and Biomedical Imaging, University of California, San Francisco, Byers Hall, Box 2532, 1700 4th Street, Suite 303D, San Francisco, CA 94158-2330, USA (janine.lupo@ucsf.edu).

Abstract

Background. Diagnostic classification of diffuse gliomas now requires an assessment of molecular features, often including IDH-mutation and 1p19q-codeletion status. Because genetic testing requires an invasive process, an alternative noninvasive approach is attractive, particularly if resection is not recommended. The goal of this study was to evaluate the effects of training strategy and incorporation of biologically relevant images on predicting genetic subtypes with deep learning.

Methods. Our dataset consisted of 384 patients with newly diagnosed gliomas who underwent preoperative MRI with standard anatomical and diffusion-weighted imaging, and 147 patients from an external cohort with anatomical imaging. Using tissue samples acquired during surgery, each glioma was classified into IDH-wildtype (IDHwt), IDH-mutant/1p19q-noncodeleted (IDHmut-intact), and IDH-mutant/1p19q-codeleted (IDHmut-codel) subgroups. After optimizing training parameters, top performing convolutional neural network (CNN) classifiers were trained, validated, and tested using combinations of anatomical and diffusion MRI with either a 3-class or tiered structure. Generalization to an external cohort was assessed using anatomical imaging models.

Results. The best model used a 3-class CNN containing diffusion-weighted imaging as an input, achieving 85.7% (95% CI: [77.1, 100]) overall test accuracy and correctly classifying 95.2%, 88.9%, 60.0% of the IDHwt, IDHmut-intact, and IDHmut-codel tumors. In general, 3-class models outperformed tiered approaches by 13.5%-17.5%, and models that included diffusion-weighted imaging were 5%-8.8% more accurate than those that used only anatomical imaging.

Conclusion. Training a classifier to predict both IDH-mutation and 1p19q-codeletion status outperformed a tiered structure that first predicted IDH-mutation, then 1p19q-codeletion. Including apparent diffusion coefficient (ADC), a surrogate marker of cellularity, more accurately captured differences between subgroups.

Key Points

- MRI and deep learning can predict new molecular subtypes of glioma with 86% accuracy.
- Classifying IDH and 1p19q mutations together was advantageous over a tiered structure.
- Diffusion-weighted imaging increased the generalization capacity of our models.

Importance of the Study

During 2019-2020, the consortium that informs the WHO has placed even greater emphasis on the delineation of glioma categories by a mutation in IDH and codeletion of 1p and 19q chromosomal arms, prioritizing these features over grade. Although radiomics and deep learning have been successful in predicting IDH-mutation status from anatomical images, less emphasis has been placed on predicting 1p19q-codeletion. This study demonstrates the benefit of: (1) using deep learning with transfer learning, (2) a single 3-class model over a 2-tiered approach that first predicts IDH-mutation then 1p19q-codeletion, and (3) including ADC maps from diffusion-weighted imaging in predicting new genetic subtypes of gliomas and evaluates the generalizability of our anatomical imaging models in an external multi-site cohort. These insights will be highly valuable for future larger, multi-site analyses of molecular subtypes.

Since the restructuring of the categorization of gliomas by the World Health Organization (WHO) in 2016 to include variations in underlying genetic and epigenetic alterations,[1] the consortium that informs the WHO has begun to place even greater emphasis on the delineation of glioma categories by a mutation in isocitrate dehydrogenase 1 and/or 2 (IDH1 and/or 2) and codeletion of 1p and 19q chromosomal arms, prioritizing these features over grade.[2–5] In contrast to the WHO 2016 guidelines that first stratify by grade and then use genetic alterations to further differentiate patients within a designated grade, the new 2021 WHO guidelines now recommend that the first diagnostic delineation relies on IDH-mutation, followed by 1p19q-codeletion status, as supported by evidence that these distinct genetic subtypes indicate drastic differences in overall survival and response to therapy.[6–9] Due to this increasing emphasis on genetic alterations as a diagnostic tool, it has become a clinical standard to perform genetic testing on tissue acquired during surgery to decide subsequent treatment.

Because genetic testing can be a costly and time-consuming process and there remains cases where resection is not recommended, an alternative, noninvasive approach for obtaining this crucial genetic information is highly attractive. With a growing body of evidence that features from magnetic resonance imaging (MRI) are predictive of genetic alterations in IDH and 1p19q-codeletion,[10–15] image analysis techniques have the potential to provide a fast, noninvasive complementary pathway for identifying genetic alterations, which is highly relevant when these molecular markers are needed rapidly to determine clinical trial eligibility, sometimes even before surgery if the trial design involves initiating the drug prior to resection. Prior knowledge of genomics derived from imaging can also help patients with less aggressive phenotypes decide the timing of their treatment and provide an additional data point if a negative IDH result on immunohistochemistry (IHC) is found without the need to undergo genetic sequencing, which is often less accessible and even more costly.

Several prior studies have implemented radiomics, machine learning, and/or deep learning to accomplish this task of classifying tumors into their genetic subtype[16–22] based on the premise that they will be better at detecting known hallmark features of each subgroup (such as the characteristic larger percentage enhancing component often with necrotic core present in IDH-wildtype tumors, the more diffuse boundaries observed in both IDH-wildtype and IDH-mutant 1p19q-codeleted lesions, and the "T2-FLAIR mismatch" sign specific to IDH-mutant 1p19q-intact gliomas)[23] as well as distinguishing subtle novel features. In 2017, Li et al reported that the automatic extraction of radiomic features using deep learning successfully predicts IDH-mutation status in grade 2 glioma; however, this study was limited by requiring a priori knowledge of the tumor grade obtained through pathological tissue evaluation, limiting its application in the presurgical setting.[18] It was also marked with an uncommon enrichment of IDH-wildtype grade II glioma in their patient cohort.[18] Since then, many studies have leveraged The Cancer Imaging Archive, either alone or together with internal datasets, to evaluate the ability of deep learning and radiomics to predict a patient's IDH-mutation.[16,19–22] All of these studies use only anatomical MRI, which is advantageous in that these images are universally acquired with standard imaging protocols and require minimal preprocessing, but lack the associated benefits of physiological imaging that more closely reflects the underlying tumor biology. Even less emphasis has been placed on predicting 1p19q-codeletion, with few studies reporting attempts to separately classify this mutation, and those that have neglected to first exclude IDH-wildtype lesions from their classification.[17,19] We hypothesize that using a 3-class model that predicts genetic subgroup, rather than individual mutation status, plus a strategy that incorporates models that have been pre-trained on classifying large, publicly available images, will improve the accuracy over a tiered approach for predicting IDH and 1p19q mutations because of the shared imaging features of these mutations.

As diffusion-weighted imaging has become a standard in mainstream routine clinical imaging of gliomas at most institutions, there is a growing body of evidence that features derived from MR diffusion-weighted imaging are predictive of both IDH-mutation and 1p19q-codeletion.[13,15,24–26] As part of this work, we also sought to evaluate whether the addition of maps of apparent diffusion coefficient (ADC) derived from diffusion-weighted imaging would improve both the accuracy and generalization to an unseen test set when included as one of the inputs to a deep convolutional neural network (CNN) trained to predict genetic subtype. Because the "T2-FLAIR mismatch" sign has been shown to identify IDH-mutated gliomas with intact 1p19q,[10–13] we hypothesize that using T2 imaging together with T2-FLAIR and ADC will improve the accuracy of the IDH-mutant, 1p19q-intact classification, whereas including post-contrast T1-weighted images, ADC, and either the T2-weighted or T2-FLAIR images will improve the classification of IDH-wildtype and IDH-mutant, 1p19q-codeleted subgroups.

Methods

Patient Characteristics and Study Design

Imaging, pathological, and clinical data from 502 adults who were newly diagnosed with a pathologically confirmed glioma at our institution between 2007 and 2019 were assessed in this retrospective, IRB-approved study. Patients were excluded if either their molecular subgroup was indeterminable (n = 87) or their preoperative MRI acquisitions did not include either T1-weighted post-contrast (T1c), T2-weighted (T2), or T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) images (n = 20) (Supplementary eFigure 1). The study design was comprised of 2 parts: hyperparameter search and model comparison. The hyperparameter search phase was first performed to select the best hyperparameters for each of our 3 deep learning models. The model comparison phase then investigated the benefits of (1) using a 2-tiered, binary classification approach as opposed to a single-tiered, 3-class classifier, and (2) including ADC images as an input channel (Supplementary eFigure 2). The best performing models were then tested on the 2019 BRAin Tumor Segmentation (BraTS) challenge[27–29] images from The Cancer Genome Atlas (TCGA) Research Network (www.cancer.gov/tcga) to determine generalization to external, multi-institutional cohorts.

