Neuro-Oncology
23(2), 304–313, 2021 | doi:10.1093/neuonc/noaa177 | Advance Access date 24 July 2020

Fully automated hybrid approach to predict the *IDH* mutation status of gliomas via deep learning and radiomics

Yoon Seong Choi, Sohi Bae, Jong Hee Chang, Seok-Gu Kang, Se Hoon Kim, Jinna Kim, Tyler Hyungtaek Rim, Seung Hong Choi, Rajan Jain, and Seung-Koo Lee

Duke-NUS Medical School, RADSC ACP, Singapore (Y.S.C.); Department of Diagnostic Radiology, Singapore General Hospital, Singapore (Y.S.C.); Department of Radiology and Research Institute of Radiological Science, Yonsei University College of Medicine, Seoul, Korea (Y.S.C., S-K.L., J.K.); Department of Radiology, National Health Insurance Service Ilsan Hospital, Goyang, Korea (S.B.); Department of Neurosurgery, Yonsei University College of Medicine, Seoul, Korea (J.H.C., S-G.K.); Department of Pathology, Yonsei University College of Medicine, Seoul, Korea (S.H.K.); Singapore Eye Research Institute, Singapore National Eye Centre, Duke-NUS Medical School, Singapore (T.H.R.); Department of Radiology, Seoul National University College of Medicine, Seoul, Korea (S.H.C.); Department of Radiology, New York University School of Medicine, New York, New York, USA (R.J.); Department of Neurosurgery, New York University School of Medicine, New York, New York, USA (R.J.)

Corresponding Author: Yoon Seong Choi, MD, PhD, Department of Diagnostic Radiology, Singapore General Hospital, Singapore, Outram Rd, Singapore 169608 (yoonseong.choi07@gmail.com).

Abstract

Background. Glioma prognosis depends on isocitrate dehydrogenase (*IDH*) mutation status. We aimed to predict the *IDH* status of gliomas from preoperative MR images using a fully automated hybrid approach with convolutional neural networks (CNNs) and radiomics.

Methods. We reviewed 1166 preoperative MR images of gliomas (grades II–IV) from Severance Hospital ($n = 856$), Seoul National University Hospital (SNUH; $n = 107$), and The Cancer Imaging Archive (TCIA; $n = 203$). The Severance set was subdivided into the development ($n = 727$) and internal test ($n = 129$) sets. Based on T1 postcontrast, T2, and fluid-attenuated inversion recovery images, a fully automated model was developed that comprised a CNN for tumor segmentation (Model 1) and CNN-based classifier for *IDH* status prediction (Model 2) that uses a hybrid approach based on 2D tumor images and radiomic features from 3D tumor shape and loci guided by Model 1. The trained model was tested on internal (a subset of the Severance set) and external (SNUH and TCIA) test sets.

Results. The CNN for tumor segmentation (Model 1) achieved a dice coefficient of 0.86–0.92 across datasets. Our hybrid model achieved accuracies of 93.8%, 87.9%, and 78.8%, with areas under the receiver operating characteristic curves of 0.96, 0.94, and 0.86 and areas under the precision-recall curves of 0.88, 0.82, and 0.81 in the internal test, SNUH, and TCIA sets, respectively.

Conclusions. Our fully automated hybrid model demonstrated the potential to be a highly reproducible and generalizable tool across different datasets for the noninvasive prediction of the *IDH* status of gliomas.

Key Points

1. Prognosis of gliomas depends on *IDH* mutation status.
2. Our hybrid model is based on convolutional neural networks that integrate shape and loci radiomics.
3. This fully automated hybrid model can predict *IDH* status across 3 datasets.

Importance of the Study

CNNs and radiomics have shown potential to be used for the noninvasive assessment of *IDH* mutation status. Herein, based on conventional MR images from 1166 patients with gliomas, we developed a fully automated hybrid model based on a CNN that integrates 2-dimensional tumor signal intensity and radiomic features from 3D tumor shape and loci. We tested the model on 3 datasets, achieving accuracies of 78.8%–93.8% and areas under the receiver operating characteristics curves of 0.86–0.96. Our results (i) show that the hybrid approach allows for the accurate prediction of *IDH* status; (ii) demonstrate the potential to use this automated process as a highly reproducible and generalizable method for noninvasive characterization of gliomas; and (iii) implicate the clinical consideration regarding the similarity between training and test data that is required for ideal generalizability of model performance.

Glioma is the most common malignant primary brain tumor in adults.[1] Gliomas have a wide range of prognoses depending on the World Health Organization (WHO) grade, with a median survival of 14 months for glioblastomas (grade IV)[2] and of more than 7 years for lower grade gliomas (grades II and III).[3] Recently, molecular subtypes, such as isocitrate dehydrogenase (*IDH*) mutation status, have been reported as important factors for the tumor behavior of gliomas.[4] Specifically, lower grade gliomas with wildtype *IDH* were reported to be similar to glioblastomas in terms of molecular profile and prognosis, while *IDH* mutated glioblastomas showed a better prognosis than *IDH* wildtype glioblastomas.[5,6] Moreover, anaplastic gliomas (grade III) with wildtype *IDH* have a worse prognosis than glioblastomas with an *IDH* mutation (grade IV).[7] As such, *IDH* status has been integrated into the 2016 WHO classification scheme for gliomas.[8] Additionally, the effect of gross total tumor resection on the prognosis of lower grade gliomas was reported to depend on the *IDH* mutation status.[9] Therefore, the preoperative prediction of *IDH* status is necessary for appropriate treatment planning.

