Two-Stage Training Framework Using Multicontrast MRI Radiomics for *IDH* Mutation Status Prediction in Glioma

Nghi C. D. Truong, PhD • Chandan Ganesh Bangalore Yogananda, PhD • Benjamin C. Wagner, BM • James M. Holcomb, BS • Divya Reddy, MS • Niloufar Saadat, MD • Kimmo J. Hatanpaa, MD, PhD • Toral R. Patel, MD • Baowei Fei, PhD • Matthew D. Lee, MD • Rajan Jain, MD • Richard J. Bruce, MD • Marco C. Pinho, MD • Ananth J. Madhuranthakam, PhD • Joseph A. Maldjian, MD

From the Departments of Radiology (N.C.D.T., C.G.B.Y., B.C.W., J.M.H., D.R., N.S., B.F., M.C.P., A.J.M., J.A.M.), Pathology (K.J.H.), and Neurologic Surgery (T.R.P.), The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390; Department of Bioengineering, The University of Texas at Dallas, Richardson, Tex (B.F.); Departments of Radiology (M.D.L., R.J.) and Neurosurgery (R.J.), New York University Grossman School of Medicine, New York, NY; and Department of Radiology, University of Wisconsin–Madison, Madison, Wis (R.J.B.). Received June 26, 2023; revision requested August 24; revision received March 18, 2024; accepted April 25. **Address correspondence to** N.C.D.T. (email: *Nghi.Truong@utsouthwestern.edu*).

A.J.M. and J.A.M. supported by the National Institutes of Health/National Cancer Institute grant no. U01CA207091 and J.A.M. supported by no. R01CA260705.

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

See also commentary by Moassefi and Erickson in this issue.

Radiology: Artificial Intelligence 2024; 6(4):e230218 • https://doi.org/10.1148/ryai.230218 • Content codes: AI NR

Purpose: To develop a radiomics framework for preoperative MRI-based prediction of isocitrate dehydrogenase *(IDH)* mutation status, a crucial glioma prognostic indicator.

Materials and Methods: Radiomics features (shape, first-order statistics, and texture) were extracted from the whole tumor or the combination of nonenhancing, necrosis, and edema regions. Segmentation masks were obtained via the federated tumor segmentation tool or the original data source. Boruta, a wrapper-based feature selection algorithm, identified relevant features. Addressing the imbalance between mutated and wild-type cases, multiple prediction models were trained on balanced data subsets using random forest or XGBoost and assembled to build the final classifier. The framework was evaluated using retrospective MRI scans from three public datasets (The Cancer Imaging Archive [TCIA, 227 patients], the University of California San Francisco Preoperative Diffuse Glioma MRI dataset [UCSF, 495 patients], and the Erasmus Glioma Database [EGD, 456 patients]) and internal datasets collected from the University of Texas Southwestern Medical Center (UTSW, 356 patients), New York University (NYU, 136 patients), and University of Wisconsin–Madison (UWM, 174 patients). TCIA and UTSW served as separate training sets, while the remaining data constituted the test set (1617 or 1488 testing cases, respectively).

Results: The best performing models trained on the TCIA dataset achieved area under the receiver operating characteristic curve (AUC) values of 0.89 for UTSW, 0.86 for NYU, 0.93 for UWM, 0.94 for UCSF, and 0.88 for EGD test sets. The best performing models trained on the UTSW dataset achieved slightly higher AUCs: 0.92 for TCIA, 0.88 for NYU, 0.96 for UWM, 0.93 for UCSF, and 0.90 for EGD.

Conclusion: This MRI radiomics-based framework shows promise for accurate preoperative prediction of *IDH* mutation status in patients with glioma.

Supplemental material is available for this article.

Glioma is a common and life-threatening type of brain tumor. The survival rates and responses to treatment of patients with glioma are influenced by the tumor's genetic and histologic characteristics. Recent studies have identified isocitrate dehydrogenase *(IDH)* mutations as a crucial factor in the development and progression of glioma. Therefore, the World Health Organization updated the brain tumor classification in 2016 to include molecular marker diagnostics and classic histologic diagnostics (1). The World Health Organization also recommends determining the *IDH* status of patients with glioma to guide the selection of appropriate treatment therapies.

The detection of *IDH* mutation status is mainly based on genetic profiling of tumor tissue acquired through biopsy or surgical resection. However, depending on the accessibility of the mass, brain tumor resection may not be safe, and biopsy-based methods may cause complications. Therefore, noninvasive alternatives are important for obtaining genetic and histologic information. Radiomics is a novel technique to extract multidimensional features of a region of interest (ROI) from medical images (2). These features can be used to develop diagnostic or predictive models for outcomes of interest. Since MRI is currently included in routine clinical care for patients with glioma, radiomics features extracted from MRI have gained substantial interest as a promising method to predict *IDH* mutation status in these patients.

Radiomics features extracted from multicontrast MRI have been combined with machine learning techniques to develop models for predicting *IDH* mutation status (3–9). Some studies have focused on a specific histologic subtype of glioma, such as low-grade glioma (7,10,11) or high-grade glioma (8,12–14). Most radiomics-based models have been tested on relatively limited patient cohorts,

Abbreviations

AUC = area under the ROC curve, ED = edema, EGD = Erasmus Glioma Database, FeTS = federated tumor segmentation, *IDH* = isocitrate dehydrogenase, NCR = necrosis, NET = nonenhancing tumor, NYU = New York University, ROC = receiver operating characteristic, ROI = region of interest, TCIA = The Cancer Imaging Archive, UCSF = University of California San Francisco Preoperative Diffuse Glioma MRI dataset, UTSW = University of Texas Southwestern Medical Center, UWM = University of Wisconsin–Madison, WT = whole tumor

Summary

A preoperative MRI radiomics-based model developed using a two-stage training framework demonstrated high performance in predicting *IDH* mutation status in patients with glioma.

Key Points

- *IDH* mutation status prediction models for glioma were developed using preoperative MRI radiomics features extracted from either the whole tumor or the combination of nonenhancing, necrosis, and edema regions, along with a multibalanced subset training strategy.
- The best performing models trained on The Cancer Imaging Archive dataset achieved area under the receiver operating characteristic curve (AUC) values ranging from 0.86 to 0.94 when tested on internal and public datasets.
- The best performing models trained on the University of Texas Southwestern Medical Center internal dataset achieved slightly higher AUC values, ranging from 0.88 to 0.96, on the internal and public test sets.

