Neuro-Oncology Advances

6(1), vdae157, 2024 | https://doi.org/10.1093/noajnl/vdae157 | Advance Access date 3 October 2024

Enhancing clinical decision-making: An externally validated machine learning model for predicting isocitrate dehydrogenase mutation in gliomas using radiomics from presurgical magnetic resonance imaging

Jan Lost, Nader Ashraf, Leon Jekel, Marc von Reppert, Niklas Tillmanns, Klara Willms, Sara Merkaj, Gabriel Cassinelli Petersen, Arman Avesta, Divya Ramakrishnan, Antonio Omuro, Ali Nabavizadeh, Spyridon Bakas, Khaled Bousabarah, MingDe Lin, Sanjay Aneja, Michael Sabel, and Mariam Aboian

All author affiliations are listed at the end of the article

Corresponding Author: Mariam S. Aboian, MD, PhD, 3401 Civic Center Blvd., Philadelphia, PA 19014, 215-590-7000, USA (mariam.aboian@gmail.com).

Abstract

Background. Glioma, the most prevalent primary brain tumor, poses challenges in prognosis, particularly in the high-grade subclass, despite advanced treatments. The recent shift in tumor classification underscores the crucial role of isocitrate dehydrogenase (IDH) mutation status in the clinical care of glioma patients. However, conventional methods for determining IDH status, including biopsy, have limitations. Exploring the use of machine learning (ML) on magnetic resonance imaging to predict IDH mutation status shows promise but encounters challenges in generalizability and translation into clinical practice because most studies either use single institution or homogeneous datasets for model training and validation. Our study aims to bridge this gap by using multi-institution data for model validation.

Methods. This retrospective study utilizes data from large, annotated datasets for internal (377 cases from Yale New Haven Hospitals) and external validation (207 cases from facilities outside Yale New Haven Health). The 6-step research process includes image acquisition, semi-automated tumor segmentation, feature extraction, model building with feature selection, internal validation, and external validation. An extreme gradient boosting ML model predicted the IDH mutation status, confirmed by immunohistochemistry.

Results: The ML model demonstrated high performance, with an Area under the Curve (AUC), Accuracy, Sensitivity, and Specificity in internal validation of 0.862, 0.865, 0.885, and 0.713, and external validation of 0.835, 0.851, 0.850, and 0.847.

Conclusions. The ML model, built on a heterogeneous dataset, provided robust results in external validation for the prediction task, emphasizing its potential clinical utility. Future research should explore expanding its applicability and validation in diverse global healthcare settings.

Key Points

- Machine learning model predicts IDH mutation in gliomas with a high AUC of 0.862
- Promising results in external validation with area under the curve of 0.835.
- Based on heterogeneous data, the model shows potential clinical utility.

Importance of the Study

In contrast to prior studies demonstrating promising results in IDH mutation prediction, this research addresses a critical gap in their translational applicability. Earlier investigations often relied on meticulously curated datasets, limiting generalizability to routine clinical scenarios. Furthermore, the absence of robust external validation undermines the confidence in the broader utility of these models. Our study stands out by employing a large, heterogeneous dataset reflective of real-world clinical practice, ensuring a more inclusive representation. The success of the ML model validated both internally and externally, accentuates its potential for clinical implementation. By mitigating the limitations of past studies, our research lays a foundation for a more widely applicable and clinically relevant tool for predicting IDH mutation status in gliomas.

Glioma is the most common primary malignant central nervous system (CNS) tumor, with 6 cases of gliomas diagnosed per 100 000 people every year in the United States.[1,2] Despite advanced treatments combining temozolomide with specialized radiotherapy, the prognosis of high-grade gliomas remains limited, with median survival spanning only 10 to 17 months.[3,4] Over the past decade, the World Health Organization (WHO) Classification of Tumors of the CNS experienced a paradigm shift away from specific histologic alterations towards molecular diagnosis of glioma, further reinforced by the recent 2021 classification.[5] This recent WHO classification recognizes isocitrate dehydrogenase (IDH) mutation status as a pivotal marker for glioma classification, therapeutic decision-making, and prognosis[6–8]; hence, it categorizes gliomas into IDH mutant and IDH wild type, which serves as a foundation for determination of their grade.[5] IDH wild-type gliomas have an aggressive and infiltrative appearance on MR imaging with evidence of blood-brain barrier breakdown and they demonstrate very low overall survival.[9,10] Conversely, IDH mutant glioma patients typically have a more favorable prognosis characterized by significantly improved survival than those with wild-type gliomas (31 vs. 15 months)[11,12] and demonstrate heterogeneous imaging characteristics on MRI, such as altering contrast enhancement intensity and tumor necrosis.