Deep convolutional neural networks (CNNs) are a representative method to exploit high-dimensional numeric information from images by learning relevant features directly from image signal intensities. CNNs have diagnostic value for predicting the *IDH* status of gliomas.[10–13] However, there are several hurdles for clinical implementation of CNNs for *IDH* status prediction; if the model prediction process is not fully automated, any operator-dependent process such as manual segmentation can be time-consuming, which limits clinical feasibility and is a source of interrater variability. Moreover, image signal intensity–based CNNs are unable to directly incorporate information from 3D tumor shapes and locations that are associated with *IDH* status.[10,11] Additionally, since image signal intensities are likely to be sensitive to different MRI protocols and machines and because they may affect the performance of CNN models, external testing is necessary to confirm model generalizability across different institutions. Although previous studies have applied CNNs to brain MR images for the prediction of molecular profiles of gliomas,[10–12,14] many studies lacked external testing for confirming model generalizability.[10,14] To mitigate these limitations, in this study, model testing for generalizability was conducted on 3 datasets, including 2 external test sets, after model training on a dataset from one institution.

The aim of this study was to predict the *IDH* status of patients with gliomas (grades II-IV) from preoperative MR images using a fully automated hybrid approach that integrates the following: (i) a CNN for automated tumor segmentation and (ii) a CNN-based classifier for *IDH* status prediction that incorporated both 2D images and the radiomic features of 3D tumor shape and loci.

Materials and Methods

This retrospective study was approved by the institutional review board of Severance Hospital, Seoul, South Korea. The requirement for informed consent was waived.

Patients

The patient enrollment process is shown in Fig. 1. A total of 1202 patients who underwent preoperative MRI for newly diagnosed gliomas (grades II–IV) from January 2006 to June 2019 at Severance Hospital were considered for inclusion. The inclusion criteria were as follows: (i) pathologically confirmed glioma, (ii) known *IDH* mutation status, (iii) preoperative MRI inclusive of postcontrast T1-weighted (T1C), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR) images, and (iv) age ≥18 years. The exclusion criteria were as follows: (i) history of biopsy or surgery for brain tumor (n = 182), (ii) the absence of T1C, T2, or FLAIR images (n = 83), and/or (iii) unknown *IDH* status (n = 81). Therefore, a total of 856 patients were enrolled from Severance Hospital (Severance set). These patients were semi-randomly allocated into development (n = 727) and internal test sets (n = 129), with stratification for *IDH* status. The development set was subsequently divided into the training (n = 596) and tuning (n = 131) sets. For external testing, a total of 107 consecutive patients from January 2017 to January 2018 from the Seoul National University Hospital (SNUH set)[15] and 203 patients from The Cancer Imaging Archive (TCIA set)[16] were enrolled in accordance with the same criteria. The list of enrolled patients from the set from TCIA is shown in eTable 1. The details of *IDH* status evaluation are provided in eDocument 1.

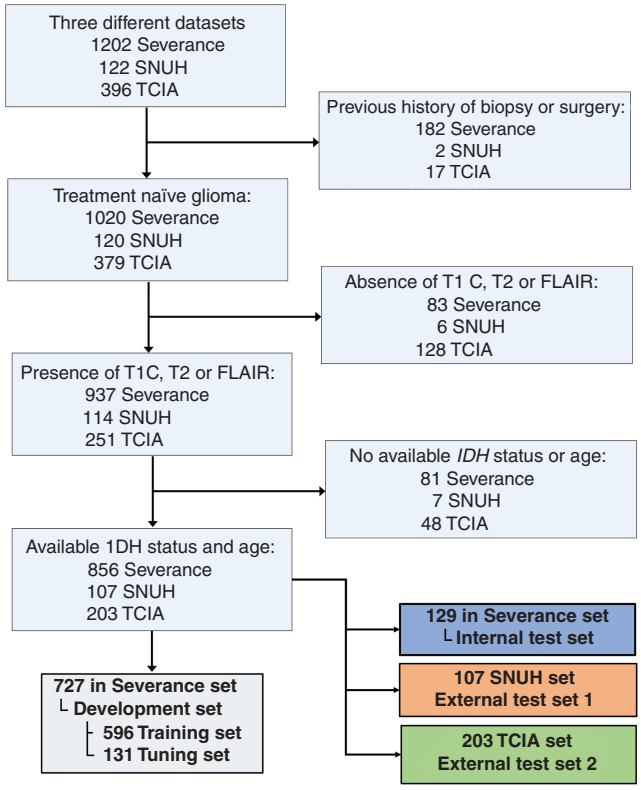

Fig. 1 Flow diagram of the study population.

Study Design

The study design is summarized in Fig. 2. Our automated hybrid model comprised 2 deep CNN models (Models 1 and 2) and a fully automated pipeline between the 2 models. A CNN for automatic tumor segmentation (Model 1) was trained to yield tumor masks based on T1C and FLAIR images, and a CNN-based binary classifier for *IDH* status (Model 2) was trained to predict the *IDH* status based on (i) the image inputs of T1C and T2 images and tumor masks from Model 1, (ii) the numeric inputs consisting of radiomic features from the tumor shape and loci extracted from the tumor masks of Model 1, and (iii) age. Model 1 and Model 2 constituted the hybrid model using the fully automated pipeline between the two models. The CNN models were developed based on the Severance development set and tested on 1 internal test set (a subset of the Severance set) and 2 external test sets (SNUH and TCIA sets).