Keywords

Glioma, Isocitrate Dehydrogenase Mutation, *IDH* Mutation, Radiomics, MRI

either from The Cancer Imaging Archive (TCIA) dataset (3) or local data (12–14), primarily using the cross-validation method. A few radiomics-based studies extended assessment of the prediction model on an independent test set (7,8,15). Although these studies have reported *IDH* prediction accuracies ranging from 72% to 97%, an extensive evaluation of the prediction models on a larger patient sample is still needed to establish their effectiveness in clinical practice.

This study focused on developing *IDH* mutation prediction models in patients with glioma using preoperative MRI radiomics features and a two-stage training framework. Unlike previous studies that focused on specific tumor subcompartments, this study extracted radiomics features from either the whole tumor (WT) or the combination of nonenhancing tumor, necrosis, and edema regions (NET + NCR + ED). This inclusive approach allowed all patients to be included in the study, regardless of whether certain tumor subcompartments were absent (eg, enhancing tumor). Relevant radiomics features for *IDH* genotyping were identified through a feature selection algorithm and used in conjunction with machine learning techniques to build prediction models. Multiple models were trained using different balanced subsets resampled from the original imbalanced training dataset and then ensembled to build the final classifier. The derived models were tested on a diverse patient sample archived from multiple institutions with varying MRI acquisition protocols, preprocessing methods, and tumor mask qualities.

Materials and Methods

Datasets

This study used retrospective MRI scans from three publicly available and three internal datasets. The public datasets included data from The Cancer Genome Atlas (16) and the Ivy Glioblastoma Atlas (17), which were both downloaded from and together referred to as TCIA (18); the University of California San Francisco Preoperative Diffuse Glioma MRI dataset (UCSF) (19); and the Erasmus Glioma Database (EGD) (20). The internal datasets were collected from three geographically distinct institutions, namely the University of Texas Southwestern Medical Center (UTSW), New York University (NYU), and the University of Wisconsin–Madison (UWM). UTSW Institutional Review Board approval was obtained with a waiver of consent for the use of retrospective data or public datasets. All internal data were anonymized, and the study was compliant with the Health Insurance Portability and Accountability Act.

Data from the patients that met the following criteria were included in this study: *(a)* newly diagnosed with glioma; *(b)* *IDH* mutation status was available; *(c)* preoperative MRI scans with T1-weighted, postcontrast T1-weighted, T2-weighted, and T2-weighted fluid-attenuated inversion recovery sequences were available; and *(d)* tumor segmentation was available.

Image Preprocessing and Multiregional Tumor Segmentation

The MRI data included in this study were collected from multiple sources and underwent distinct preprocessing tools due to the unavailability of raw data for uniform processing. However, the preprocessing pipeline of all datasets consisted of standardized steps commonly used for multimodal glioma analysis, including registering to a common anatomic space with a voxel resolution of 1 × 1 × 1 mm^3, correcting for bias field distortion, coregistering MRI scans to a template atlas, and removing all nonbrain tissue (skull stripping) from the image. The federated tumor segmentation (FeTS) (21) tool was used to preprocess the TCIA and internal datasets. The UCSF dataset underwent preprocessing using multiple publicly available tools, including the Advanced Normalization Tools (22) and the brain masking tool (23). For the EGD dataset, all scans were registered using Elastix, version 5.0.0 (3CX) (24), and the skull was stripped by HD-BET (25).

Tumor segmentation masks for the TCIA and three internal datasets were obtained using the FeTS tool. FeTS segmented the tumor into three subcompartments: the necrotic tumor core (NCR, label 1), the NET and peritumoral edematous/invaded tissue (NET/ED, label 2), and the enhancing part of the tumor (ET, label 4). These automated masks were used directly to extract radiomics features without manual correction.

For the UCSF dataset, a different automated segmentation tool based on the multimodal brain tumor segmentation challenge algorithm was used to obtain automated tumor segmentation masks (26). These masks were then corrected manually by a group of annotators and approved by a neuroradiologist with more than 15 years of experience. However, the segmentation

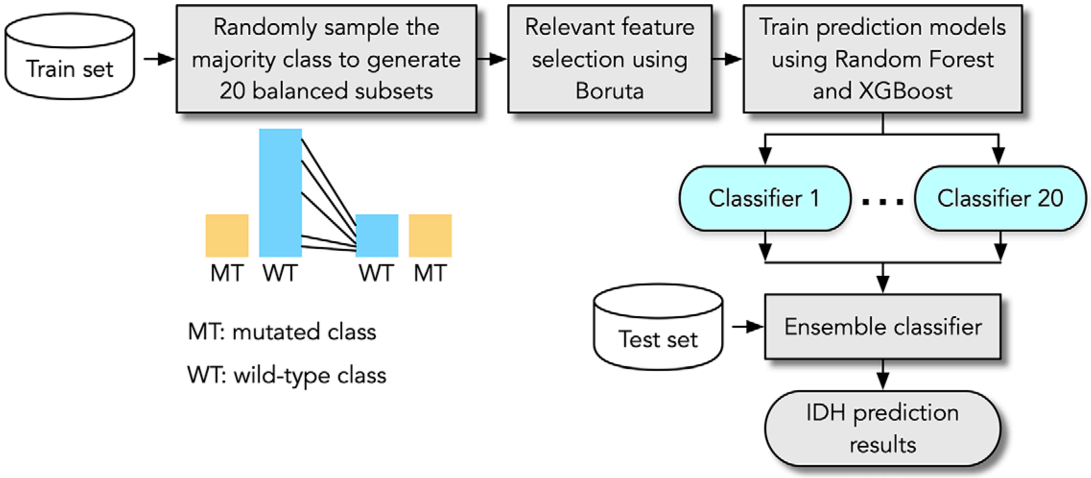
Train set
Randomly sample the majority class to generate 20 balanced subsets
Relevant feature selection using Boruta
Train prediction models using Random Forest and XGBoost
Classifier 1
Classifier 20
MT
WT
WT
MT
MT: mutated class
WT: wild-type class
Test set
Ensemble classifier
IDH prediction results

Figure 1: Flowchart of proposed MRI radiomics-based framework for predicting isocitrate dehydrogenase *(IDH)* mutation status in gliomas.

labels for the UCSF dataset were slightly different from those of the TCIA and internal datasets. The tumor in the UCSF dataset was segmented into three subcompartments: the NET and necrotic tumor core (NET/NCR, label 1), ED (ED, label 2), and enhancing tumor (ET, label 4).