Assessment of Genetic Alterations

IDH-mutation status for UCSF cases was evaluated by Sanger sequencing of IDH1 and IDH2 genes or by IHC (IDH1R132H, H09, Dianova GmbH, Hamburg, Germany) using standard techniques (details provided in Supplementary eDocument 1). Negative IDH-mutation results based on IHC in lower-grade gliomas or patients 55 years of age or younger at diagnosis were either validated by sequencing or excluded. All patients with a GBM pathological diagnosis and negative IDH-mutation based on IHC who were >55 years of age were considered IDH-wildtype, per the guidelines of the European Association for Neuro-Oncology.[30] IDH-mutation status for all TCGA cases was assessed via Sanger sequencing. Confirmed negative IDH1 and IDH2 mutated samples were classified as IDH-wildtype ("IDHwt"). Our IDH-mutated tumors were further classified based on either 1p19q-codeletion status or ATRX alterations for UCSF cases, or solely 1p19q-codeletion status for TCGA data. Since tumors with 1p/19q-codeletion ("IDHmut-codel") almost invariably have IDH and TERT promoter mutations and are almost mutually exclusive with ATRX mutations, IDHwt gliomas and IDH-mutant ("IDHmut") gliomas with ATRX alterations were not tested for 1p/19q-codeletion unless it was performed clinically. ATRX was assessed by IHC (HPA001906, Sigma Aldrich, St. Louis, MO, USA) performed at the UCSF Brain Tumor Research Center using previously published methods,[31] while the presence of a 1p/19q-codeletion was determined with clinical FISH assays. IDHmut tumors that either had an ARTX alteration or were lacking a 1p19q-codeletion were classified as 1p/19q-intact ("IDHmut-intact").

Image Acquisition and Processing

All patients underwent MRI examinations on a 3T Discovery MR750 scanner (GE Healthcare, Waukesha, WI, USA) using an eight-channel phased-array head coil prior to surgical resection. The imaging protocol included T2-weighted FLAIR and fast spin echo (FSE) images, along with 3D T1-weighted IR-SPGR imaging pre- and post- the injection of a gadolinium-based contrast agent. Diffusion tensor images (DTI) were obtained in the axial plane with b = 1000 s/mm^2 and either 6 gradient directions and 4 excitations or 24 gradient directions and 1 excitation or b = 2000 s/mm^2 and 55 gradient directions (repetition-time [TR]/echo-time [TE] = 1000/108 ms, voxel size = 1.7-2.0 × 1.7-2.0 × 2.0-3.0 mm). To calculate the ADC maps, a pipeline that utilized components of FMRIB's Diffusion Toolkit was applied to estimate relevant diffusion parameters from the DWI and DTI data as previously described.[32]

All images from the UCSF cohort were registered to the T1c image volume using either FMRIB's FSL Linear Image Registration Tool (FLIRT) or 3D Slicer's BRAINSFit tool with B-spline warping and resampled to an identical 1-mm isotropic spatial coordinate.[32–34] Brain masks were derived using the Brain Extraction Tool (BET) (FSL, FMRIB) and were visually verified to have worked properly.[32] All images were subjected to signal intensity normalization through a multistep process: (i) the images were multiplied by the brain mask, (ii) pixels above the 99.9th percentile were thresholded to the pixel intensity denoting the 99.9th percentile; (iii) the mean was subtracted from each pixel and the result divided by the standard deviation; (iv) the images were scaled to lie between a value of 0 and 1 by subtracting the minimum and dividing by the difference between the maximum and the minimum pixels. The T2-lesion (T2L), defined as hyperintense signal on FLAIR images, and contrast-enhancing lesion (CEL), or hyperintense signal on the T1c images that was not enhancing on the original T1-weighted images, had been previously contoured for the UCSF dataset using either 3D Slicer, or in-house developed software.[35] The 2019 BraTS data from the TCGA cohort were already preprocessed, segmented, and curated as part of this publicly available imaging dataset as described previously.[27–29]

Input images containing tumor were automatically selected and processed to form multi-contrast RGB colormaps according to Figure 1. Masks of the 3D segmented lesion volumes were first used to automatically select the image slice containing the largest tumor area in each direction. Additional slices spaced 5 mm apart were added in each direction until the edge of the tumor mask was reached. Images were also automatically cropped to a rectangular bounding box surrounding each lesion of

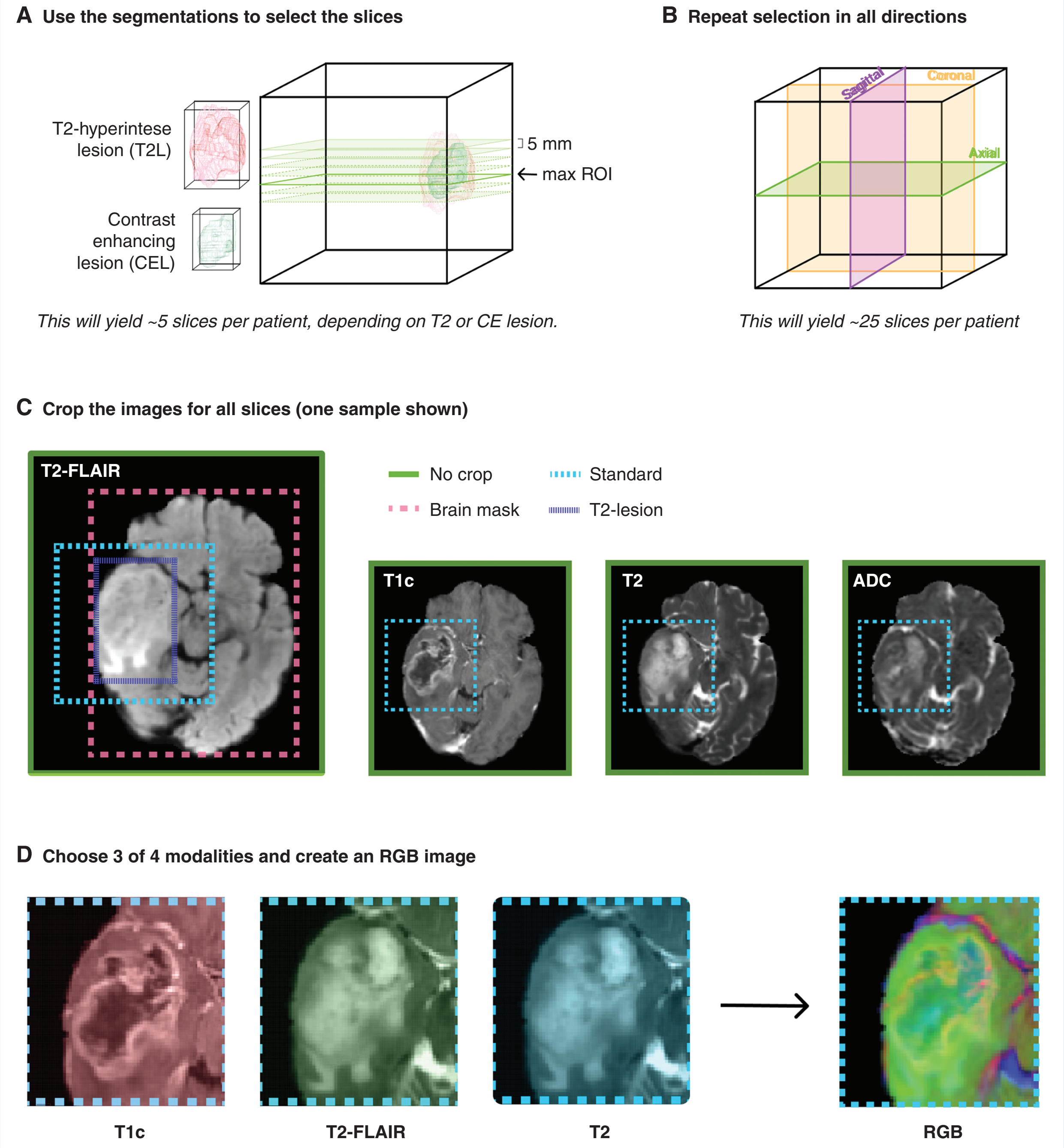
A Use the segmentations to select the slices
T2-hyperintese lesion (T2L)
Contrast enhancing lesion (CEL)
5 mm
max ROI
This will yield ~5 slices per patient, depending on T2 or CE lesion.
B Repeat selection in all directions
Sagittal
Coronal
Axial
This will yield ~25 slices per patient
C Crop the images for all slices (one sample shown)
T2-FLAIR
No crop
Standard
Brain mask
T2-lesion
T1c
T2
ADC
D Choose 3 of 4 modalities and create an RGB image
T1c
T2-FLAIR
T2
RGB