In the field of neuro-oncology, imaging plays a crucial role in determining the location of the tumor, defining tumor borders, and providing key information for establishing the diagnosis. Nonetheless, the gold standard for the definitive diagnosis of gliomas occurs via a stereotactic biopsy or tumor resection, followed by histopathological analysis and molecular profiling. This approach, however, has its drawbacks, including the risks of complications, high costs, and potential misclassification due to the spatial heterogeneity of gliomas leading to sampling bias.[13–15] In addition, the necessity for a biopsy of the tumor limits the availability of neoadjuvant therapy trials to patients, which can potentially influence treatment outcomes. Recent studies suggest that targeted therapy of IDH mutated gliomas is gaining interest and might be an increasingly important staple of glioma treatment in the future.[16] Some studies have also reported that the reliability of IDH mutation testing through pathology is challenged by technical limitations.[17–19]

Considering the pitfalls of conventional methods for determining the IDH mutation status, magnetic resonance imaging (MRI) emerges as a valuable tool in this context by offering a noninvasive technique for the preoperative evaluation of IDH mutation, aiding in directing targeted biopsies, opening consideration for neoadjuvant therapy, screening for potential genetic mutations, and tailoring treatment strategies.[20] However, IDH mutation status lacks a distinct radiological imaging profile.[21] IDH mutant gliomas might exhibit characteristics like T2-FLAIR mismatch sign,[22] reduced enhancement, lower blood flow on perfusion-weighted images, increased mean diffusion values, smaller sizes, and a tendency to occur in the frontal lobe.[23] Despite these trends, the sensitivity and specificity for identification of IDH mutation by MRI ranges from 56% to 100% and from 51% to 100%, respectively.[24] In cases with a positive T2-FLAIR mismatch sign, the specificity has been reported to be 100%. Hence, there is a growing need for sensitive and specific techniques to accurately predict IDH mutation status in gliomas through MRI.

Recent advances in machine learning (ML) show promising results in several tasks such as tumor segmentation,[25] predicting overall survival,[26] differentiating gliomas from other CNS malignancies such as metastases[27] and lymphomas,[28] as well as classifying glioma grades[29] and molecular subtypes.[15,30] While much of the literature in neuro-oncology that applies ML techniques demonstrates high area under the curve (AUC) values within internal datasets, the application of these models to external validation sets remains limited.[15]

A significant concern in current research is the reliance on small, meticulously selected ("cherry-picked") datasets without artifacts and well-curated imaging sequences, which poses challenges in applying these findings to the more diverse and complex data encountered in everyday clinical practice. A recent systematic review of ML algorithms for identification of IDH mutation by MRI demonstrated AUC and accuracy of 0.88 and 85% in internal validation and 0.86 and 87% in very limited external validation data sets, reporting studies with incomplete data and high bias.[15] Recognizing these limitations, our study aims to bridge this gap.

By employing a large, annotated dataset encompassing external validation and heterogeneous data reflective of routine clinical scenarios, we aim to set a new standard for the clinical application of ML in neuro-oncology. Given the pivotal role of IDH in both clinical management and prognosis of glioma patients, our approach not only aligns with the current academic focus on noninvasive IDH prediction

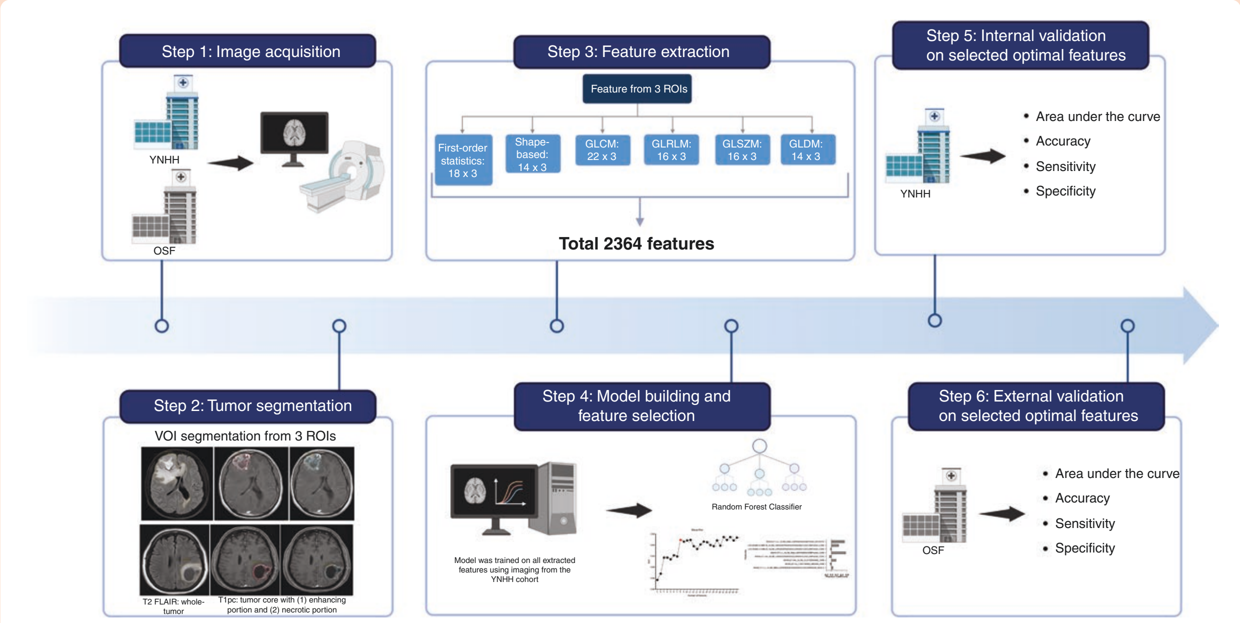
Step 1: Image acquisition
YNHH
OSF
Step 2: Tumor segmentation
VOI segmentation from 3 ROIs
T2 FLAIR: whole-tumor
T1pc: tumor core with (1) enhancing portion and (2) necrotic portion
Step 3: Feature extraction
Feature from 3 ROIs
First-order statistics: 18 x 3
Shape-based: 14 x 3
GLCM: 22 x 3
GLRLM: 16 x 3
GLSZM: 16 x 3
GLDM: 14 x 3
Total 2364 features
Step 4: Model building and feature selection
Model was trained on all extracted features using imaging from the YNHH cohort
Random Forest Classifier
Step 5: Internal validation on selected optimal features
YNHH
Area under the curve
Accuracy
Sensitivity
Specificity
Step 6: External validation on selected optimal features
OSF
Area under the curve
Accuracy
Sensitivity
Specificity