Finally, for the EGD dataset, only WT masks were available. These masks were segmented either manually or automatically using a convolutional neural network–based method (27). The automated masks in the EGD dataset were not manually corrected.

Radiomics Feature Extraction

The brain tumor radiomics features were extracted using the PyRadiomics Python package (28). To ensure that all patients were included in the study, regardless of the presence or absence of any tumor subcompartment, two ROIs were defined: the WT and the nonenhancing and edematous (NET + NCR + ED) region. Radiomics features were extracted from both ROIs for each of the four MRI sequences (T1-weighted, postcontrast T1-weighted, T2-weighted, and T2-weighted fluid-attenuated inversion recovery).

Different types of radiomics features were extracted from the brain tumor, including shape, first-order, and texture features. Specifically, we extracted 14 shape features, 18 first-order features, 22 gray-level co-occurrence matrix features, 16 gray-level run-length matrix features, 16 gray-level size zone matrix features, 14 gray-level dependence matrix features, and five neighboring gray-tone difference matrix features.

In addition to the original images, 12 derived images for each MRI sequence were also generated to extract additional radiomics features. These derived images were obtained using different filters, including wavelet filtering at four levels, Laplacian of Gaussian filtering with four levels of σ (from 2 to 5 mm), square, square root, logarithm, and exponential operators. Each of these derived images emphasizes different characteristics of the original images, thereby enriching the extracted radiomics features. In total, we extracted 1197 radiomic features from each ROI and each MRI sequence. The detailed description of the feature extraction process is presented in Appendix S1. PyRadiomics settings, as well as image types, feature classes, and feature names used for feature extraction, are summarized in Tables S3 and S4.

Data Balancing

The overall prevalence of *IDH*-mutated tumors in our cohort was approximately 27%, which is lower than *IDH* wild type. Hence, to enhance classification performance on the unbalanced dataset, a two-stage training framework was implemented (Fig 1). In the first stage, multiple balanced training subsets were generated by randomly sampling the majority class (ie, *IDH* wild-type instances were randomly sampled to achieve a 1:1 ratio with the mutated cases). Then, a set of prediction models was trained using these subsets. In the second stage, we ensembled the prediction models by averaging the prediction probabilities of *IDH* mutation status. This approach was designed to produce an efficient classifier with improved accuracy and reduced bias toward the majority class.

Feature Selection and *IDH* Mutation Classification

Boruta feature selection (29) was employed to identify the most relevant radiomics features for predicting *IDH* mutations. The Boruta method operates by generating random shadow features, which serve as a reference point for the actual features. The most relevant features were identified by assessing how frequently they outperform these shadow features. In our framework, Boruta was used on multiple balanced subsets of the training data. The most frequently selected features across these subsets were then considered the relevant feature set to train the classifier models. The radiomics features from each individual ROI (WT or [NET + NCR + ED]), as well as the combined features of both ROIs, were passed through Boruta to determine three sets of the most pertinent features.

The *IDH* mutation prediction models were then trained using two classifiers, random forest (30) and XGBoost (31). RF and XGBoost are both ensemble methods, which enable them to capture complex data relationships and handle high-dimensional

Table 1: Patient Characteristics of Different Datasets

Variable	TCIA (*n* = 227)	UCSF (*n* = 495)	EGD (*n* = 456)	UTSW (*n* = 356)	NYU (*n* = 136)	UWM (*n* = 174)	Total (*n* = 1844)
IDH status							
Mutated	94 (41.4)	103 (20.8)	150 (32.9)	102 (28.7)	23 (16.9)	19 (10.9)	491 (26.6)
Wild type	133 (58.6)	392 (79.2)	306 (67.1)	254 (71.3)	113 (83.1)	155 (89.1)	1353 (73.4)
Sex							
Female	103 (45.4)	199 (40.2)	171 (37.5)	145 (40.7)	63 (46.3)	0	681 (36.9)
Male	108 (47.6)	296 (59.8)	284 (62.3)	193 (54.2)	73 (53.7)	0	954 (51.7)
Unknown	16 (7.0)	0	1 (0.2)	18 (5.1)	0	174 (100)	209 (11.3)
Age (y)							
> 65	45 (19.8)	154 (31.1)	136 (29.8)	101 (28.4)	33 (24.3)	0	469 (25.4)
≤ 65	166 (73.1)	341 (68.9)	280 (61.4)	233 (65.4)	53 (39.0)	0	1073 (58.2)
Unknown	16 (7.0)	0	40 (8.8)	22 (6.2)	50 (36.8)	174 (100)	302 (16.4)
Median	54 (41–63)	59 (47–68)	56.5 (45–69)	56.5 (44–67)	61 (43–68)	…	…
Segmentation							
Manual	0	495 (100)	0	0	0	0	495 (26.8)
FeTS	227 (100)	0	0	356 (100)	136 (100)	174 (100)	893 (48.4)
Other	0	0	456 (100)	0	0	0	456 (24.7)

Note.—Data are presented as numbers of patients with percentages in parentheses or medians with IQRs in parentheses. EGD = Erasmus Glioma Database, FeTS = federated tumor segmentation, *IDH* = isocitrate dehydrogenase, NYU = New York University, TCIA = The Cancer Imaging Archive, UCSF = University of California San Francisco Preoperative Diffuse Glioma MRI dataset, UTSW = University of Texas Southwestern Medical Center, UWM = University of Wisconsin–Madison.

feature spaces effectively. Additionally, they exhibit resilience to outliers, making them robust choices for this study. *IDH* prediction models were trained on multiple balanced subsets derived from either TCIA (*IDH* mutated, *n* = 94; *IDH* wild type, *n* = 133) or UTSW (*IDH* mutated, *n* = 102; *IDH* wild type, *n* = 256) data. These two datasets were selected as the training data since they had a sufficient number of mutated cases and were preprocessed by FeTS with the same tumor subregion annotations approach. The trained models were tested on all other held-out datasets. UTSW data were included in the test set for the models derived from the TCIA data used as the training set, and vice versa. The parameters for the Boruta feature selection method, as well as the random forest and XGBoost classifiers, can be found in Table S5.