Fig. 1 Schematic of image processing strategy. (A) Segmented contrast-enhancing (CEL) or T2 (T2L) lesions were used to automatically select the slices by first selecting the central slice with maximum area and expanding every 5 mm until the boundary of the lesion was reached. (B) This process was repeated in each direction: axial, coronal, and sagittal. (C) Images were then cropped to either the brain mask, the T2L, a standard size, or were not cropped. (D) Three of the four sequences of interest (T2-FLAIR, T1c, T2, and ADC) were placed in the R, G, and B channels of a color image that was used as the input to the network.

interest. For each of the 4 cropping strategies shown in Figure 1C, combinations of 3 image modalities were then merged to create a multi-contrast RGB image for each lesion slice, where each image contrast was saved in as a red, green, or blue color channel, with values ranging from 0-256. The creation of the multi-contrast RGB image allowed us to take advantage of transfer learning from a pre-trained network that had exceptionally high accuracy in classifying a large-scale dataset of millions of color images (ImageNet).

Baseline Models From Clinical Metrics

Since age and the presence or absence of contrast enhancement are known predictors of IDH-mutation status, and anatomical MR images of the brain can predict age with high accuracy,[36,37] we first used logistic regression models in the scikit-learn package to establish an interpretable baseline prediction accuracy for which to compare our models.[38] Age and the presence of contrast enhancement were included as independent variables and used in: (1) a tiered binomial logistic regression structure to predict IDH-mutation status followed by 1p19q-codeletion status and (2) a 3-class multinomial logistic regression model to directly predict molecular subgroup. The presence of contrast enhancement was automatically quantified as having a CEL volume greater than 150 mm^2, the cutoff for the lower 10th percentile, and included in the first-tier and 3-class models. In the 3-class model, multinomial loss fit was minimized across the entire probability distribution to balance class weights. We also tested whether the 2 datasets (UCSF and TCGA) differed significantly using the Mann-Whitney *U* test for age and the χ^2 test for categorical variables sex, mutation status, and the presence of contrast enhancement.

Hyperparameter Search

In order to find a reasonable starting point to train our models, we first searched through a set of randomly generated hyperparameters that included various learning parameters, VGG and ResNet model architectures, whether or not the network was pre-trained on ImageNet, and image preprocessing strategies such as the extent of cropping, to find a reasonable starting point for each of our 3 classification models: IDH-mutation only, 1p19q-codeletion only, and 3-class molecular subgroups.[39–41] Two different slice-combining paradigms were also compared: (1) pooling slices for a single prediction per patient as performed in MRNet from Bien et al[42]; and (2) treating each slice individually while training and combining slice predictions afterward as described by Chang et al.[21] The individual hyperparameters that were tested are listed in Supplementary eTable 3, while hyperparameters that remained fixed were the learning rate cycling strategy ("One Cycle"),[43] the optimization algorithm ("Adam"),[44] and the weight decay coefficient (0.01). During this phase, model inputs were restricted to the T2, T2-FLAIR, and T1c images, and the 1p19q-codeletion experiments began with the model pre-trained on the IDH-mutation status classification. Overall model accuracy, defined as the total number of patients a model predicted correctly divided by the total number of patients evaluated, was used to evaluate the models' efficacy. The top hyperparameter sets from each outcome were rerun 5 times with different seeds to account for stochasticity introduced during gradient descent. The set of hyperparameters with the best performance based on the mean overall classification accuracy of the 5 seeds on the validation set that also had training and validation loss curves that were steadily decreasing or stable, was then chosen. The details describing the model development phase, including the hyperparameter search space, can be found in the Supplementary Materials (Supplementary eDocument 2 and Supplementary eTable 3).

Model Comparison

Using the set of hyperparameters for the top performing model for each classification experiment determined during the model development phase, we investigated the impact of using a 2-tiered vs single 3-class structure approach and the addition of ADC as one of the input image channels. As in the model development phase, training and validation loss plots were visually compared in order to ensure that there was appropriate reduction in validation loss as the model trained and to prevent overfitting at a certain epoch (examples of which can be found in Supplementary eFigure 6). For the 2-tiered approach, IDH-mutation status was predicted in the first tier, while 1p19q-codeletion was then classified from the IDH-mutated tumors in the second tier. The final class accuracy for the IDHwt subgroup was determined from the output of the first tier, while the class accuracy of predicting the other 2 subgroups was determined by the prediction accuracies of the second tier. For each classification approach (2-tiered and 3-class), ADC maps were then included as 1 of the 3 input channels (in place of either the T1c, T2, or T2-FLAIR image volumes) and trained with the same set of hyperparameters run with 5 different seeds. Confidence intervals (CI) were calculated using bootstrapping, whereby slices were randomly selected with replacement until the number of original slices was achieved, which results in approximately 2/3 of the slices (with 1/3 repeated) for each bootstrapped sample. For each repetition, the per-patient prediction was calculated from these slices. This process was repeated 1000 times for each of the training, validation, and test sets in order to generate the 2.5th and 97.5th percentile of these repetitions, which were reported as the 95% CI.

Model explanation was performed using GradCAM, a heatmap-based feature attribution method. In contrast to methods that use "guided" back-propagation as a part of feature attribution, GradCAM has been validated in the deep learning literature to assign feature importance to areas of the image.[45] This allowed for quick visual confirmation that our models were behaving as expected by extracting features in the areas that align with human interpretation. We used these GradCAM maps to help interpret our best and worst predicted patient examples.

Model Generalization

In order to evaluate whether our developed anatomical models were able to generalize to data from multiple institutions acquired using different scanners and acquisition parameters that result in variations in image contrast and resolution, the publicly available TCGA dataset together with the post-processing and labeling performed for the BraTS challenge were used to establish an independent dataset for testing. The BraTS imaging dataset was preprocessed with the same specifications as our data with expert segmentations, but because this dataset did not include diffusion data, we were only able to validate our best anatomical 2-tiered and 3-class models with this dataset.

Results

Characteristics of the Study Sample

The clinical characteristics of the entire dataset, consisting of 384 patients from UCSF and 147 patients from the TCGA dataset, are summarized in Table 1. While sex was not statistically significantly different between the 2 cohorts (59% male in UCSF vs 52% male in TCGA), patients at UCSF were statistically significantly younger than patients in the TCGA dataset, with mean age of 47.4 ± 15.3 years compared to 53 ± 14.9 years (*P* < .001; Supplementary eFigure 3). UCSF and TCGA datasets also significantly differed in the proportion of IDH-mutation status, with 269 (62%) mutated UCSF patients and 56 (22.8%) mutated TCGA patients (*P* < .001), and the frequency of enhancement, with 206 (47%) UCSF patients enhancing and 120 (82%) TCGA patients enhancing (*P* < .001).

Baseline Models From Clinical Metrics

The detailed results of baseline clinical logistic regression models using solely age and the presence of contrast enhancement and trained on UCSF patients are presented with the average and class accuracies for the 3-class and 2-tiered results shown in Supplementary eTables 1 and 2, respectively. The tiered logistic regression achieved 67% [0.65, 1.00] overall accuracy, while the 3-class logistic regression achieved 71% [0.57, 1.00] overall accuracy on the UCSF test sets, which served as the basis for comparison for our deep learning models. However, for both the tiered and 3-class baseline models, the bootstrapped CI were very wide for every metric, with the best prediction accuracy achieved for IDHwt tumors (95% [0.86, 1.00]/91% [0.83, 0.94] for the UCSF/TCGA test sets) and worst for the IDHmut-codel group (40% [0.29, 0.82]/23% [0.06, 0.44] for the UCSF/TCGA test sets).

Hyperparameter Search

The impact of top features on patient validation accuracy is shown in Figure 2A–C (see Supplementary eFigure 4 for all hyperparameter search results). The first comparison, evaluating whether pooling slice predictions was advantageous compared to making slice-by-slice predictions, demonstrated that taking the mean of slice-by-slice predictions achieved higher patient-level validation accuracies compared with pooling (Figure 2A). The rest of the hyperparameter experiments were performed using slice-by-slice predictions only. The most notable findings from subsequent analyses was that in both 3-class and IDH-mutation only experiments, cropping to the T2-hyperintense lesion decreased performance (Figure 2B), while using models that were pre-trained on ImageNet improved performance (Figure 2C). The VGG-16 architecture performed the best for both the 3-class and IDH-mutation status tier of the 2-tiered model, while ResNet-18 outperformed other architectures in the 1p19q mutation tier. The results from the rest of the hyperparameter search can be found in Supplementary eFigure 4.