Figure 1. Proposed pipeline for IDH mutation status prediction. A graphical illustration of the overall pipeline which includes 6 parts: (1) image acquisition; (2) tumor segmentation; (3) feature extraction; (4) model building and feature selection; (5) internal validation (including model retraining); and (6) external validation. Abbreviations: YNHH, Yale New Haven Hospital; OSF, outside facility; VOI, volume of interest; ROI, region of interest; GLCM, gray-level co-occurrence matrix; GLRLM, gray-level run-length matrix; GLSZM, gray-level size zone matriX; GLDM, gray-level dependence matrix; T1pc, T1 post-contrast enhancement; FLAIR, fluid-attenuated inversion recovery.

but also contributes valuable insights into the real-world applicability of ML in clinical settings. Furthermore, our findings, grounded in a comprehensive and practical dataset, promise to provide a more relatable and transferable benchmark for comparison with ongoing academic efforts in this rapidly evolving field.

Methods

Study Design

This retrospective study was conducted in alignment with the Helsinki Declaration and approved in November 2021 by the Yale New Haven Hospital (YNHH) Institutional Review Board at our institute (IRB protocol ID 2000029055).

Figure 1 shows the overall workflow of our approach, which includes 6 parts: (1) image acquisition; (2) tumor segmentation; (3) feature extraction; (4) model building and feature selection; (5) internal validation (including model retraining); and (6) external validation. We used the Checklist for Evaluation of Radiomic Research (CLEAR), which can be found in Supplementary Material Appendix 1.[31]

Subjects

Patients were categorized based on the location where their imaging scans were performed. Specifically, those who underwent imaging at YNHH were assigned as such and utilized for training and as an internal validation dataset. In contrast, patients whose imaging was conducted at national hospitals outside of YNHH were categorized under the "outside facility" (OSF) group and reserved for external validation, effectively serving as our testing cohort. This stratification strategy was critical in ensuring the robustness and generalizability of our predictive model, as it allowed for the evaluation of the model's performance in diverse clinical settings.

A total of 1031 patients (652 YNHH and 379 OSF) with available imaging and pathological data from January 2012 to December 2019 were initially collected in our glioma database, with clinical information collected from our hospital's radiation oncology registry. The necessary MRI and pathology data were extracted from the hospital's electronic medical records.

To develop an ML pipeline to specifically predict IDH mutation status in glioma, patients were selected based on the following criteria: [I] grades 2–4 diffuse glioma according to the 2021 WHO classification criteria with histopathologically confirmed IDH mutation status, and [2] availability of conventional pretreatment MRI scans consisting at least of both contrasts enhanced T1 weighted imaging (T1pc) and T2 fluid-attenuated inversion recovery (FLAIR). Patients were excluded according to the following criteria: [I] absence of IDH mutation status ($n = 252$), [II] absence of T1pc or T2 FLAIR ($n = 89$), and [III] WHO grade 1 tumor ($n = 20$) or absence of information about tumor grade ($n = 93$).

Figure 2 illustrates the flow diagram of the study population.