Statistical Analysis

The performance of the prediction models was assessed using several metrics, including accuracy, sensitivity, specificity, precision, F1 score, and the area under the receiver operating characteristic curve (AUC). To compute the AUC, the prediction probabilities of the mutated class were used to construct the receiver operating characteristic (ROC) curve. Various probability thresholds were applied to classify *IDH* status into either the mutated or wild-type class, resulting in a series of (1 – specificity, sensitivity) points forming the ROC curve. The AUC was calculated as the area under the ROC curve. The CI of the AUC was calculated using the DeLong method (32). The covariance of sensitivity and (1 – specificity) across all possible classification thresholds was first computed. This covariance was used to estimate the variance of the AUC. The CI was then derived using the standard normal distribution. Statistical metrics were calculated using the Scikit-learn package, version 1.2.1, in Python, version 3.8. No statistical significance testing was conducted in this study.

Results

Patient Characteristics

A total of 1844 patients across all datasets were included in this study (TCIA, 227 patients; UCSF, 495 patients; EGD, 456 patients; UTSW, 356 patients; NYU, 136 patients; and UWM, 174 patients). Table 1 summarizes the patient characteristics and *IDH* status of all datasets.

Performance of Radiomics Models for *IDH* Mutation Status Prediction

The prediction performance of different models trained by TCIA and UTSW datasets is detailed in Tables 2 and 3, respectively. The 95% CIs for the AUC values are reported in Tables S7 and S8. Three models were compared: two models built from multicontrast features extracted from a single ROI (WT or [NET + NCR + ED]) and one model built from the combined features of these two ROIs. Because the EGD dataset had only WT masks, the reported results were only for the WT.

For models trained on the TCIA dataset, the highest AUC values were 0.89 (95% CI: 0.86, 0.92) for UTSW, 0.86 (95% CI: 0.80, 0.93) for NYU, 0.93 (95% CI: 0.89, 0.97) for UWM, 0.94 (95% CI: 0.92, 0.96) for UCSF, and 0.88 (95% CI: 0.85, 0.91) for EGD test sets, all obtained using the random forest classifier. The ROC curves for the combined test sets, including UTSW, NYU, UWM, and UCSF, are visually represented in

Table 2: Summary of the *IDH* Mutation Status Prediction Performance of Different Models Trained on the TCIA Dataset

Model/	WT						NET + NCR + ED						Combined					
Classifier	ACC	SEN	SPE	PRE	F1	AUC	ACC	SEN	SPE	PRE	F1	AUC	ACC	SEN	SPE	PRE	F1	AUC
UTSW (MT: 102, WT: 254, TT: 356)																		
RF	81.5	76.5	83.5	65.0	70.3	0.88	82.9	83.3	82.7	65.9	73.6	0.89	85.1	76.5	88.6	72.9	74.6	0.89*
XGB	82.3	84.3	81.5	64.7	73.2	0.89	82.6	76.5	85.0	67.2	71.6	0.88	82.0	76.5	84.3	66.1	70.9	0.86
NYU (MT: 23, WT: 113, TT: 136)																		
RF	85.3	56.5	91.2	56.5	56.5	0.83	86.8	65.2	91.2	60.0	62.5	0.83	90.4	60.9	96.5	77.8	68.3	0.86*
XGB	83.8	56.5	89.4	52.0	54.2	0.81	87.5	43.5	96.5	71.4	54.1	0.86	91.2	56.5	98.2	86.7	68.4	0.82
UWM (MT: 19, WT: 155, TT: 174)																		
RF	83.9	78.9	84.5	38.5	51.7	0.90	90.8	73.7	92.9	56.0	63.6	0.93*	88.5	84.2	89.0	48.5	61.5	0.92
XGB	82.8	84.2	82.6	37.2	51.6	0.91	92.0	73.7	94.2	60.9	66.7	0.92	90.8	68.4	93.5	56.5	61.9	0.90
UCSF (MT: 103, WT: 392, TT: 495)																		
RF	84.4	89.3	83.2	58.2	70.5	0.93	89.9	74.8	93.9	76.2	75.5	0.90	91.3	82.5	93.6	77.3	79.8	0.94*
XGB	83.4	90.3	81.6	56.4	69.4	0.93	89.3	69.9	94.4	76.6	73.1	0.88	88.7	67.0	94.4	75.8	71.1	0.92
Overall accuracy of UTSW, NYU, UWM, and UCSF																		
RF	83.5	80.2	84.5	58.1	67.3	0.90	87.5	77.3	90.2	68.0	72.3	0.90	88.9	78.1	91.7	71.7	74.8	0.92*
XGB	83.0	84.2	82.7	56.7	67.8	0.90	87.4	70.4	92.0	70.2	70.3	0.89	87.2	70.0	91.8	69.8	69.9	0.89
EGD (MT: 150, WT: 306, TT: 456)																		
RF	73.8	90.6	65.7	56.2	69.4	0.88*	…	…	…	…	…	…	…	…	…	…	…	…
XGB	73.2	93.3	63.4	55.4	69.5	0.86	…	…	…	…	…	…	…	…	…	…	…	…

Note.—Sensitivity (SEN) and specificity (SPE) correspond to the accuracy of the mutated and wild-type classes, respectively. ACC = accuracy, AUC = area under the receiver operating characteristic curve, ED = edema, EGD = Erasmus Glioma Database, F1 = F1 score, MT = number of mutated cases, NCR = necrosis, NET = nonenhancing tumor, NYU = New York University, PRE = precision, RF = random forest, TCIA = The Cancer Imaging Archive, TT = total cases, UCSF = University of California San Francisco Preoperative Diffuse Glioma MRI dataset, UTSW = University of Texas Southwestern Medical Center, UWM = University of Wisconsin–Madison, WT = whole tumor (number of wild-type cases), XGB = XGBoost.
* Indicates the highest AUC achieved for each dataset.

Figure 2A. The models leveraging features extracted from both the WT and (NET + NCR + ED) ROIs appeared to exhibit slightly better performance.