Model Comparison

Two-tiered vs single 3-class classifier. —When using anatomical images only, the best 3-class model resulted in an overall patient accuracy of 84.6% [0.826, 1.0], 82.0% [0.740, 1.0], and 81.6% [0.735, 1.0] for the training, validation and testing sets of UCSF data, with individual test class accuracies of 90.5% [0.90, 0.95], 77.8% [0.579, 0.737], and 70%

Table 1. Clinical and Demographic Characteristics of the Patient Population Included in This Study

		UCSF, n = 384		TCGA, n = 147		UCSF and TCGA Difference	
Age		Mean	SD	Mean	SD	*P*-value (Mann-Whitney *U* test)	*P*-value (χ^2 test)
		47.7	15.6	53	14.9	.007	n/a
Sex		Male	Female	Male	Female		
(n)		231	153	76	70	.05	.088
IDH status		IDHwt	IDHmut	IDHwt	IDHmut		
(n)		151	233	87	59	<.0001	<.0001
WHO Grade	Mutation status	Non-enhancing	Enhancing	Non-enhancing	Enhancing		<.0001
2	IDH-wildtype	0	1	2	0		
	IDH-mutation + 1p19q intact	69	4	11	8		
	IDH-mutation + 1p19q-codeletion	63	8	3	3		
3	IDH-wildtype	6	1	2	7		
	IDH-mutation + 1p19q intact	54	5	7	14		
	IDH-mutation + 1p19q-codeletion	6	7	1	6		
4	IDH-wildtype	0	143	0	76		
	IDH-mutation + 1p19q intact	1	13	0	6		
	IDH-mutation + 1p19q-codeletion	0	2	0	0		

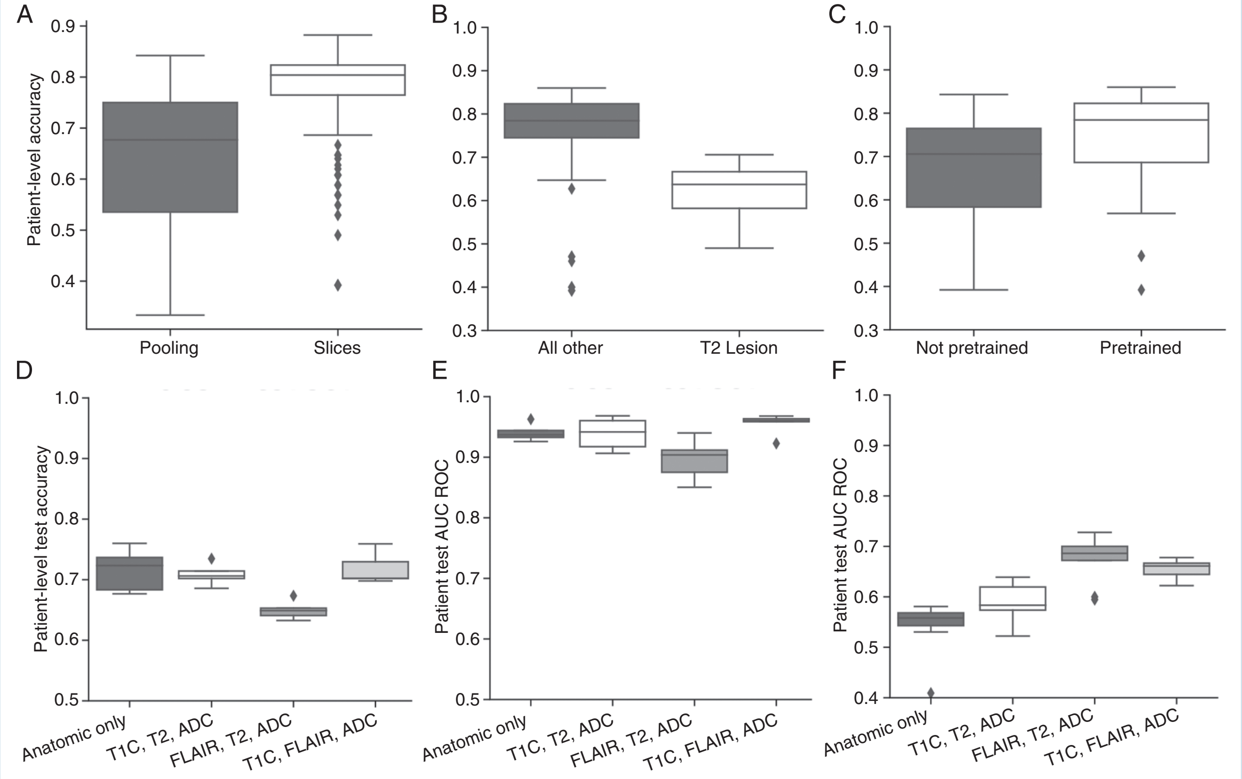

Fig. 2 Main insights from the hyperparameter search and effects of including ADC as an input image. (A–C) Hyperparameter search: (A) A slice-by-slice prediction approach improved the ability to achieve high accuracy on the validation data compared to average pooling of slices. (B) Cropping to the T2-lesion reduced the validation accuracy of the 3-class model. (C) Pre-training increased validation accuracy of the model. (D–F) Benefit of including ADC as an input image: (D) For the 3-class models, lower generalization accuracy was observed when ADC replaced T1c, while replacing the T2 image with ADC achieved the best performance. (E) For the IDH-mutation tier, test accuracy was slightly improved using T1c, T2-FLAIR, and ADC as input images. (F) For the 1p19q mutation tier, replacing T1c with ADC significantly improved test accuracy.

[0.60, 0.90] for the IDHwt, IDHmut-intact, and IDHmut-codel subgroups, respectively. The final model parameters are shown in Supplementary eTable 4. Although the best performing 2-tiered structure resulted in a higher overall patient accuracy in training (94% [0.902, 1.0]) and relatively similar accuracy as the 3-class model in validation (84.0% [0.780, 1.0]), this model did not generalize as well to the UCSF test set (69.4% [0.633, 1.0] accuracy). Detailed class accuracies for each training, validation, and testing cohort along with confusion matrices are shown in Supplementary eTables 5 and 7, while the final predictions of the 2-tiered vs 3-class models are shown in Table 2 and Figure 3A. However, when evaluating the ability of each approach to generalize to the multi-institutional TCGA data, the 2-tiered structure outperformed the 3-class model (81.6% [0.769, 1.0] compared to 68.7% [0.678, 1.0] overall accuracy). Although both of these approaches were able to predict the IDHwt group with high accuracy (91.2% [0.868, 0.912] for 2-tiered and 95.6% [0.956, 0.967] for 3-class), the 2-tiered model was also able to predict the IDHmut-intact subgroup with 86.0% [0.744, 0.857] accuracy while the 3-class model accuracy was only 32.6% [0.279, 0.372] for this subtype. Both approaches failed at correctly predicting any of the tumors in the IDHmut-codel subgroup.