Image Acquisition and Processing

The details of the image acquisition parameters are summarized in eFigure 1 and eTable 2. The most commonly used imaging parameters and their proportions in each dataset are as follows; magnetic field strength: 3 T in the Severance (98.2%) and SNUH (85.0%) sets and 1.5 T in the set from TCIA (55.7%); manufacturer: Philips in the Severance set (98.1%), Siemens in the SNUH set (79.4%), and General Electric in the set from TCIA (73.4%); T1C slice thickness: ≤1.0 mm in the Severance (83.1%) and SNUH (100.0%) sets and 2.0–3.0 mm in the set from TCIA (36.5%); T2 slice thickness: 6.0–7.0 mm in the Severance (45.9%) and SNUH (50.5%) sets and 4.0–5.0 mm in the set from TCIA (32.5%); FLAIR slice thickness: 6.0–7.0 mm in the Severance (54.2%) and SNUH (88.8%) sets and 2.0–3.0 mm in the set from TCIA (42.4%). The detailed image processing methods are described in eDocument 2. Briefly, T1C, T2, and FLAIR images were registered to an identical 1-mm isovoxel spatial coordinate. The images were subjected to signal intensity normalization and resampling to sizes of 128 × 128 × 128. The ground-truth whole tumor was defined as high-signal intensity on FLAIR images and was segmented by a neuroradiologist and confirmed by another neuroradiologist (S.B. and Y.S.C., with 4 and 7 years of experience in neuroradiology, respectively).

CNN for Tumor Segmentation (Model 1)

For the CNN for tumor segmentation, the modified 3D U-shaped CNN architecture proposed by Kickingereder and Isensee et al[17] was revised to use T1C and FLAIR images with sizes of 128 × 128 × 128 as network inputs and yield whole tumor segmentation only. The details of the architecture and training process are described in eDocument 3.

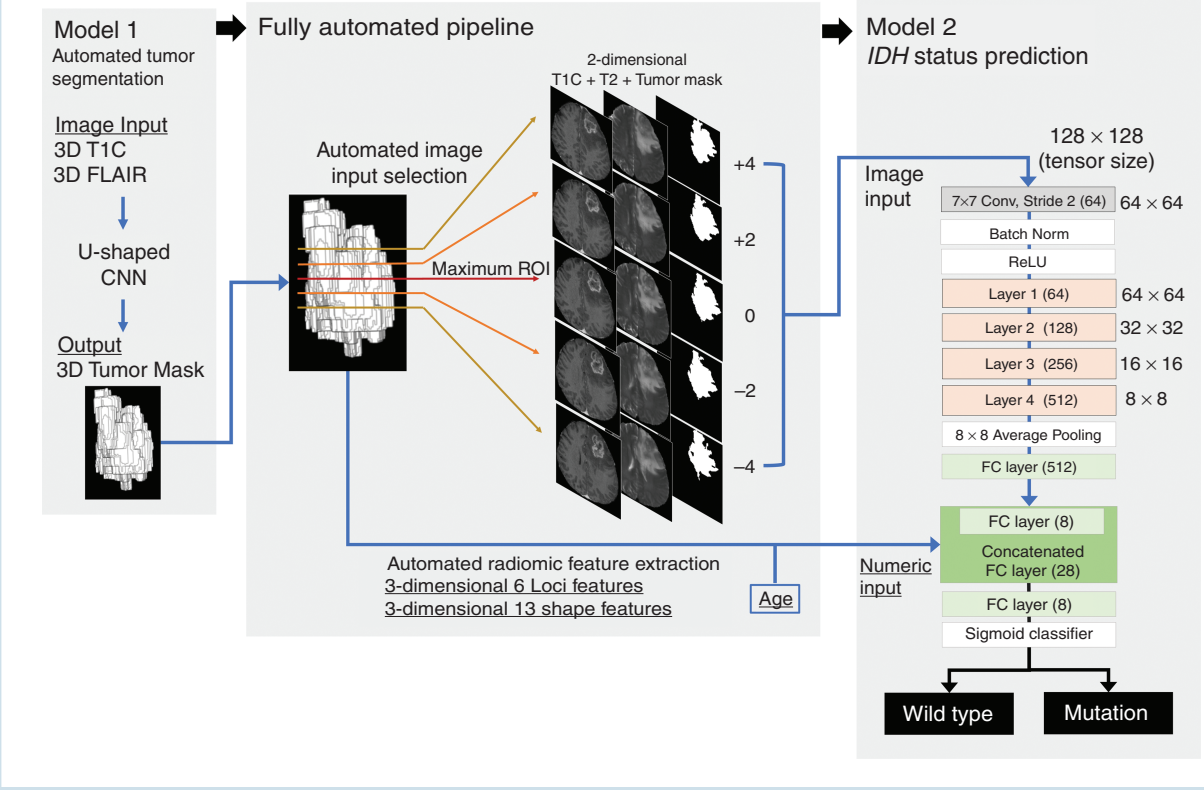

Fig. 2 Fully automated hybrid model for IDH status prediction. In Model 2, layers 1–4 consisted of 3, 4, 6, and 3 residual blocks, with each block containing the 3 × 3 convolutions twice.

CNN Classifier for IDH Status Prediction (Model 2)

The Model 2 architecture is shown in Fig. 2. Briefly, our CNN classifier for *IDH* status prediction was derived from the well-known 34-layer ResNet architecture[18] (hereinafter referred as the conventional ResNet) shown in eFigure 2 that contains the initial 7 × 7 convolution and layers 1–4 comprising 3, 4, 6, and 3 residual blocks with each residual block having the 3 × 3 convolution twice. For the hybrid approach, additional fully connected layers were added to the conventional ResNet to build Model 2, which used the additional numeric inputs along with the image inputs.