Classifiers trained on the UTSW dataset appeared to perform slightly better than those trained with the TCIA dataset. Specifically, the highest AUC values were 0.92 (95% CI: 0.88, 0.95) for TCIA, 0.88 (95% CI: 0.81, 0.94) for NYU, 0.96 (95% CI: 0.93, 0.99) for UWM, 0.93 (95% CI: 0.91, 0.95) for UCSF, and 0.90 (95% CI: 0.87, 0.93) for EGD test sets, achieved mainly by XGBoost. Features extracted from the (NET + NCR + ED) ROI or the combination of features from both the WT and (NET + NCR + ED) ROIs generally led to improved AUC in most of the test datasets compared with using only the WT features. Model prediction performance on

Table 3: Summary of the *IDH* Mutation Status Prediction Performance of Different Models Trained on the UTSW Dataset

	WT						NET + NCR + ED						Combined					
Model/Classifier	ACC	SEN	SPE	PRE	F1	AUC	ACC	SEN	SPE	PRE	F1	AUC	ACC	SEN	SPE	PRE	F1	AUC
TCIA (MT: 94, WT: 133, TT: 227)																		
RF	78.0	77.7	78.2	71.6	74.5	0.83	83.3	85.1	82.0	76.9	80.8	0.92*	83.3	80.9	85.0	79.2	80.0	0.89
XGB	78.0	75.5	79.7	72.4	74.0	0.83	84.6	87.2	82.7	78.1	82.4	0.91	83.7	83.0	84.2	78.8	80.8	0.90
NYU (MT: 23, WT: 113, TT: 136)																		
RF	86.0	65.2	90.3	57.7	61.2	0.87	90.4	60.9	96.5	77.8	68.3	0.87	90.4	60.9	96.5	77.8	68.3	0.88*
XGB	88.2	69.6	92.0	64.0	66.7	0.86	90.4	60.9	96.5	77.8	68.3	0.88	91.2	65.2	96.5	78.9	71.4	0.86
UWM (MT: 19, WT: 155, TT: 174)																		
RF	92.5	78.9	94.2	62.5	69.8	0.88	92.5	52.6	97.4	71.4	60.6	0.95	94.8	73.7	97.4	77.8	75.7	0.90
XGB	93.7	78.9	95.5	68.2	73.2	0.87	94.3	63.2	98.1	80.0	70.6	0.96*	96.0	78.9	98.1	83.3	81.1	0.93
UCSF (MT: 103, WT: 392, TT: 495)																		
RF	89.1	80.6	91.3	70.9	75.5	0.91	89.5	74.8	93.4	74.8	74.8	0.91	90.7	77.7	94.1	77.7	77.7	0.93
XGB	88.5	72.8	92.6	72.1	72.5	0.91	89.3	76.7	92.6	73.1	74.9	0.92	91.1	80.6	93.9	77.6	79.0	0.93*
Overall accuracy of TCIA, NYU, UWM, and UCSF																		
RF	86.8	77.8	89.5	68.9	73.1	0.89	88.8	75.7	92.7	75.7	75.7	0.92	89.7	77.0	93.5	78.0	77.5	0.91
XGB	87.1	74.1	91.0	71.1	72.5	0.89	89.3	78.2	92.6	76.0	77.1	0.92	90.4	79.9	93.5	78.6	79.3	0.92*
EGD (MT: 150, WT: 306, TT: 456)																		
RF	80.2	91.9	74.5	63.7	75.3	0.89	…	…	…	…	…	…	…	…	…	…	…	…
XGB	82.6	89.9	79.1	67.7	77.2	0.90*	…	…	…	…	…	…	…	…	…	…	…	…

Note.—Sensitivity (SEN) and specificity (SPE) correspond to the accuracy of the mutated and wild-type classes, respectively. ACC = accuracy, AUC = area under the receiver operating characteristic curve, ED = edema, EGD = Erasmus Glioma Database, F1 = F1 score, MT = number of mutated cases, NCR = necrosis, NET = nonenhancing tumor, NYU = New York University, PRE = precision, RF = random forest, TCIA = The Cancer Imaging Archive, TT = total cases, UCSF = University of California San Francisco Preoperative Diffuse Glioma MRI dataset, UTSW = University of Texas Southwestern Medical Center, UWM = University of Wisconsin–Madison, WT = whole tumor (number of wild-type cases), XGB = XGBoost.
* Indicates the highest AUC achieved for each dataset.

the EGD dataset showed a slight improvement when trained using the UTSW dataset compared with the TCIA dataset (AUC: 0.88 [95% CI: 0.85, 0.91] for TCIA and 0.90 [95% CI: 0.87, 0.93] for UTSW).

Analysis of Radiomics Features Contributions

Figure 3 provides an overview of the contributions of various image types, feature classes, and MRI sequences to the relevant feature set. Predominantly, the original images, squared images, and Laplacian of Gaussian images were the main contributors among the image types. Similarly, first-order statistics, gray-level co-occurrence matrix, and gray-level size zone matrix were the most selected feature classes. Furthermore, Figure 4 presents the most relevant features along with their important scores, which are quantified by the average of the absolute Shapley additive explanations (33) values, offering a clear insight into their impact.

Discussion

We developed a preoperative MRI radiomics-based framework for predicting the *IDH* mutation status of gliomas. The prediction models were trained using the TCIA or UTSW datasets and tested on several independent test sets, including the NYU, UWM, UCSF, and EGD datasets. For the mod-