Benefit of adding ADC to the model.—We next investigated whether the addition of ADC was advantageous compared with using anatomical images only as inputs to the model for both the 3-class (Supplementary eTable 6) and 2-tiered (Supplementary eTable 8) approaches. In both the 3-class and IDH models (first tier), the best performance was achieved when ADC maps were used along with T1c and T2-FLAIR images as inputs, while replacing the T1c image with ADC decreased performance from using anatomical imaging alone (Figure 2D and E). For the 1p19q mutation classification (second tier) of the 2-tiered approach, however, the best performance was achieved when ADC replaced the T1c image (Figure 2F), resulting in a 10% increase in test accuracy for this tier, and a 20% increase in accuracy for the IDHmut-codel subgroup (Table 2). Although, in general, including ADC in both models outperformed their anatomical imaging-only counterparts, the best generalization power to the test set was achieved with a 3-class model that replaced T2-weighted images with ADC maps. The final overall patient accuracies achieved were 86.0% [0.839, 1.0], 80.0% [0.720, 1.0], and 85.7% [0.771, 1.0] on training, validation, and UCSF test sets, with final test set class accuracies of 95.2% [0.857, 0.952] for

Table 2 Summary of Classification Results for All Models

2-Tiered Logistic Regression	IDHwt	IDHmut-intact	IDHmut-codel	Overall	3-Class Logistic Regression	IDHwt	IDHmut-intact	IDHmut-codel	Overall
Training					*Training*				
Accuracy	93.80%	62.60%	51.50%	72.30%	Accuracy	95.50%	54.20%	53.00%	70.20%
Confidence interval	[0.895, 0.973]	[0.547, 0.707]	[0.412, 0.621]	[0.681, 1.0]	Confidence interval	[0.921, 0.983]	[0.46, 0.62]	[0.429, 0.634]	[0.656, 1.0]
Validation					*Validation*				
Accuracy	100.00%	70.00%	45.50%	76.00%	Accuracy	100.00%	55.00%	54.50%	72.00%
Confidence interval	[1.0, 1.0]	[0.524, 0.857]	[0.222, 0.727]	[0.66, 1.0]	Confidence interval	[1.0, 1.0]	[0.364, 0.739]	[0.286, 0.818]	[0.620, 1.0]
Test					*Test*				
Accuracy	95.00%	47.40%	50.00%	67.30%	Accuracy	95.20%	61.10%	40.00%	71.40%
Confidence interval	[0.864, 1.0]	[0.538, 0.889]	[0.125, 0.667]	[0.653, 1.0]	Confidence interval	[0.857, 1.0]	[0.364, 0.739]	[0.286, 0.818]	[0.571. 1.0]
TCGA					*TCGA*				
Accuracy	90.10%	25.60%	23.10%	65.30%	Accuracy	91.20%	23.30%	23.10%	65.30%
Confidence interval	[0.825, 0.935]	[0.321, 0.575]	[0.071, 0.444]	[0.63, 1.0]	Confidence interval	[0.916, 1.0]	[0.118, 0.333]	[0.059, 0.444]	[0.596, 1.0]
2-tiered anatomical	IDHwt	IDHmut-intact	IDHmut-codel	Overall	3-class anatomical	IDHwt	IDHmut-intact	IDHmut-codel	Overall
Training					*Training*				
Accuracy	97.30%	95.30%	86.40%	94.00%	Accuracy	92.90%	95.30%	53.00%	84.60%
Confidence interval	[0.946, 0.973]	[0.916, 0.963]	[0.742, 0.848]	[0.902, 1.0]	Confidence interval	[0.902, 0.938]	[0.925, 0.962]	[0.492, 0.591]	[0.826, 1.0]
Validation					*Validation*				
Accuracy	89.50%	90.00%	63.60%	84.00%	Accuracy	100.00%	65.00%	81.80%	82.00%
Confidence interval	[0.895, 0.947]	[0.850, 0.900]	[0.455, 0.636]	[0.780, 1.0]	Confidence interval	[0.947, 1.0]	[0.5, 0.65]	[0.727, 0.909]	[0.74, 1.0]
Test					*Test*				
Accuracy	81.00%	77.80%	30.00%	69.40%	Accuracy	90.50%	77.80%	70.00%	81.60%
Confidence interval	[0.714, 0.810]	[0.684, 0.842]	[0.100, 0.400]	[0.633, 1.0]	Confidence interval	[0.9, 0.95]	[0.579, 0.737]	[0.6, 0.9]	[0.735, 1.0]
TCGA					*TCGA*				
Accuracy	91.20%	86.00%	0.00%	81.60%	Accuracy	95.60%	32.60%	0.00%	68.70%
Confidence interval	[0.868, 0.912]	[.744, 0.857]	[0.0, 0.077]	[0.769, 1.0]	Confidence interval	[0.956, 0.967]	[0.279, 0.372]	[0.0, 0.077]	[0.678, 1.0]
2-Tiered With ADC	IDHwt	IDHmut-intact	IDHmut-codel	Overall	3-Class With ADC	IDHwt	IDHmut-intact	IDHmut-codel	Overall
Training					*Training*				
Accuracy	97.30%	96.30%	80.30%	93.00%	Accuracy	92.00%	94.40%	62.10%	86.00%
Confidence interval	[0.946, 0.973]	[0.916, 0.963]	[0.712, 0.803]	[0.891, 1.0]	Confidence interval	[0.911, 0.938]	[0.916, 0.963]	[0.561, 0.652]	[0.839, 1.0]
Validation					*Validation*				
Accuracy	89.50%	90.00%	45.50%	80.00%	Accuracy	94.70%	80.00%	54.50%	80.00%
Confidence interval	[0.895, 0.947]	[0.800, 0.900]	[0.364, 0.545]	[0.760, 1.0]	Confidence interval	[0.895, 0.947]	[0.7, 0.8]	[0.364, 0.545]	[0.72, 1.0]
Test					*Test*				
Accuracy	81.00%	83.30%	50.00%	75.50%	Accuracy	95.20%	88.90%	60.00%	85.70%
Confidence interval	[0.700, 0.800]	[0.684, 0.842]	[0.221, 0.500]	[0.633, 1.0]	Confidence interval	[0.857, 0.952]	[0.778, 0.944]	[0.4, 0.6]	[0.771, 1.0]

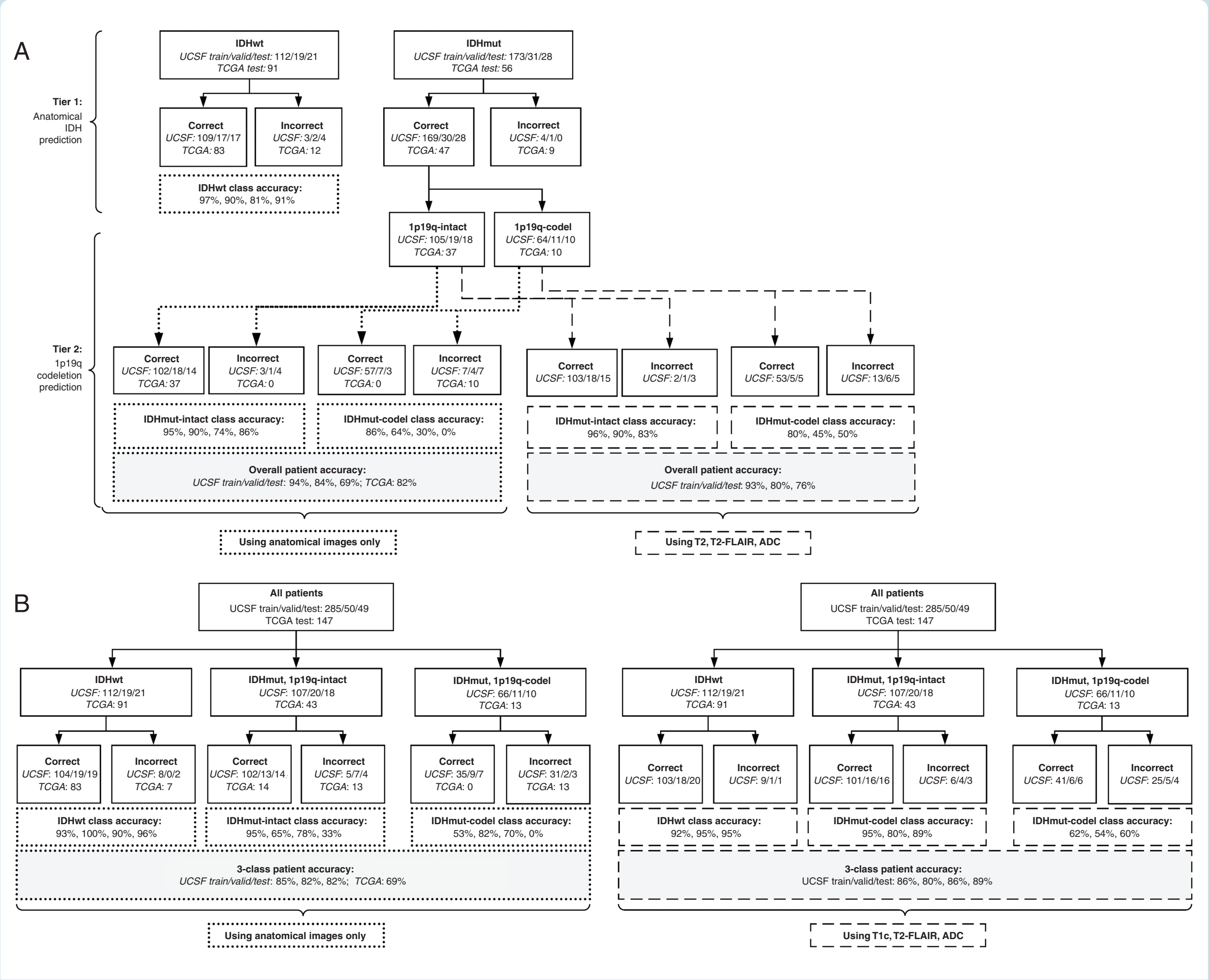

Fig. 3 Final patient accuracy and class accuracies of the final four models: (A) A 2-tiered structure with all anatomical images (left) and ADC replacing the T1c image in the second tier (right). (B) A 3-class structure with anatomical images only (left) and ADC replacing T2 as the third channel (right). An increase in the IDHmut-codel accuracy and overall accuracy was observed when ADC maps were included as input images. The best model was the 3-class model that included ADC.