Finally, data from 584 patients (219 females and 365 males; age range 5–95) met the inclusion criteria. The cohort was split into 377 cases from YNHH used for model training, and

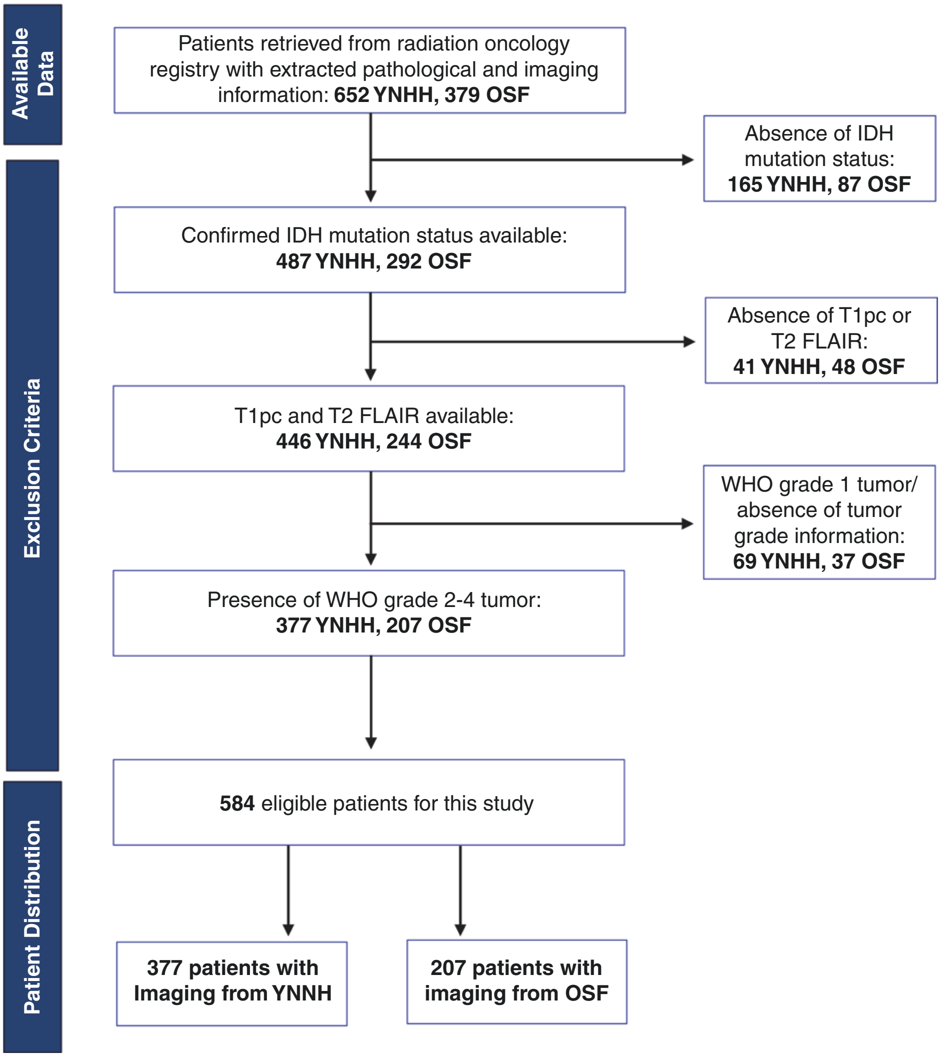
Available Data
Exclusion Criteria
Patient Distribution
Patients retrieved from radiation oncology registry with extracted pathological and imaging information: 652 YNHH, 379 OSF
Absence of IDH mutation status: 165 YNHH, 87 OSF
Confirmed IDH mutation status available: 487 YNHH, 292 OSF
Absence of T1pc or T2 FLAIR: 41 YNHH, 48 OSF
T1pc and T2 FLAIR available: 446 YNHH, 244 OSF
WHO grade 1 tumor/ absence of tumor grade information: 69 YNHH, 37 OSF
Presence of WHO grade 2-4 tumor: 377 YNHH, 207 OSF
584 eligible patients for this study
377 patients with Imaging from YNNH
207 patients with imaging from OSF

Figure 2. Patient selection workflow. Patient information was retrieved from the radiation oncology registry and screened for imaging and pathological information in the hospital's electronic medical records. Abbreviations: YNHH, Yale New Haven Hospital; OSF, outside facility; IDH, isocitrate dehydrogenase; T1pc, contrast-enhanced T1 weighted imaging; FLAIR, fluid-attenuated inversion recovery; WHO, World Health Organization.

207 cases under the OSF group for external validation of the prediction pipeline. The clinical characteristics of the selected patients from YNHH and OSF are summarized in Table 1.