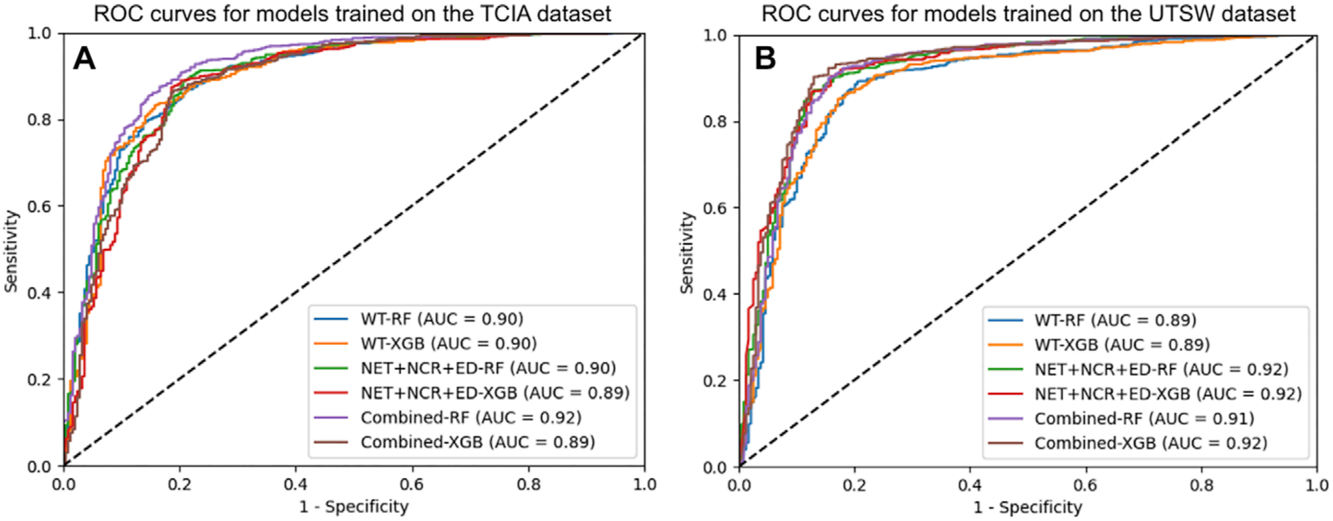

Figure 2: Receiver operating characteristic (ROC) curves for the random forest and XGBoost (XGB) models trained using relevant features from the whole tumor; the nonenhancing tumor (NET), necrosis (NCR), and edema (ED) region of interest; and the combined features from both regions of interest. **(A)** ROC curves for the combined test sets from the University of Texas Southwestern Medical Center (UTSW), New York University (NYU), University of Wisconsin–Madison, and University of California San Francisco Preoperative Diffuse Glioma MRI dataset (UCSF) obtained by models trained on The Cancer Imaging Archive (TCIA) dataset. **(B)** ROC curves for the combined test sets from TCIA, UTSW, NYU, and UCSF obtained by models trained on the UTSW dataset. AUC = area under the receiver operating characteristic curve, RF = random forest.

els trained on the TCIA dataset, the best overall AUC values were 0.92 (95% CI: 0.90, 0.93) on 1038 patients from the UTSW, NYU, UWM, and UCSF datasets, using combined features from both ROIs, and 0.88 (95% CI: 0.85, 0.91) on 456 patients of the EGD dataset, using the WT features. The models trained on the UTSW dataset appeared to perform slightly better than the TCIA-trained models likely due to the larger number of wild-type cases in the UTSW dataset and the two-stage training framework. The ROC curves showed improvement for the models trained by features from the (NET + NCR + ED) ROI or the combined features from both ROIs. Random forest and XGBoost algorithms performed comparably for both trained datasets.

Our model's *IDH* prediction performance was assessed on an independent testing cohort comprising more than 1500 patients, marking it as one of the most extensive radiomics-based investigations. Previous studies have typically dealt with smaller datasets. For example, Lu et al (15) trained their radiomics-based prediction model on 306 patients from local institutions and tested it on 108 patients from The Cancer Genome Atlas, resulting in an AUC of 0.88. Li et al (3) employed a training cohort of 118 patients and a validation cohort of 107 patients, achieving the highest AUC of 0.96 when incorporating both radiomics and age as features. Some deep learning–based studies have incorporated slightly larger testing cohorts. van de Voort et al (34) trained their model on 1508 patients and tested on 221 patients from The Cancer Genome Atlas datasets, achieving an AUC of 0.90. Wu et al (35) attained an AUC of 0.87 when tested on 234 patients from an internal independent dataset. Our study, by leveraging a testing cohort exceeding 1500 patients, surpasses the prior studies in terms of dataset size, further bolstering the robustness and generalizability of our *IDH* prediction model.

Previous studies have revealed that training machine learning models on imbalanced data may result in prediction bias. Although various techniques have been used to address this issue, bias has remained a persistent problem. For instance, Li et al (13) addressed this problem by oversampling the mutated class, which comprised less than 10% of the cases in the training dataset, achieving an accuracy of 70% for the mutated class and 99% for the wild-type class. Another radiomics-based study (36) balanced the proportions of glioblastomas and anaplastic gliomas in the training dataset to account for the minority of *IDH*-mutated tumors rather than the proportions of mutated and wild-type tumors. This study reported an accuracy of 42% for the mutated class and 100% for the wild-type class.

In our study, we proposed a two-stage training framework that involves training multiple models on balanced subsets obtained by resampling the original imbalanced dataset. This approach resulted in 79.9% accuracy for the mutated class and 93.5% for the wild-type class, compared with the corresponding accuracy of 73.6% and 94.9% when training the models on balanced data using the synthetic minority oversampling technique (37), in which the mutated class was augmented to match the number of wild-type cases from the original training data. The full results obtained by the synthetic minority oversampling technique are presented in Tables S9 and S10. Thus, the two-stage training approach helped reduce bias and improve the accuracy of the mutated class.

The proposed radiomics-based framework extracted features from either the WT or a combination of NET, NCR, and ED subregions, making it suitable for any glioma grade. This approach differs from previous studies that focused on specific tumor subcompartments. By adopting this inclusive approach, all patients could be included in the study, regardless of whether certain tumor subcompartments were absent. Using these two tumor ROIs also enables the models to be trained