IDHwt, 88.9% [0.778, 0.944] for IDHmut-intact, and 60.0% [0.4, 0.6] for IDHmut-codel subgroups (Figure 3B;Table 2).

Visualization and interpretation.—Figure 4 illustrates representative GradCAM images for the best (>90% confidence that the genetic alteration was ground truth) and worse (>50% confidence that the genetic alteration was other than its ground truth) predictions for the best 3-class model with ADC, T1c, and T2-FLAIR images as inputs. The most prominent features of IDH-wildtype gliomas were a ring-enhancing lesion surrounding a necrotic core, with elevated T2-hyperintensity on the T2 FLAIR image and reduced, more heterogeneous ADC values. Correctly predicted IDHmut-1p19qcodel tumors exhibited uniformly elevated T2-hyperintensity on T2-FLAIR images with more diffuse regions of heightened ADC and T1-hypointensity, while IDHmut-intact tumors tended to be larger with mostly elevated but textured ADC, moderate T2-FLAIR-hyperintensity, and the most hypointense on T1-weighted images. Incorrectly predicted IDHwt tumors were often non-enhancing, while IDHmut-codel tumors were frequently predicted as IDHmut-intact because the network was not looking at the right part of the image.

Discussion

In this study, we systematically investigated different sets of hyperparameters to achieve the optimal deep learning framework and MRI sequences for jointly identifying the IDH-mutation and 1p19q-codeletion status of a glioma patient prior to surgery. To our knowledge, this is the first study to: (a) classify molecular subgroup using imaging and deep learning; (b) investigate the impact of including ADC; (c) incorporate a pre-training strategy that includes the generation of an RGB color image from 3 grayscale MR images; and (d) thoroughly evaluate differences between various deep learning strategies through extensive hyperparameter searching complemented by training/validation loss curves and a feature attribution technique.

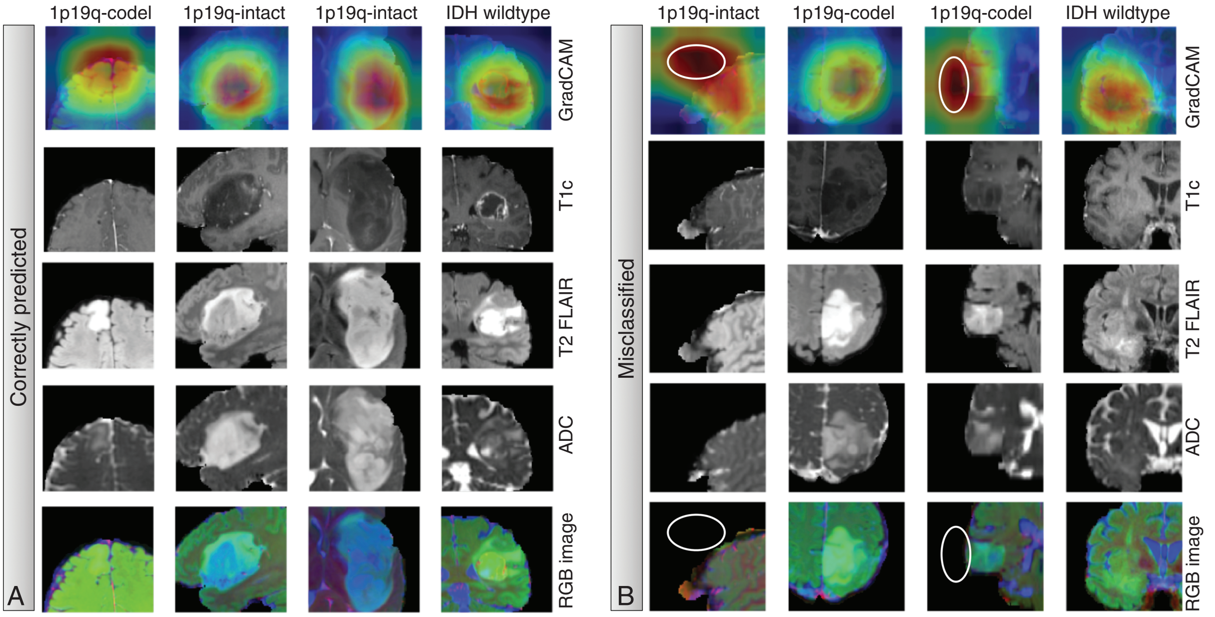

Fig. 4 Visualization of imaging features and GradCAM analysis of the best (A) and worst (B) predicted patients with the final 3-class model that included ADC. GradCAM maps of the worst predicted patients often indicated that these models were looking outside of the tumor region when making their decision (white ellipses), compared to looking at the lesion for all of the correctly predicted patients. Top column lists true groupings. IDHwt and IDHmut-intact were misclassified as IDHmut-codel gliomas, while the IDHmut-codel examples were predicted as IDHmut-intact.

Our best performing model was a 3-class, VGG-16 model pre-trained on ImageNet that included T1c, T2-FLAIR, and ADC images as inputs and predicted patients in our UCSF test set with promising overall (85.7%) and individual class accuracy (IDHwt: 95.2% [95% CI = (0.857, 0.952)]; IDHmut-intact: 88.9% [95% CI = (0.778, 0.944)]; IDHmut-codel: 60.0% [95% CI = (0.40, 0.60)]). A 3-class model approach was advantageous compared to a tiered strategy that first predicted IDH, and then 1p19q-codeletion mutations. Adding ADC as one of the input images increased generalization to test sets for both the 3-class models and second-tier 1p19q models. All deep learning models outperformed the corresponding logistic regression baseline models using age and the presence of contrast enhancement alone, indicating that imaging features can provide additional insight into genetic alterations. Our GradCAM analyses confirmed that the final algorithm was in fact learning features derived from tumor regions and not surrounding areas.

As age and the presence of contrast-enhancing tumors are known predictors of IDH-mutation status, we constructed logistic regression models using these variables to serve as a benchmark for our models to outperform. This approach also ensured that the deep learner was more than a complex detector of age or the presence of contrast enhancement.[46] Using contrast enhancement as an input to our baseline logistic regression models improved their generalization to the UCSF validation, UCSF test, and TCGA test sets for the 3-class and IDH model to 70%-72% overall accuracy. These final patient and class accuracies served as a benchmark for which to compare our deep learning models.

Before implementing our hypothesis-driven comparisons on the influence of ADC and modeling approach, we performed an extensive hyperparameter search to determine the optimal set of network training parameters for each set of experiments. First, 2 different slice-combining paradigms were compared: (1) pooling slices for a single prediction per patient[42]; and (2) treating each slice individually in training and then afterwards combining slice predictions.[21] In (1), all slices from a single patient are used in the same batch such that the number of slices becomes the effective batch size. Average or max pooling is then employed in the final layer to condense all slices into a single feature vector which generates a single prediction per patient. As a result, a single value is back propagated through the network for each patient after the loss is calculated. In contrast, (2) treats each slice independently such that a batch often contains slices from many patients; in turn, backpropagating gradients on a slice-by-slice basis and calculating a final patient-level prediction only after training is complete. Updating network weights based on individual slices resulted in better training/validation loss curves, as well as an increase in the overall patient-level accuracy as shown in Figure 2A. A marked decrease in prediction capability for both the 3-class and 2-tiered settings was observed when cropping to the T2 lesion compared with no cropping or cropping to a standard-sized rectangle. This is in line with the notion that the location of the lesion within the brain is associated with IDH-mutation status.[47] Although the ResNet-18 architecture most frequently resulted in models with higher average accuracy, the VGG-16 architecture had achieved the highest accuracy when other hyperparameters were optimized. This is not all that surprising given VGG-16's 3 terminal fully connected layers that could be helpful in capturing the heterogenous characteristics present in these lesions by allowing for different interactions among features, and the fact that the benefits of the residual connections in ResNet architectures typically are not realized until an order of magnitude of more data is used in training.