Imaging Data Protocol

The incorporation of data from both YNHH-affiliated hospitals and external facilities introduces a diversity of imaging protocols within the ML pipeline. The predominant imaging protocol involves the utilization of 3T imaging systems from Siemens Vervio (Siemens Healthineers AG). Specifically, for FLAIR imaging, the parameters include TE (Echo Time) ranging from 82 to 112 milliseconds, TR (Repetition Time) set at 9000 milliseconds, Inversion Time of 2500 milliseconds, and a slice thickness of 4–5 mm. T1-weighted images were acquired using 2 different pulse sequences: Post Gadolinium Gradient Echo (GRE) with TE ranging from 2.48 to 3.09 milliseconds, TR between 1600 and 1900 milliseconds, Inversion Time set at 900 milliseconds, and a slice thickness of 0.9–1 mm. Post Gadolinium Spin Echo (PGSE) with TE varying from 2.48 to 9.3 milliseconds, TR spanning 277 to 5000 milliseconds, Inversion Time set at 900 milliseconds, and a slice thickness ranging from 1 to 5 mm. Notably, gadolinium-based contrast agents were administered for the T1-weighted images. DICOM metadata extraction was facilitated through the application programming interface of the PACS provider (Visage 7 by Visage Inc.).

Tumor Segmentation

For accurate tumor delineation, we employed a U-Net Deep Learning segmentation tool, integrated into our local PACS System.[32] The UNETR was trained using volumetric patches of size $128 \times 128 \times 64$ with a spacing of $1.5 \times 1.5 \times 2\ \text{cm}^3$

Table 1. Clinical Characteristics of All Included Patients ($N = 584$)

Variable	YNHH ($n = 377$)	OSF ($n = 207$)	P-value
Age (years), mean ± SD	58.3 ± 17.5	58.8 ± 15.5	.722[a]
Sex, n (%)			.982[b]
Female	142 (37.7)	77 (37.2)	
Male	235 (62.3)	130 (62.8)	
IDH mutation status, n (%)			.653[b]
Mutated	43 (11.4)	27 (13.0)	
Wild-type	334 (88.6)	180 (87.0)	
Tumor enhancing status, n (%)			.059[b]
Enhancing	323 (85.7)	164 (79.2)	
Non-enhancing	54 (14.3)	43 (20.8)	
WHO tumor grade, n (%)			.589[b]
Grade 2	29 (7.7)	21 (10.1)	
Grade 3	43 (11.4)	24 (11.6)	
Grade 4	305 (80.9)	162 (78.3)	
Imaging sequences, n (%)			
FLAIR, PGSE, GRE	279 (74.0)	91 (44.0)	
FLAIR, PGSE	30 (8.0)	94 (45.4)	
FLAIR, GRE	68 (18.0)	22 (10.6)	

a: Assessed using an independent samples t-test.
b: Analyzed using chi-squared tests.
Abbreviations: YNHH, Yale New Haven Hospital; OSF, outside facility; IDH, isocitrate dehydrogenase; WHO, World Health Organization; FLAIR, fluid-attenuated inversion recovery; PGSE, pulsed gradient spin echo contrast-enhanced T1 weighted imaging; GRE, gradient echo contrast-enhanced T1 weighted imaging.

using trilinear interpolation with sampling of the patches. The z-direction was resampled from 1 to 2, compared to a conventional Convolutional Neural Network, which resulted in a batch size of 128 × 128 × 64. The UNETR tool automatically segments glioma tumors on FLAIR imaging, using a combination ofT1pc and FLAIR sequences as inputs. The segmentation process identified 3 key volumes of interest (VOI): whole tumor portion (including peritumoral edema) on FLAIR, tumor core (encompassing necrotic and enhancing portions), and exclusively necrotic tumor portion. These segmented VOIs were subsequently mapped onto theT1pc sequences.

Manual adjustments were made by a neuroradiologist to the segmented VOIs to ensure precision, particularly in challenging cases with complex tumor morphology. As part of our inclusion criteria, FLAIR segmentation with peritumoral edema was present in all the cases; however, core and necrotic segmentation onT1pc were present in 487 (83.4%) and 432 (74.0%) cases, respectively. This segmentation methodology aligns with the latest advancements in glioma MRI analysis, ensuring high fidelity in tumor delineation crucial for subsequent feature extraction.

Feature Extraction

Radiomic features were preprocessed and extracted using an open-source Python tool Pyradiomics (version 3.0.1).[33] These features encompassed 3 primary categories: (1) volume and shape, (2) intensity, and (3) texture. To further enhance the feature set, advanced image processing techniques were employed. Retrieval performance with the inclusion of 8 high- and low-pass wavelet filters for each of these feature classes was applied to the original images to generate transform-domain images. The following original Pyradiomics features were extracted from each of the 3 VOI, if present: 18 first-order statistics, 14 shape-based features, 22 gray-level co-occurrence matrices, 16 gray-level run length matrices, 16 gray-level size zone matrices, and 14 gray-level dependence matrices. Finally, 2364 features were extracted from the MR imaging of each patient.

ML Pipeline

Our ML approach aims to reliably predict IDH mutation from the extracted radiomic features of MR Images. Our binary prediction pipeline involves an extreme gradient boosting (XGBoost) algorithm for model training.[34] In the preprocessing phase, categorical features are converted, and numerical features are normalized using MinMaxScaler from the scikit library. Additionally, correlated features were removed using a custom preprocessing transformer. The resulting features are used in our XGBoost pipeline.