The Model 2 image input comprised axial T1C and T2 images and tumor masks of 128 × 128 size. To extract comprehensive 2D signal intensity information from the tumor, the axial slice with the maximum tumor area was automatically selected as the "maximum tumor image," and the other 4 images were extracted from 4 upper (+4), 2 upper (+2), 2 lower (–2), and 4 lower (–4) slices from the maximum tumor images. Thus, 5 axial slices per patient were automatically selected based on tumor segmentation and were considered as individual samples in model development and testing. The Model 2 numeric inputs were selected among a total of 24 features consisting of age and shape- and loci-based radiomic features (eDocument 4) that were automatically extracted from the tumor masks, according to the univariate *t*-test in terms of *IDH* status in the development set.

The details of the Model 2 training process are described in eDocument 5. Briefly, the conventional ResNet was first trained using the image inputs as a "warm-up training." Then, the pretrained weights from the layers that were close to image inputs were imported to Model 2 and fixed, and the rest of the layers of Model 2 were fine-tuned using both image and numeric inputs.

Fully Automated Pipeline and Hybrid Model

Our automated hybrid model was built by connecting Model 1 for segmentation with Model 2 for *IDH* status prediction using the fully automated pipeline. Our trained automated hybrid model was tested in 1 internal test set (a subset of the Severance set) and 2 external test sets (the SNUH and TCIA sets). The internal test set, as well as the SNUH and TCIA sets, were separated and not revealed during model development. The first step of automated hybrid model testing was to obtain automatic tumor segmentation of the test samples from Model 1. These segmentations were subsequently used for the selection of the 2D image inputs (5 axial slices) of Model 2 and the extraction of 3D shape- and loci-based radiomic features that were used along with age as the numeric inputs of Model 2. All processes were automated, and the trained models and code for image processing are available at the following link: https://github.com/yoonchoi-neuro/automated_hybrid_IDH.

Model Explanation

To understand which part of the image inputs are relevant for *IDH* status prediction, ablation analysis was conducted[19] and saliency maps were generated, which were assessed by 2 neuroradiologists (S.B. and J.K.) based on a survey,[20] as described in eDocument 6. The variable importance of numeric inputs was calculated based on the shape/loci radiomics classifier that is mentioned later, using the "varImp" function of "caret" R package.

Statistical Analysis

The performance of Model 1 for tumor segmentation was measured using the dice similarity coefficient that measures the extent of spatial overlap between 2 binary segmentation masks (ie, ground-truth segmentation from radiologists and automatic segmentation from Model 1) and ranges from 0 (no overlap) to 1 (perfect agreement).[21] The diagnostic performance of the automated hybrid model that connected Models 1 and 2 was measured in terms of accuracy, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC), using the "PRROC" R package. The 95% confidence intervals (CIs) of the AUROC and AUPRC values were calculated from 2000 iterations of bootstrapping with the predicted probabilities from the models. The probability threshold for the accuracy calculation was set to 0.5; thus, a predicted probability of ≥0.5 was classified as an *IDH* mutation, and other values were classified as *IDH* wildtype. The diagnostic accuracy of the automated hybrid model was measured for both individual samples and the mean probability from 5 samples per patient. Beside the automated hybrid model and the conventional ResNet, to assess the diagnostic value of the numeric features, the shape/loci radiomics classifier was built based on the numeric inputs of Model 2 using the random forest algorithm and 10-fold cross-validation within the development set using the "caret" R package.[22] Additionally, the diagnostic model that used age alone was built using logistic regression on the development set. The diagnostic performance on the test sets was measured for the conventional ResNet with image inputs, the shape/loci radiomics classifier with numeric inputs, and the prediction using age alone. To evaluate whether the image-integrated model (ie, the automated hybrid model) predicts *IDH* status more accurately than clinical factors alone (ie, age), accuracy and AUROCs were compared via an exact binomial test (with the "binom.test" R function) and a method described by Delong et al.[23] A *P*-value of <0.05 was considered to indicate a statistically significant difference. All statistical analyses were conducted using the R software, v3.4.4.

Results

Characteristics of the Study Population

The clinical characteristics of a total of 1166 patients from the Severance Hospital, TCIA, and SNUH are summarized in Table 1. In terms of *IDH* status, the SNUH set showed no difference compared with the Severance set ($P > 0.999$), whereas the set from TCIA had a significantly higher prevalence of *IDH* mutation than the Severance set ($P < 0.001$). In terms of WHO grade, the SNUH set had a significantly higher proportion of grade II gliomas ($P = 0.005$) and the set from TCIA showed a nonsignificant tendency toward a higher proportion of lower grade gliomas ($P = 0.054$) than the Severance set. No significant difference was found in terms of age ($P = 0.194$ for the Severance vs SNUH sets; and $P = 0.818$ for the Severance vs TCIA sets) or sex ($P = 0.721$ for the Severance vs SNUH sets; and $P = 0.305$ for the Severance vs TCIA sets).

Model Performance

Model 1 (CNN for tumor segmentation) was achieved through 46 epochs of training and yielded dice coefficients of 0.91 ± 0.04, 0.92 ± 0.01, and 0.86 ± 0.08 in the internal test, SNUH, and TCIA sets, respectively. The pretrained conventional ResNet with image inputs was achieved through 111 epochs of training. Among the 24 numeric features of shape and loci features and age, 20 features showed significant differences according to *IDH* status and were used as the numeric inputs of Model 2. After a part of the weights were imported from the pretrained conventional ResNet, Model 2 was achieved through 72 epochs of fine-tuning. The performance of the automated hybrid model on test sets is summarized in Table 2 and Fig. 3. With the mean probabilities from 5 samples per patient, our model achieved accuracies of 93.8%, 87.9%, and 78.8%, with AUROCs of 0.96 (95% CI: 0.93–0.99), 0.94 (95% CI: 0.89–0.97), and 0.86 (95% CI: 0.80–0.91) and AUPRCs of 0.88 (95% CI: 0.72–0.98), 0.82 (95% CI: 0.65–0.94), and 0.81 (95% CI: 0.71–0.88) in the internal test, SNUH, and TCIA sets, respectively.