We hypothesized that a single 3-class model that was trained to predict molecular subgroup by classifying both IDH-mutation status and 1p19q-codeletion simultaneously would outperform a 2-tiered cascaded approach because: (1) the second tier predicting 1p19q-codeletion had a limited number of patients from which to learn imaging features; and (2) learned imaging features could be shared between tasks. Our results supported this hypothesis, regardless of whether or not ADC was included in our models. The reduced overall training accuracy of the 3-class model also suggests that the model was less likely to overfit when capturing features of 3-classes, boosting its performance on the test set compared to the 2-tiered approach. Supplementary eFigure 5 shows class accuracies plotted from the 3-class experiments that were performed during the hyperparameter search phase, depicting the tradeoff between a model's ability to predict IDHmut-codel patients and IDHmut-intact patients correctly. As the ability to predict IDHmut-codel patients increased, the prediction accuracy for IDHmut-intact patients diminished, while the ability to predict IDHwt patients remained stable. This result implies that even in the multiclass setting, the power of deep learning models to discriminate the 1p19q-codeletion was still limited. Although the 2-tiered approach more accurately classified the TCGA cohort compared to the 3-class model, the second tier incorrectly predicted all of the TCGA IDH-mutated patients as 1p19q-intact, further supporting improved generalizability with the 3-class model. This poor generalization of the IDHmut-codel class is most likely due to the fact that 80% of the UCSF IDHmut-codel lesions were non-enhancing, whereas only 31% of the IDHmut-codel patients from the TCGA dataset had non-enhancing lesions. Heterogeneous scan parameters field strengths, and vendor system platforms resulting in images of varying quality and contrasts in the TCGA data that were not seen in training, may also play a role.

Using ADC in place of an anatomical imaging sequence conferred an advantage in test accuracy for both the tiered and 3-class settings (Table 3). This advantage was particularly evident when comparing experiments predicting 1p19q-codeletion (Figure 3), where we observe the greatest generalization power in models using ADC together with T2 and T2-FLAIR. This finding was expected given that the mismatch in the T2 and T2-FLAIR signals contains imaging features specific to IDHmut-intact patients. When comparing the 3-class models with and without ADC, a more balanced result between IDH-mutated classes was achieved with ADC: 70% IDHmut-codel and 74% IDHmut-intact accuracy compared with the 60% IDHmut-codel and 84% IDHmut-intact accuracy. In contrast, including ADC as a modality in place of either T1c, T2, or T2-FLAIR images did not confer an advantage in predicting IDH-mutation status alone, despite prior evidence that features derived from diffusion-weighted imaging can help differentiated IDH-mutation status.[15,24,25] This result, however, does not mean that ADC is not valuable, but rather that the loss of another more informative sequence outweighed the benefit of ADC. For both the 3-class and IDH model tier, replacing the T1c images with ADC substantially decreases generalization power to the UCSF test set as expected given the known association between the presence of contrast enhancement and IDH-wildtype tumors. Although a limitation of our study is that we were not able to validate these findings on the multi-institutional external cohort, the promise of incorporating ADC into deep learning models that predict molecular subgroup, especially for the 1p19q-mutation, is still clear from the results presented and a valuable contribution to the scientific community.

Despite the implications of patient management associated with 1p19q-codeletion status, the vast majority of prior studies have focused exclusively on IDH-mutation prediction. In 2018, a radiomics-based machine learning algorithm utilized the BraTS portion of TCGA data to predict 1p19q-codeletion vs intact patients and achieved 80% accuracy.[20] The validation set used to assess this accuracy, however, consisted of only 5 subjects. Two other studies since reported a deep learning-based 3- or 5-fold cross-validation with remarkable 93.4% accuracies for the prediction of 1p19q-codeletion in 2019 and 2020.[17,19] However, one of these studies also lacked a separate test set for generating this metric and for both overall accuracy measures included IDH-wildtype tumors in the 1p19q-intact class, artificially boosting baseline accuracy to either 88% or 65% even if all of the 1p19q-codeleted tumors were predicted incorrectly. There was also no specification about whether early stopping was employed, which in our experience, when used in conjunction with cross-validation approaches, results in a >30% drop in accuracy between the validation and test sets. Without reporting an independent test set or at least the loss curves observed during training and validation as shown in Supplementary eFigure 7, it is not possible to assess whether a model would work on unseen data. Although van der Voort et al[48] used the BraTS images of the TCGA dataset as an external test cohort to validate their radiomics-based machine learning model of 1p19q-codeletion classification in low-grade glioma, and demonstrated clinical relevance by comparing model results to the predictions of expert clinicians, they also did not first stratify by IDH before predicting 1p19q mutation status, elevating their accuracy in 1p19q-intact patients by 25% which translated to a 0.73 AUC ROC on the external test set. In contrast, our study is the first analysis that aims to predict 1p19q-codeletion for patients already determined to be IDH-mutant, without first segregating based on tumor grade obtained through pathology. Although the sample size of our IDH-mutated subgroups is still limited and our external test set does not reflect the same distribution of 1p19q-codeleted to intact patients as our internal dataset, it is still the first deep learning study that attempts to validate models incorporating 1p19q-codeletion status within IDH-mutant gliomas on an external, multi-institutional cohort.

Our results from the UCSF test cohort and downstream GradCAM analysis indicated that a deep learning model is in fact learning from signals in tumor regions and it is possible to learn generalizable imaging features when the patient samples are of similar outcome distribution. GradCAMs provide some amount of interpretability of an otherwise "black-box" CNN by displaying a combination of semantically meaningful features in the form of a heatmap, which can be thought of as a map of where the network is looking in the image to draw its predictions. Our analysis included the generation of GradCAM heatmaps for both well- and poorly predicted patients for the best 3-class model, including ADC maps in place of T2-weighted images. The GradCAM heatmaps in Figure 4 provide confidence in our results because the network focuses on the lesion in all correctly predicted tumors, while in the patients who were misclassified, the GradCAM heatmaps show that the network often became confused by other parts of the image outside the lesion boundaries that confound the overall prediction. Although GradCAMs provide insight into where the model is looking, they do not attribute feature importance and have limited spatial resolution based on the size of the final output layer of the chosen model.

In conclusion, we created a model that was able to generalize to unseen data with a promising overall accuracy of 86%, and individual class accuracies of 95% for IDHwt, 90% for IDHmut-intact, and 60% for IDHmut-codel subgroups. From our extensive hyperparameter search during model development, we derived insights that support the use of a network that has been pre-trained on ImageNet for classification tasks combined with a slice-by-slice approach for updating weights in training and a cropping strategy that extends beyond the boundaries of the T2-lesion. Using this framework, we concluded that classifying both IDH and 1p19q mutations together in a single step was advantageous compared to implementing a tiered structure that first predicted IDH-mutation status before 1p19q-codeletion using 2 separate binary models. The addition of ADC increased the generalization capacity of our models regardless of the modeling structure chosen, highlighting the utility of incorporating diffusion-weighted imaging, that more closely reflects underlying tumor biology, in future multisite analyses of molecular subtypes. Although the goal of this study was to utilize routine clinical sequences for this analysis, additional image contrasts, such as rCBV and SWI that have been shown to highlight unique features specific to 1p19q-codeleted tumors, should be incorporated in the future to improve class accuracy of this subgroup. This, along with more data with different acquisition parameters, an evaluation of feature attribution for different images, and integrating newer network architectures and training approaches will be essential for further improving classification accuracy. Although larger studies that focus on accumulating enough IDH-mutant patients are still desperately needed to improve the accuracy of 1p19q-codeleted gliomas for implementation in clinical practice, the insights gleaned from this study will be highly valuable once such datasets become publicly available.

Supplementary Material

Supplementary material is available at *Neuro-Oncology* online.

Keywords

ADC | convolutional neural network | deep learning | diffusion-weighted imaging | glioma subtype

Funding

This work was supported by NIH/NCI grant P01CA118816 and a T32 Ruth L. Kirschstein Institutional National Research Service Award.

Acknowledgments

We would like to acknowledge the patients and their families for their participation in this study and the numerous prior lab members, clinical coordinators, and technologists for their help with scanning, scheduling, and compiling data for this study.

Conflict of interest statement. The authors have no conflicts of interest to report with respect to the content of this manuscript.