To address the heterogeneous nature of our data, we handle class imbalance using SMOTE (Synthetic Minority Over-sampling Technique), a widely used oversampling technique. The pipeline comprises a 5-fold cross-validation with 10 iterations, ensuring robust model evaluation. The reproducibility seed is set to 123 to ensure consistent results across runs. During the model training process, we employ Randomized SearchCV for hyperparameter tuning. The features used for prediction are then ranked by their median importance for binary classification. The best 25 features are visualized in an elbow plot, where diminishing returns are observed to find the optimal number of features for this model. The Pearson Coefficient is then analyzed for the optimal number of features. The resulting features are then used to re-train the XGBoost model.

The re-trained model is then used for external validation. The data used for external validation from outside hospitals is preprocessed in the same way as the training cohort. Lastly, the pipeline the performance metrics mean AUC, accuracy, sensitivity, and specificity with their respective 95% confidence interval (CI) for each of the 3 steps (Internal validation, internal validation after feature selection, and external validation).

IDH Mutation Assessment

The IDH mutation status was retrieved from our institution's internal electronic medical record pathology reports. Tumor samples were biopsied from a region in the brain that had been confirmed through imaging at YNHH. Specimens with a minimum of 50% of tumor cells in a microdissection target were accepted for analysis. The initial determination of IDH mutation status for all samples involved the utilization of immunohistochemistry (IHC), employing a specific clinically validated antibody (DIA-H09; Dianova GmbH, Hamburg, Germany) targeting the IDH1 R132H mutation.[35] For cases in which IDH1 R132H mutation was not detected via IHC, and patients were younger at the time of diagnosis, bidirectional Sanger sequencing was performed using the BigDye Terminator Kit on ABI3130/3730 (Applied Biosystems).

Statistical Analysis

Data for this study were organized and recorded in a Microsoft Excel spreadsheet (Version 16.81), utilizing a combination of descriptive statistics and tabulated presentations. Statistical analyses were performed to derive insights from the collected patient data, employing both the student's *t*-test for comparing means and the chi-square test for assessing associations between categorical variables. Statistical significance was determined by *P*-values < .05. Analysis was conducted using GraphPad Prism software (Version 10.0.03, GraphPad Software, LLC).

Results

Clinical Characteristics of Patients

The study evaluated the clinical characteristics of 584 patients (Table 1), divided into 2 cohorts: training and internal validation YNNH group with 377 patients (37.7% females and 62.3% males) and external validation OSF group with 207 patients (37.2% females and 62.8% males). The mean age for the YNHH cohort was 58.3 ± 17.46 years, while for the OSF cohort, it was 58.78 ± 15.45 years. IDH mutation status revealed 43 (11.4%) and 334 (88.6%) patients of the YNHH group with IDH mutant (IDHmut) and IDH wild type (IDHwt) gliomas, respectively. Similarly, the OSF group comprised 27 (13.0%) and 180 (87.0%) patients with IDHmut and IDHwt gliomas, respectively. The majority of the patients in the YNHH group ($n = 323$, 85.7%) and OSF group ($n = 164$, 79.2%) had an enhancing tumor on T1pc. In terms of imaging sequences, most patients ($n = 279$, 74.0%) in the YNHH cohort had FLAIR, PGSE, and GRE sequences, whereas this combination was less common in the OSF group ($n = 91$, 44.0%). The FLAIR and PGSE combination were available for 30 (8.0% of YNHH group) 94 (45.4% of OSF group) patients, and the FLAIR and GRE combination for 68 (18.0% of YNHH group), and 22 (10.6% of OSF group) of patients.

There were no statistically significant differences in terms of clinical factors such as age ($P = .722$), sex ($P = .982$), IDH mutation status ($P = .653$), tumor enhancing status ($P = 0.059$), WHO tumor grade ($P = 0.589$) between the 2 groups.

Feature Selection

After extracting 2364 radiomic features from the imaging sequences, we employed an elbow plot analysis to determine the optimal number of features. This analysis revealed that the model's predictive performance, as measured by the AUC, reached an optimal point at 8 features, where the inclusion of additional features resulted in negative or diminishing returns (Figure 3).

Further examination of these 8 features using the Pearson correlation coefficient analysis demonstrated varying degrees of correlation with the outcome variable, with some features showing stronger associations than others (Figure 4). The 8 features selected through this process consist of 7 textures and 1 intensity feature. The detailed characteristics and implications of these 8 radiomic features are thoroughly documented in Supplementary Appendix 2.