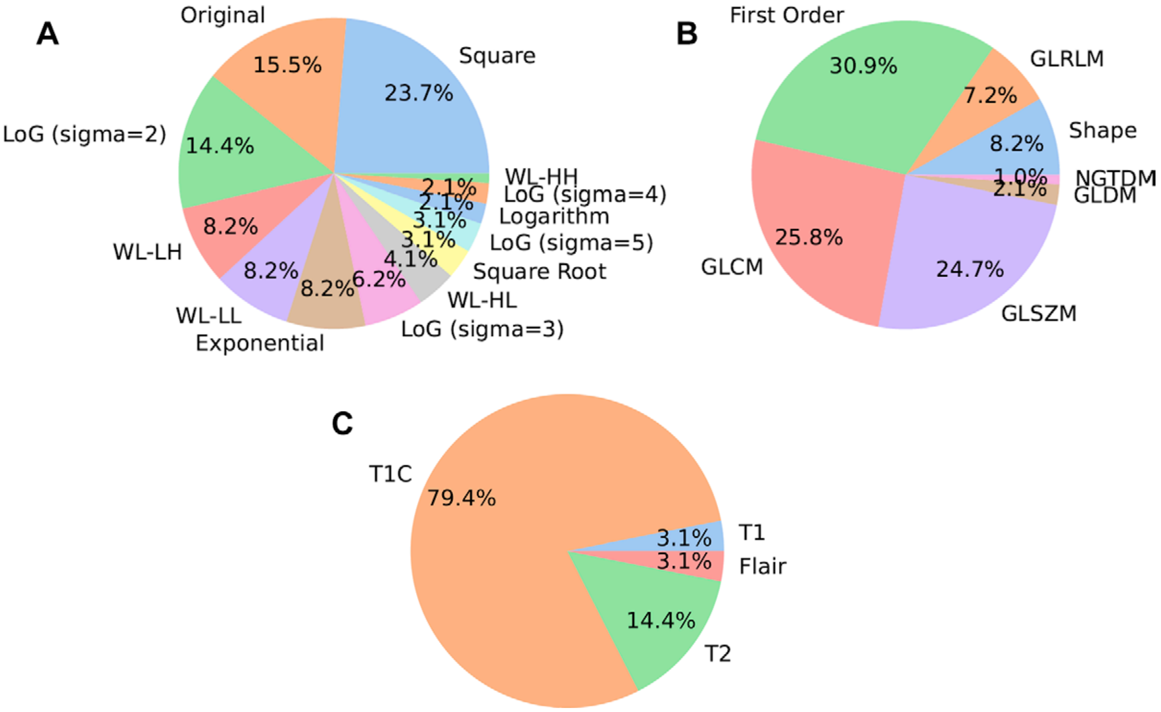

Figure 3: Summary of the most frequently selected image types, feature classes, and MRI sequences. **(A)** The most selected image types. **(B)** The most selected feature classes. **(C)** The percentage of MRI sequences identified as relevant features. GLCM = gray-level co-occurrence matrix, GLDM = gray-level dependence matrix, GLRLM = gray-level run-length matrix, GLSZM = gray-level size zone matrix, LoG = Laplacian of Gaussian, NGTDM = neighboring gray-tone difference matrix, WL-HH = wavelet filtering with high-pass filters, WL-HL = wavelet filtering with high-pass and low-pass filters, WL-LH = wavelet filtering with low-pass and high-pass filters, WL-LL = wavelet filtering with low-pass filters.

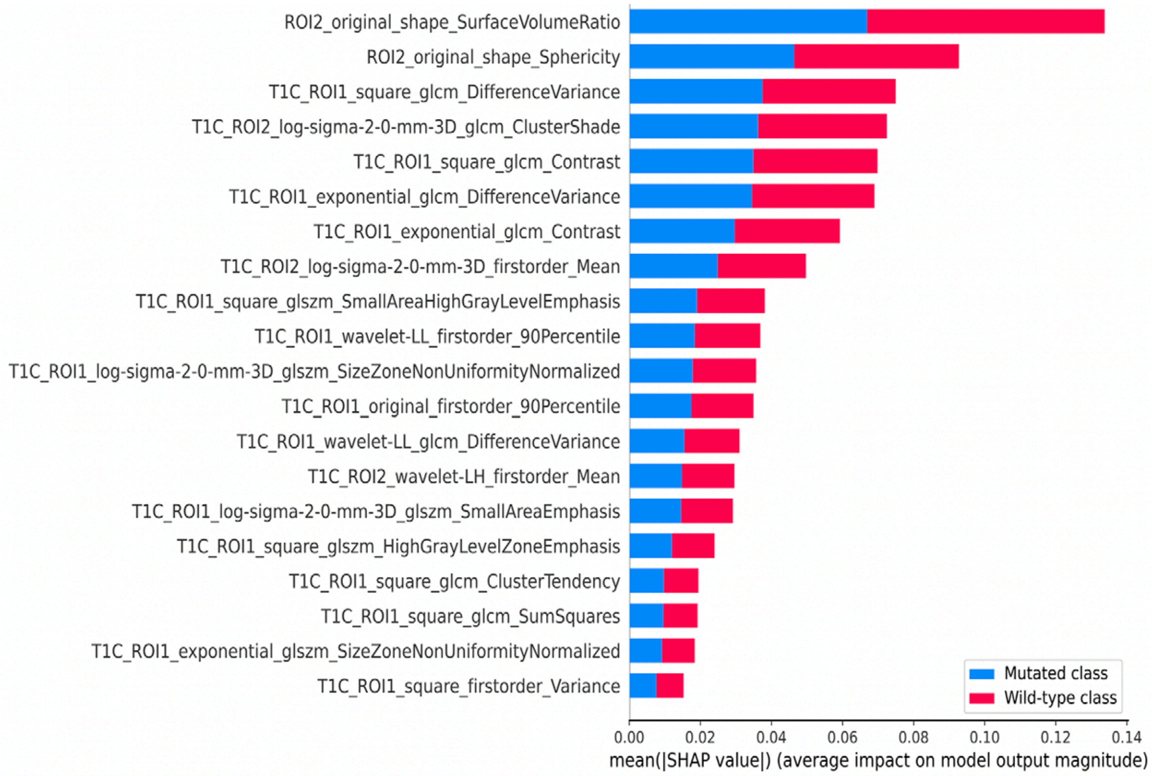

Figure 4: Chart of the top 20 relevant features with their importance scores measured by Shapley additive explanations (SHAP) values. The feature names are labeled as region of interest 1 (ROI1) and ROI2, representing the whole tumor and the nonenhancing tumor (NET), necrosis (NCR), and edema (ED) region of interests, respectively. glcm = gray-level co-occurrence matrix, glszm = gray-level size zone matrix, log = Laplacian of Gaussian, ROI = region of interest, wavelet-LH = wavelet filtering with low-pass and high-pass filters, wavelet-LL = wavelet filtering with low-pass filters.

and tested across various datasets with different definitions of the tumor subregions.

Figures 3 and 4 succinctly outline the distinctive contributions of different image types, feature classes, and MRI sequences to the relevant feature set. These features, derived from the combined region, offer a more holistic perspective of the tumor, potentially uncovering patterns that might be overlooked when focusing solely on individual tumor subcompartments. Notably, two shape features extracted from the (NET + NCR + ED) region exhibit high importance scores (Fig 4). The integration of these features has the potential to enhance the performance of the prediction models.

Our study had limitations. The main limitation of the radiomics-based approach is the need for a tumor mask, and the reliability of radiomic features depends on the accuracy of the tumor segmentation. However, brain tumor segmentation techniques have recently undergone substantial advancements, and many automatic brain tumor segmentation tools are now available. In our study, both the TCIA and internal data (UTSW, NYU, and UWM) were segmented using FeTS without manual correction. Although the radiomics features were extracted directly from the masks generated by FeTS, we achieved high prediction accuracies when testing on a large patient sample.