The performances of the conventional ResNet, shape/loci radiomics classifier, and the prediction using age alone are summarized in Table 3. The conventional ResNet, shape/loci radiomics classifier, and prediction using age alone achieved accuracies of 73.5%–92.2%, 75.4%–85.3%, and 68.5%–72.1%; AUROCs of 0.81–0.95, 0.84–0.90, and 0.74–0.81; and AUPRCs of 0.74–0.87, 0.65–0.85, and 0.44–0.68, respectively, in the test sets. The automated hybrid model was superior to prediction using age alone across all datasets (all $P < 0.05$ for accuracy and AUROC, eTable 3).

Model Explanation

The ablation analysis results are shown in eTable 4. Compared with the performance of the original conventional ResNet, in the ablation analysis with a decreased number of axial images per patient, decreased number of image sequences, or masking of the nontumor brain tissue area, we observed decreased diagnostic performance, with the accuracy, AUROC, and AUPRC in the ranges of 67.8%–86.8%, 0.71–0.93, and 0.62–0.84, respectively. When the tumor area was masked from image inputs, the lowest diagnostic performance was yielded with accuracies of 58.4%–72.1%, AUROCs of 0.55–0.71, and AUPRCs of 0.37–0.47.

Table 1 Patient characteristics

	Severance Set (*n* = 856)	SNUH Set (*n* = 107)	*P*-value (Severance vs SNUH)	TCIA Set (*n* = 203)	*P*-value (Severance vs TCIA)
Age	52.9 ± 15.4	50.9 ± 15.5	0.194	52.7 ± 14.9	0.818
Sex			0.721		0.305
Male	492 (57.5)	64 (59.8)		108 (53.2)	
Female	364 (42.5)	43 (40.2)		95 (46.8)	
***IDH* status**			>0.999		<0.001
Wildtype	617 (72.1)	77 (72.0)		119 (58.6)	
Mutation	239 (27.9)	30 (28.0)		84 (41.4)	
WHO grade			0.005[b]		0.054[b]
II	175 (20.5)	14 (13.1)		48 (23.7)	
IDH wildtype [a]	37 (21.1)	5 (35.7)		3 (6.2)	
IDH mutation + 1p/19 non-codeletion	80 (45.7)	1 (7.1)		28 (58.3)	
IDH mutation + 1p/19 codeletion	58 (33.2)	8 (57.2)		17 (35.5)	
III	169 (19.7)	35 (32.7)		52 (25.6)	
IDH wildtype	89 (52.7)	15 (42.9)		17 (32.7)	
IDH mutation + 1p/19 non-codeletion	32 (18.9)	12 (34.3)		24 (46.2)	
IDH mutation + 1p/19 codeletion	48 (28.4)	8 (22.8)		11 (21.1)	
IV	512 (59.8)	58 (54.2)		103 (50.7)	
IDH wildtype	491 (95.9)	57 (98.3)		99 (96.1)	
IDH mutation + 1p/19 non-codeletion	16 (3.1)	1 (1.7)		3 (2.9)	
IDH mutation + 1p/19 codeletion	4 (0.8)	0 (0.0)		1 (1.0)	
IDH mutation + 1p/19q status not specified	1 (0.2)	0 (0.0)		0 (0.0)	

[a] Number in parentheses represents the proportion of each molecular subgroup within a specific WHO grade.
[b] *P*-value from the chi-square test to compare WHO grade constituents between the two cohorts.

The saliency map survey results and representative cases are shown in eTable 5 and eFigure 3. On the T1C saliency maps, the main activation areas were the enhancing tumor areas if the tumor had enhancing portions and the tumor periphery along the margin, regardless of the presence of enhancement. On T2 saliency maps, the main activation area involved the entire tumor area. Additionally, outside the tumors, the peritumoral area along the tumor margin was included in the activation area on both the T1C and T2 saliency maps. The variable importance of the numeric features and their different distributions according to *IDH* status are shown in eFigures 4 and 5. Among the 20 numeric features, age, frontal lobe location, tumor sphericity, parietal lobe location, and thalamus location were the most important features.

Discussion

We included a total of 1166 patients with gliomas and developed an automated hybrid model to predict *IDH* status. Our model incorporated information from 2D tumor signal intensity, 3D tumor shape and location, and age into one CNN along with CNN-based automated tumor segmentation and a fully automated pipeline without any operator-dependent processes. After developing the model based on one institutional dataset, our automated hybrid model allowed for *IDH* status prediction across different cohorts, MR scanners, and imaging protocols, with AUROCs ranging from 0.86 to 0.96.

An *IDH* gene mutation confers a better prognosis and treatment response of gliomas, independent of the histologic grade,[5,7,24] which led to the integration of *IDH* status for glioma classification in the 2016 update to the WHO classification of tumors of the central nervous system.[8] Recent studies have shown that gross total resection is more beneficial for *IDH*-mutant astrocytomas than gliomas with other molecular subtypes.[9] Although maximum tumor resection is the standard treatment regardless of *IDH* status, preoperative prediction of *IDH* status may potentially help in planning for treatments including surgery. This led to the investigation of radiological findings to predict *IDH* status.[25,26] However, such visually assessed qualitative radiological findings are predisposed to interobserver variability and are only able to provide information that is perceivable by visual assessment by humans; a previous study conducted a visual assessment according to the Visually Accessible Rembrandt Images (VASARI) annotations and texture analyses on MR images

Table 2 Diagnostic performance of the hybrid model for the prediction of *IDH* status

Dataset	Accuracy	AUROC (95% CI)			AUPRC (95% CI)		
Per slice							
Internal test set	90.50%	0.96	0.95	0.97	0.88	0.82	0.93
SNUH set	84.90%	0.93	0.9	0.95	0.81	0.74	0.87
TCIA set	77.10%	0.84	0.81	0.86	0.79	0.74	0.83
Per patient[a]							
Internal test set	93.80%	0.96	0.93	0.99	0.88	0.72	0.98
SNUH set	87.90%	0.94	0.89	0.97	0.82	0.65	0.94
TCIA set	78.80%	0.86	0.8	0.91	0.81	0.71	0.88

AUROC, area under the receiver operating characteristics curve; AUPRC, area under the precision-recall curve.
[a]Since each patient yielded 5 tumor slices, the diagnostic accuracy per patient was calculated from the mean value of the 5 predicted probabilities per patient.