Performance of the Classification Model

Table 2 shows the classification results of the different combinations of sequences used in the model. We compared different runs using the available segmentation sequences. The combined approach of "Whole_Flair + Core + Necrotic" sequences showed notable performance metrics across various validation phases and achieved the highest overall performance in external validation (Table 2). Therefore, this combination will be the focus of our analysis. In the training phase, the model achieved an accuracy (ACC) of 0.885 and a sensitivity (SEN) of 0.926, indicating high efficacy in correctly identifying positive cases. The specificity (SPEC), however, was relatively lower at 0.562, suggesting a moderate rate of correctly identifying negative cases. The AUC was 0.846, with a 95% CI (0.827, 0.865),

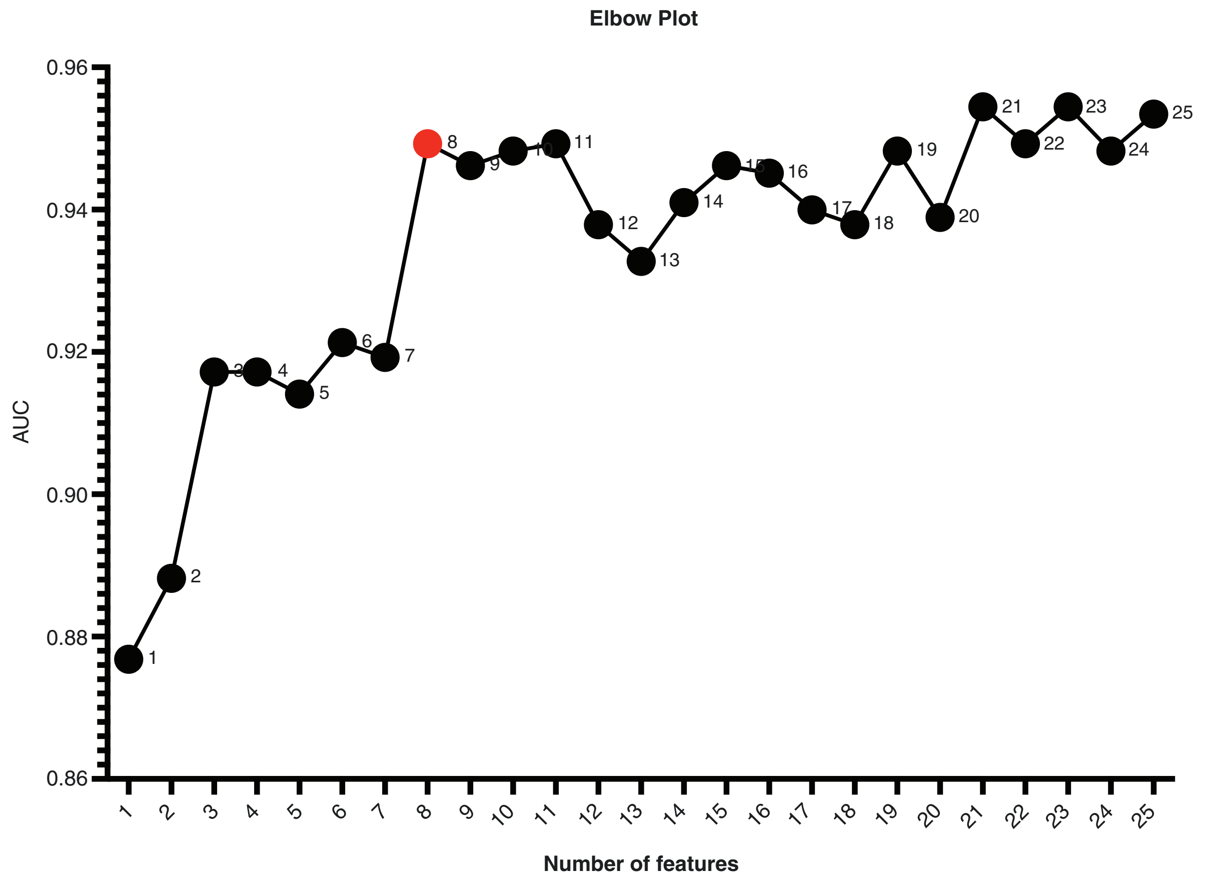

Figure 3. Elbow plot. Elbow plot visualizing the top 25 features ranked by median importance score. Eight is the point of optimal feature selection, as additional features stagnate performance and result in negative or diminishing returns.