Authorship statement. Study design: J.M.L. and J.C. Data collection: M.L. and D.N. Pathology analysis: J.J.P. and A.S. Image processing and analysis: J.C., J.M.L., P.A.L., M.P.O., and T.L.L. Model training and optimization: J.C., Y.I., and V.P. Statistical analysis: J.C., A.M.M., and P.C. Data interpretation: J.M.L., J.C., S.M.C., J.J.P., A.M.M., J.E.V.M., and P.A.L. Manuscript preparation: J.C., J.M.L., A.M.M., J.J.P., S.M.C., P.A.L., and J.E.V.M.

References

1. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* 2016;131(6):803–820.
2. Brat DJ, Aldape K, Colman H, et al. cIMPACT-NOW update 5: recommended grading criteria and terminologies for IDH-mutant astrocytomas. *Acta Neuropathol.* 2020;139(3):603–608.
3. Yeaney GA, Brat DJ. What every neuropathologist needs to know: update on cIMPACT-NOW. *J Neuropathol Exp Neurol.* 2019;78(4):294–296.
4. Louis DN, Ellison DW, Brat DJ, et al. cIMPACT-NOW: a practical summary of diagnostic points from Round 1 updates. *Brain Pathol.* 2019;29(4):469–472.

5. Louis DN, Wesseling P, Aldape K, et al. cIMPACT-NOW update 6: new entity and diagnostic principle recommendations of the cIMPACT-Utrecht meeting on future CNS tumor classification and grading. *Brain Pathol.* 2020;30(4):844–856.
6. van den Bent MJ, Brandes AA, Taphoorn MJ, et al. Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951. *J Clin Oncol.* 2013;31(3):344–350.
7. Cairncross G, Wang M, Shaw E, et al. Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402. *J Clin Oncol.* 2013;31(3):337–343.
8. Nobusawa S, Watanabe T, Kleihues P, Ohgaki H. IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas. *Clin Cancer Res.* 2009;15(19):6002–6007.
9. Eckel-Passow JE, Lachance DH, Molinaro AM, et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *N Engl J Med.* 2015;372(26):2499–2508.
10. Patel SH, Poisson LM, Brat DJ, et al. T2-FLAIR mismatch, an imaging biomarker for IDH and 1p/19q status in lower-grade gliomas: a TCGA/TCIA project. *Clin Cancer Res.* 2017;23(20):6078–6085.
11. Broen MPG, Smits M, Wijnenga MMJ, et al. The T2-FLAIR mismatch sign as an imaging marker for non-enhancing IDH-mutant, 1p/19q-intact lower-grade glioma: a validation study. *Neuro Oncol.* 2018;20(10):1393–1399.
12. Jain R, Johnson DR, Patel SH, et al. "Real world" use of a highly reliable imaging sign: "T2-FLAIR mismatch" for identification of IDH mutant astrocytomas. *Neuro Oncol.* 2020;22(7):936–943.
13. Lee MK, Park JE, Jo Y, Park SY, Kim SJ, Kim HS. Advanced imaging parameters improve the prediction of diffuse lower-grade gliomas subtype, IDH mutant with no 1p19q codeletion: added value to the T2/FLAIR mismatch sign. *Eur Radiol.* 2020;30(2):844–854.
14. Chen L, Voronovich Z, Clark K, et al. Predicting the likelihood of an isocitrate dehydrogenase 1 or 2 mutation in diagnoses of infiltrative glioma. *Neuro Oncol.* 2014;16(11):1478–1483.
15. Park YW, Han K, Ahn SS, et al. Prediction of IDH1-mutation and 1p/19q-codeletion status using preoperative MR imaging phenotypes in lower grade gliomas. *AJNR Am J Neuroradiol.* 2018;39(1):37–42.
16. Bangalore Yogananda CG, Shah BR, Vejdani-Jahromi M, et al. A novel fully automated MRI-based deep-learning method for classification of IDH mutation status in brain gliomas. *Neuro Oncol.* 2020;22(3):402–411.
17. Yogananda CGB, Shah BR, Yu FF, et al. A novel fully automated MRI-based deep-learning method for classification of 1p/19q co-deletion status in brain gliomas. *Neurooncol Adv.* 2020;2(1):vdaa066.
18. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Sci Rep.* 2017;7(1):5467.
19. Chang P, Grinband J, Weinberg BD, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR Am J Neuroradiol.* 2018;39(7):1201–1207.
20. Lu CF, Hsu FT, Hsieh KL, et al. Machine learning-based radiomics for molecular subtyping of gliomas. *Clin Cancer Res.* 2018;24(18):4429–4436.
21. Chang K, Bai HX, Zhou H, et al. Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging. *Clin Cancer Res.* 2018;24(5):1073–1081.
22. Choi YS, Bae S, Chang JH, et al. Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. *Neuro Oncol.* 2021;23(2):304–313.
23. Lasocki A, Anjari M, Örs Kokurcan S, Thust SC. Conventional MRI features of adult diffuse glioma molecular subtypes: a systematic review. *Neuroradiology.* 2021;63(3):353–362.
24. Eichinger P, Alberts E, Delbridge C, et al. Diffusion tensor image features predict IDH genotype in newly diagnosed WHO grade II/III gliomas. *Sci Rep.* 2017;7(1):13396.
25. Cui Y, Ma L, Chen X, Zhang Z, Jiang H, Lin S. Lower apparent diffusion coefficients indicate distinct prognosis in low-grade and high-grade glioma. *J Neurooncol.* 2014;119(2):377–385.
26. Villanueva-Meyer JE, Wood MD, Choi BS, et al. MRI features and IDH mutational status of grade II diffuse gliomas: impact on diagnosis and prognosis. *AJR Am J Roentgenol.* 2018;210(3):621–628.
27. 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*, arXiv:1811.02629v3 , November 5, 2018, preprint: not peer reviewed.
28. 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.
29. 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.
30. Weller M, van den Bent M, Hopkins K, et al.; European Association for Neuro-Oncology (EANO) Task Force on Malignant Glioma. EANO guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. *Lancet Oncol.* 2014;15(9):e395–e403.
31. Heaphy CM, de Wilde RF, Jiao Y, et al. Altered telomeres in tumors with ATRX and DAXX mutations. *Science.* 2011;333(6041):425.
32. Duarte-Carvajalino JM, Sapiro G, Harel N, Lenglet C. A framework for linear and non-linear registration of diffusion-weighted MRIs using angular interpolation. *Front Neurosci.* 2013;7:41.
33. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage.* 2002;17(2):825–841.
34. Johnson HJ, Harris G, Williams K. BRAINSFit: Mutual Information Registrations of Whole-Brain 3D Images, Using the Insight Toolkit. *Insight* J 2007;57(1):1–10.
35. Kikinis R, Pieper SD, Vosburgh KG. 3D slicer: a platform for subject-specific image analysis, visualization, and clinical support. In: Jolesz FA, ed. *Intraoperative Imaging and Image-Guided Therapy.* New York, NY: Springer New York; 2014:277–289.
36. Zhang B, Chang K, Ramkissoon S, et al. Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro Oncol.* 2017;19(1):109–117.
37. Sajedi H, Pardakhti N. Age prediction based on brain MRI image: a survey. *J Med Syst.* 2019;43(8):279.
38. Pedregosa F, Varoquaux F, Gramfort G, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–2830.
39. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*, arXiv:1409.1556v6, September 4, 2014, preprint: not peer reviewed.
40. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20–25, 2009; Miami, FL, USA: 248–255.
41. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27–30, 2016; Las Vegas, NV, USA: 770–778.
42. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med.* 2018;15(11):e1002699.
43. Smith LN. A disciplined approach to neural network hyper-parameters: part 1 – learning rate, batch size, momentum, and weight decay. *arXiv*, arXiv:1803.09820v2, March 26, 2018, preprint: not peer reviewed.
44. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv*, arXiv:1412.6980v9, December 22, 2014, preprint: not peer reviewed.

45. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. arXiv, arXiv:1810.03292v3, October 8, 2018, preprint: not peer reviewed.

46. Cole JH, Poudel RPK, Tsagkrasoulis D, et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage.* 2017;163:115–124.

47. Qi S, Yu L, Li H, et al. Isocitrate dehydrogenase mutation is associated with tumor location and magnetic resonance imaging characteristics in astrocytic neoplasms. *Oncol Lett.* 2014;7(6):1895–1902.

48. van der Voort SR, Incekara F, Wijnenga MMJ, et al. Predicting the 1p/19q codeletion status of presumed low-grade glioma with an externally validated machine learning algorithm. *Clin Cancer Res.* 2019;25(24):7455–7462.