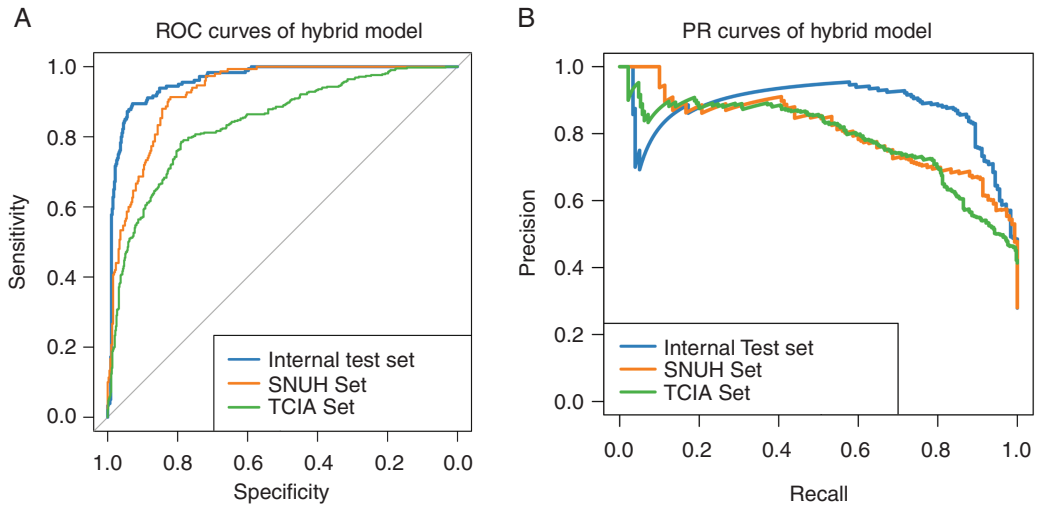

Fig. 3 Performance of the hybrid model in the prediction of IDH status. ROC, receiver operating characteristic; PR, precision-recall.

of lower grade gliomas and reported that the diagnostic performance of qualitative VASARI features (AUROC mean, 0.73 ± 0.02) was lower than that of quantitative texture analysis (AUROC mean, 0.86 ± 0.01).[26]

CNNs and radiomics analysis are representative quantitative methods for image analysis and are capable of extracting high-dimensional and abstract numeric information beyond what is perceivable via the visual assessment of a given image. Although CNNs and radiomics analysis have shown excellent performance for the prediction of *IDH* status,[10–13] both CNN- and radiomics-based classifiers have major obstacles when it comes to clinical implementation. First, robust tumor segmentation is a major challenge for both CNN-based and radiomics classifiers. Although semi-automatic segmentation has shown greater reproducibility than manual segmentation,[27] automatic segmentation is still mandatory to achieve the ideal reproducibility. Second, for radiomic features, a major cause of limited reproducibility is the lack of a standard method for the computation of intensity features (ie, first order and texture features), in terms of the ranges of intensity and the number of bins to discretize intensities.[28–30] In contrast to intensity features, shape and loci features are independent from parameters used for intensity feature computation; thus, shape and loci features can be stable if robust tumor segmentation can be achieved. Third, although CNNs eliminate the steps of feature computation and selection by using convolutions to capture the key features directly from the images,[31] signal intensity–based CNNs are not able to directly capture the 3D shape and location of tumors that are reported to be distinct depending on *IDH* status.[26,32–34]

A few studies have applied CNNs for *IDH* status prediction.[10–12] One study with a total of 259 patients from a TCIA set developed a CNN model that used 2D image inputs selected using a pretrained algorithm for tumor segmentation.[10] Another study with 214 patients from a TCIA set developed a fully automated network that performs

Table 3 Diagnostic performance of the conventional ResNet, shape/loci radiomic classifier, and age in the prediction of the *IDH* status

	Dataset	Accuracy	AUROC (95% CI)			AUPRC (95% CI)		
Conventional ResNet (per slice)	Internal test set	88.70%	0.94	0.92	0.96	0.86	0.8	0.91
	SNUH set	81.50%	0.9	0.87	0.92	0.79	0.73	0.85
	TCIA set	74.40%	0.79	0.77	0.82	0.72	0.68	0.77
Conventional ResNet (per patient)[a]	Internal test set	92.20%	0.95	0.9	0.98	0.87	0.72	0.97
	SNUH set	84.10%	0.91	0.84	0.96	0.81	0.67	0.91
	TCIA set	73.50%	0.81	0.74	0.87	0.74	0.63	0.83
Shape/loci radiomics classifier	Internal test set	85.30%	0.9	0.83	0.96	0.85	0.74	0.93
	SNUH set	79.40%	0.87	0.8	0.93	0.65	0.48	0.85
	TCIA set	75.40%	0.84	0.78	0.9	0.77	0.67	0.86
Age	Internal test set	72.10%	0.74	0.63	0.84	0.44	0.32	0.59
	SNUH set	74.80%	0.81	0.71	0.89	0.51	0.36	0.72
	TCIA set	68.50%	0.77	0.7	0.84	0.68	0.57	0.8