In conclusion, we present an MRI radiomics-based approach for predicting the *IDH* mutation status in both low-grade and high-grade gliomas. *IDH* prediction models were built based on a set of relevant radiomics features extracted from multicontrast MR images and two ROIs. The random forest and XGBoost methods were used as classifiers. A two-stage training strategy was adopted to address the unbalanced training data. The models were trained on either the TCIA or UTSW dataset and tested on the independent data, yielding promising prediction accuracy across a large and diverse patient sample. Future research may focus on improving performance of the *IDH* prediction models by incorporating patient demographic characteristics, as suggested in Jiang et al (38), and implementing a rigorous quality assurance procedure to ensure that the segmentation data meet rigorous standards of accuracy and reliability.

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

Disclosures of conflicts of interest: **N.C.D.T.** No relevant relationships. **C.G.B.Y.** No relevant relationships. **B.C.W.** No relevant relationships. **J.M.H.** No relevant relationships. **D.R.** No relevant relationships. **N.S.** No relevant relationships. **K.J.H.** No relevant relationships. **T.R.P.** No relevant relationships. **B.F.** No relevant relationships. **M.D.L.** No relevant relationships. **R.J.** No relevant relationships. **R.J.B.** Supported by NIH 1R01LM013151-01A1. **M.C.P.** No relevant relationships. **A.J.M.** Supported by NIH/NCI U01CA207091. **J.A.M.** Supported by NIH/NCI U01CA207091 and R01CA260705.

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 (Berl) 2016;131(6):803–820.
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278(2):563–577.
3. Li Y, Ammari S, Lawrance L, et al. Radiomics-Based Method for Predicting the Glioma Subtype as Defined by Tumor Grade, IDH Mutation, and 1p/19q Codeletion. Cancers (Basel) 2022;14(7):1778.
4. Liu X, Li Y, Li S, et al. IDH mutation-specific radiomic signature in lower-grade gliomas. Aging (Albany NY) 2019;11(2):673–696.
5. Tan Y, Zhang S-T, Wei J-W, et al. A radiomics nomogram may improve the prediction of IDH genotype for astrocytoma before surgery. Eur Radiol 2019;29(7):3325–3337.
6. Wu S, Meng J, Yu Q, Li P, Fu S. Radiomics-based machine learning methods for isocitrate dehydrogenase genotype prediction of diffuse gliomas. J Cancer Res Clin Oncol 2019;145(3):543–550.
7. Yu J, Shi Z, Lian Y, et al. Noninvasive IDH1 mutation estimation based on a quantitative radiomics approach for grade II glioma. Eur Radiol 2017;27(8):3509–3522.
8. 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.
9. 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.
10. Ding H, Huang Y, Li Z, et al. Prediction of IDH Status Through MRI Features and Enlightened Reflection on the Delineation of Target Volume in Low-Grade Gliomas. Technol Cancer Res Treat 2019;18:1533033819877167.
11. 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.
12. Lee MH, Kim J, Kim S-T, et al. Prediction of IDH1 Mutation Status in Glioblastoma Using Machine Learning Technique Based on Quantitative Radiomic Data. World Neurosurg 2019;125:e688–e696.
13. Li Z-C, Bai H, Sun Q, et al. Multiregional radiomics profiling from multiparametric MRI: Identifying an imaging predictor of IDH1 mutation status in glioblastoma. Cancer Med 2018;7(12):5999–6009.
14. Wang Q, Zhang J, Li F, Xu X, Xu B. Diagnostic performance of clinical properties and conventional magnetic resonance imaging for determining the IDH1 mutation status in glioblastoma: a retrospective study. PeerJ 2019;7:e7154.
15. Lu J, Xu W, Chen X, Wang T, Li H. Noninvasive prediction of IDH mutation status in gliomas using preoperative multiparametric MRI radiomics nomogram: A mutlicenter study. Magn Reson Imaging 2023;104:72–79.
16. Ceccarelli M, Barthel FP, Malta TM, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell 2016;164(3):550–563.
17. Puchalski RB, Shah N, Miller J, et al. An anatomic transcriptional atlas of human glioblastoma. Science 2018;360(6389):660–663.
18. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26(6):1045–1057.
19. Calabrese E, Villanueva-Meyer JE, Rudie JD, et al. The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. Radiol Artif Intell 2022;4(6):e220058.
20. van der Voort SR, Incekara F, Wijnenga MMJ, et al. The Erasmus Glioma Database (EGD): Structural MRI scans, WHO 2016 subtypes, and segmentations of 774 patients with glioma. Data Brief 2021;37:107191.
21. Pati S, Baid U, Edwards B, et al. The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research. Phys Med Biol 2022;67(20):204002.
22. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 2011;54(3):2033–2044.
23. Calabrese E, Villanueva-Meyer JE, Cha S. A fully automated artificial intelligence method for non-invasive, imaging-based identification of genetic alterations in glioblastomas. Sci Rep 2020;10(1):11852.
24. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29(1):196–205.
25. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp 2019;40(17):4952–4964.
26. Bakas S, Reyes M, Jakab A, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. ArXiv 1811.02629 [preprint] https://arxiv.org/abs/1811.02629. Posted November 5, 2018. Accessed January 18, 2023.
27. van der Voort SR, Incekara F, Wijnenga MMJ, et al. WHO 2016 subtyping and automated segmentation of glioma using multi-task deep learning. ArXiv

2010.04425 [preprint] https://arxiv.org/abs/2010.04425. Posted October 9, 2020. Accessed January 18, 2023.

28. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77(21):e104–e107
29. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J Stat Softw 2010;36(11):1–13.
30. Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
31. Chen TQ, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016; 785–794.
32. 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.
33. Chaddad A, Zinn PO, Colen RR. Radiomics texture feature extraction for characterizing GBM phenotypes using GLCM. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). IEEE, 2015; 84–87.
34. van der Voort SR, Incekara F, Wijnenga MMJ, et al. Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. Neuro Oncol 2023;25(2):279–289.
35. Wu J, Xu Q, Shen Y, Chen W, Xu K, Qi XR. Swin Transformer Improves the IDH Mutation Status Prediction of Gliomas Free of MRI-Based Tumor Segmentation. J Clin Med 2022;11(15):4625.
36. Alis D, Bagcilar O, Senli YD, et al. Machine learning-based quantitative texture analysis of conventional MRI combined with ADC maps for assessment of IDH1 mutation in high-grade gliomas. Jpn J Radiol 2020;38(2):135–143.
37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(1):321–357.
38. Jiang S, Zanazzi GJ, Hassanpour S. Predicting prognosis and IDH mutation status for patients with lower-grade gliomas using whole slide images. Sci Rep 2021;11(1):16849.