AUROC, area under the receiver operating characteristics curve; AUPRC, area under the precision-recall curve.
[a]Since each patient yielded 5 tumor slices, the diagnostic accuracy per patient was calculated from the mean value of the 5 predicted probabilities per patient.

tumor segmentation and *IDH* status prediction simultaneously, based on whole brain 3D T2 images.[12] However, these 2 studies performed cross-validation only to test model generalizability without performing external testing on a separate dataset.[12] Another study enrolled a total of 496 multi-institutional patients from 3 different datasets to develop a CNN model with 2D image inputs, and it is the only study that conducted external testing for model generalizability, to the best of our knowledge.[11] Nonetheless, the authors of this study used manual tumor segmentation to select the tumor slice for network input and did not confirm the improved diagnostic performance of the CNN-based model compared with that of age alone; the accuracies in the independent external testing ranged 67.1–79.0%, 77.5–84.5%, and 77.7–84.1% for prediction by the CNN model only, CNN model combined with age, and age alone, respectively.[11] To the best of our knowledge, a fully automated model with its generalizability tested on multiple tests with a varying degree of similarity to the development set has not been well established. Given that not many institutions can be sufficiently equipped with a large patient cohort and technical infrastructure that are required for deep learning model development, importing the externally trained model from a large institution may be a more viable option than participating in multi-institutional model development from the beginning. Hence, the practical model should be generalizable across institutions without being trained on samples from those institutions and be reproducible via automation. In this study, we attempted to simulate the feasible clinical scenario that a fully automated model can be developed at a large institution and applied to datasets from various institutions.

Model 2 of our hybrid model is different from the previously reported image-based CNNs[10–12] based on the fact that 3D tumor shape, loci, and age were integrated into one CNN that yielded constantly better model performance across datasets than the image only–based CNN. The analysis results for model explanation implicate that the relevant features for our model prediction are in line with those in previous reports. The average age of patients with *IDH*-mutant gliomas is several years lower than that of patients with *IDH* wildtype gliomas.[5,26,35] *IDH* mutant gliomas occupy the frontal lobe,[34] whereas their *IDH* wildtype counterparts frequently occupy the parietal lobe with little frontal lobe involvement.[32] *IDH* wildtype gliomas have a higher proportion of enhancement and more irregular tumor boundaries than *IDH* mutant gliomas.[26,33] Apparent diffusion coefficient as a tumor cellularity index was reported to be distinct in terms of *IDH* status, and T2-weighted signal intensity was reported to be related to tumor cellularity.[36–38]

Our results imply that our automated hybrid model can be generalized across different MRI scanners, image protocols, and cohorts, and it consistently yielded better performance than the use of age alone. Nevertheless, it is noteworthy that our hybrid model performance varied depending on test sets; it was highest in the internal test set, followed by the SNUH set and the set from TCIA. Regarding the most commonly used imaging parameters, the SNUH set used the same 3T magnetic strength and image spatial resolutions that were similar or higher compared with those of the Severance set. Contrarily, in the set from TCIA, the magnetic strength was 1.5 T, and some image spatial resolutions were lower than that of the Severance set, including the slice thickness of T1C that was used as a registration template in imaging processing. Thus, the SNUH set represents a dataset that is similar to the development set regarding imaging protocol, *IDH* mutation proportion, and patient ethnicity (homogeneously Korean), whereas the set from TCIA represents an extremely heterogeneous dataset that consists of patients of multiple ethnicities from multiple institutions with different imaging protocols and proportions of *IDH* mutation compared with the development set. Thus, our results imply that the degree of similarity between the development and test sets, regarding cohorts and imaging protocols, should be

considered for excellent model performance. Given the good model performance in the SNUH set with a difference in scanners, our results also imply that if cohort characteristics are similar, the scanner-related difference can be substantially overcome by the standardization of image protocols across institutions to enhance model generalizability.[39]

This study has 3 major limitations that merit discussion. First, advanced MRI techniques such as perfusion and diffusion weighted imaging were not considered; however, our goal was to develop a feasible model based on conventional MR images that are widely available, and using advanced MRI may limit model feasibility. Second, only whole tumor segmentation was used to extract shape and loci features, without separate consideration for contrast enhancing tumors and necrosis. However, multiple segmentations from one tumor may yield redundant tumor shape and loci information that do not further enhance the performance of the model. Third, although Model 2 incorporated the tumor image, shape and loci, and age into one CNN, our model is not an end-to-end model; Model 1 for tumor segmentation and Model 2 for *IDH* status prediction were separately trained and combined afterward. However, image preprocessing, Models 1–2, and the pipeline in between are completely based on open-source modules that can be integrated into a Python-based pipeline, and a test sample can be automatically run through each of the steps of our model. Thus, our approach to combine the 2 models is unlikely to be a barrier to clinical feasibility and convenience. Moreover, by separately developing Models 1–2, each model may have the flexibility to be used as an independent building block in the partial absence of the required MRI sequences or for other applications in neuro-oncology.

In conclusion, we developed a model from deep learning and radiomics that can reliably predict the *IDH* status of gliomas using a fully automated process based on conventional MR imaging. Our model has the potential to be used more widely as a practical tool with high reproducibility and generalizability for the noninvasive characterization of gliomas to support individualized treatment planning.

Supplementary Material

Supplementary data are available at *Neuro-Oncology* online.

Keywords

convolutional neural network | glioma | isocitrate dehydrogenase mutation | magnetic resonance imaging | radiomics

Funding

None.

Acknowledgments

We thank PhD candidate Yohan Jun and Professor Dosik Hwang for their help in searching for useful resources to build CNN models.

Authorship statement. Conception and design: Yoon Seong Choi. Collection and assembly of data: Yoon Seong Choi, Sohi Bae, Seung Hong Choi, Jong Hee Chang, Seok-Gu Kang, Se Hoon Kim, Rajan Jain. Data analysis and interpretation: Yoon Seong Choi, Tyler Hyungtaek Rim. Manuscript writing: Yoon Seong Choi, Tyler Hyungtaek Rim. Final approval of the manuscript: Yoon Seong Choi, Sohi Bae, Jong Hee Chang, Seok-Gu Kang, Se Hoon Kim, Jinna Kim, Tyler Hyungtaek Rim, Seung Hong Choi, Rajan Jain, Seung-Koo Lee. Accountable for all aspects of the work: Yoon Seong Choi.

Conflict of interest statement. None.

References

1. Ostrom QT, Gittleman H, Xu J, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2009-2013. *Neuro Oncol.* 2016;18(suppl_5):v1–v75.
2. Stupp R, Hegi ME, Mason WP, et al; European Organisation for Research and Treatment of Cancer Brain Tumour and Radiation Oncology Groups; National Cancer Institute of Canada Clinical Trials Group. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* 2009;10(5):459–466.
3. Suzuki H, Aoki K, Chiba K, et al. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat Genet.* 2015;47(5):458–468.
4. Riemenschneider MJ, Jeuken JW, Wesseling P, Reifenberger G. Molecular diagnostics of gliomas: state of the art. *Acta Neuropathol.* 2010;120(5):567–584.
5. Yan H, Parsons DW, Jin G, et al. IDH1 and IDH2 mutations in gliomas. *N Engl J Med.* 2009;360(8):765–773.
6. 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.
7. Hartmann C, Hentschel B, Wick W, et al. Patients with IDH1 wild type anaplastic astrocytomas exhibit worse prognosis than IDH1-mutated glioblastomas, and IDH1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas. *Acta Neuropathol.* 2010;120(6):707–718.
8. 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.
9. Patel SH, Bansal AG, Young EB, et al. Extent of surgical resection in lower-grade gliomas: differential impact based on molecular subtype. *AJNR Am J Neuroradiol.* 2019;40(7):1149–1155.

10. 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.
11. 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.
12. Yogananda CGB, 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.
13. 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.
14. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual deep convolutional neural network predicts MGMT methylation status. *J Digit Imaging.* 2017;30(5):622–628.
15. Choi KS, Choi SH, Jeong B. Prediction of IDH genotype in gliomas with dynamic susceptibility contrast perfusion MR imaging using an explainable recurrent neural network. *Neuro Oncol.* 2019;21(9):1197–1209.
16. Network CGAR. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med.* 2015; 372(26):2481–2498.
17. Kickingereder P, Isensee F, Tursunova I, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 2019;20(5):728–740.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:770–778.
19. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*: Springer; 2015:234–241.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision* 2017:618–626.
21. 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.
22. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw.* 2008; 28(5).
23. 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.
24. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science.* 2008;321(5897):1807–1812.
25. Darlix A, Deverdun J, Menjot de Champfleur N, et al. IDH mutation and 1p19q codeletion distinguish two radiological patterns of diffuse low-grade gliomas. *J Neurooncol.* 2017;133(1):37–45.
26. Zhou H, Vallières M, Bai HX, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro Oncol.* 2017;19(6):862–870.
27. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One.* 2014;9(7):e102107.
28. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol.* 2019;20(7):1124–1137.
29. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys.* 2018;102(4):1143–1158.
30. Li Q, Bai H, Chen Y, et al. A fully-automatic multiparametric radiomics model: towards reproducible and prognostic imaging signature for prediction of overall survival in glioblastoma multiforme. *Sci Rep.* 2017;7(1):14331.
31. Korfiatis P, Erickson B. Deep learning can see the unseeable: predicting molecular markers from MRI of brain gliomas. *Clin Radiol.* 2019;74(5):367–373.
32. Arita H, Kinoshita M, Kawaguchi A, et al. Lesion location implemented magnetic resonance imaging radiomics for predicting IDH and TERT promoter mutations in grade II/III gliomas. *Sci Rep.* 2018;8(1):11773.
33. Mazurowski MA, Clark K, Czarnek NM, Shamsesfandabadi P, Peters KB, Saha A. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data. *J Neurooncol.* 2017;133(1):27–35.
34. Tejada Neyra MA, Neuberger U, Reinhardt A, et al. Voxel-wise radiogenomic mapping of tumor location with key molecular alterations in patients with glioma. *Neuro Oncol.* 2018;20(11):1517–1524.
35. 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.
36. Chang PD, Malone HR, Bowden SG, et al. A multiparametric model for mapping cellularity in glioblastoma using radiographically localized biopsies. *AJNR Am J Neuroradiol.* 2017;38(5):890–898.
37. van Lent DI, van Baarsen KM, Snijders TJ, Robe PA. Radiological differences between subtypes of WHO 2016 grade II-III gliomas: a systematic review and meta-analysis. *Neuro-Oncology Advances.* 2020;2(1):vdaa044.
38. Maynard J, Okuchi S, Wastling S, et al. World Health Organization grade II/III glioma molecular status: prediction by MRI morphologic features and apparent diffusion coefficient. *Radiology.* 2020:191832(epub).
39. Ellingson BM, Bendszus M, Boxerman J, et al; Jumpstarting Brain Tumor Drug Development Coalition Imaging Standardization Steering Committee. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro Oncol.* 2015;17(9):1188–1198.