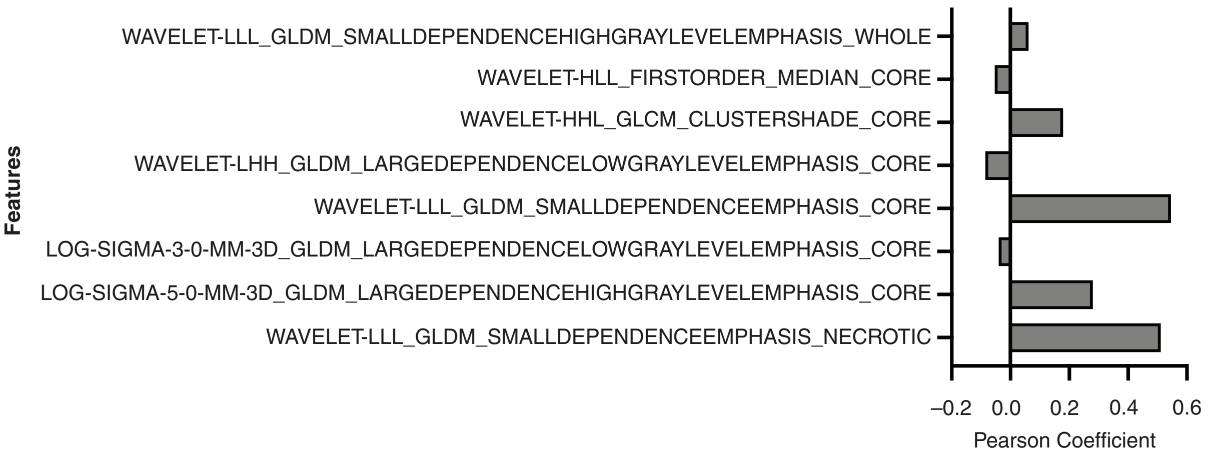

Figure 4. Extracted optimal features and weights. Abbreviations: LLL, low low subband filter; GLDM, gray-level dependence matrix; GLCM, gray-level co-occurrence matrix.

reflecting strong model performance. During the internal validation phase, there was a slight decrease in ACC to 0.865 and in SEN to 0.885, but an increase in SPEC to 0.713, indicating improved identification of negative cases. The AUC remained high at 0.862, with a 95% CI (0.842, 0.881). In the external validation phase, the model's performance metrics further evolved, with the ACC slightly decreasing to 0.851 and SEN to 0.850. However, SPEC saw a significant

Table 2. Evaluation Metrics in the Training and Validation Phases

Sequences	Phase	ACC	SEN	SPEC	AUC	95% CI
Whole FLAIR + Core + Necrotic	Training	0.885	0.926	0.562	0.846	(0.827, 0.865)
	Internal validation	0.865	0.885	0.713	0.862	(0.842, 0.881)
	External validation	0.851	0.850	0.847	0.835	(0.698, 0.978)
Core + Necrotic	Training	0.885	0.910	0.684	0.857	(0.836, 0.876)
	Internal validation	0.872	0.891	0.733	0.891	(0.877, 0.905)
	External validation	0.821	0.833	0.733	0.805	(0.706, 0.884)
Whole FLAIR + Necrotic	Training	0.871	0.908	0.591	0.842	(0.823, 0.860)
	Internal validation	0.862	0.872	0.786	0.872	(0.854, 0.889)
	External validation	0.739	0.744	0.707	0.749	(0.646, 0.906)
Whole FLAIR + Core	Training	0.891	0.928	0.594	0.859	(0.839, 0.878)
	Internal validation	0.850	0.864	0.739	0.864	(0.842, 0.883)
	External validation	0.792	0.789	0.813	0.814	(0.701, 0.973)

Each modality had a different number of optimal features since the optimal number of features suggested by the elbow plot varies from the data used for prediction. AUC, area under the curve.

improvement to 0.847, aligning more closely with the SEN. The AUC was 0.835, within a broader 95% CI (0.698, 0.978). Also, the F1-Score, generally seen as more biased resistant than conventional performance metrics, was noticeably high with 0.908 (Supplementary Appendix 3).

Discussion

The development of our ML model to predict IDH mutation status in glioma represents a crucial step in developing personalized medicine approaches in patients who are treated for these aggressive tumors with limited therapeutic options. The primary focus of our research is the development of an ML model on a clinically relevant dataset and the external validation of the ML model in a heterogeneous dataset. This approach is vital for ensuring the model's clinical applicability and generalizability.[36,37] The high sensitivity and specificity achieved in our model, particularly in the external validation phase, highlight its potential for clinical deployment. Such performance is crucial for models intended for real-world clinical use, as they must reliably identify both IDH mutant and wild-type gliomas in diverse patient populations.[38] Especially in regard to clinical decision-making, aiding in targeted therapy of IDH mutated gliomas, robust predictive models might be of rising interest in the future.[16]

Our study's external validation in a diverse clinical setting aligns with recent trends in healthcare modeling, emphasizing the importance of model generalization beyond the initial development environment.[39,40]The systematic allocation of patients into distinct cohorts for internal and external validation addresses the critical need for models that can adapt to varied clinical scenarios,[41,42] especially in a clinical context where diagnostic accuracy is paramount.[43] This stratification strategy, as seen in our study, enhances the robustness of the predictive model, a key aspect that has been highlighted in recent studies.[44,45]The application of our model to an external dataset has provided insights into its performance variability, an aspect often overlooked in ML studies.

The model's high sensitivity in the training phase and its gradual improvement in specificity from the training to the external validation phase emphasize its potential in medical diagnostics. The slight decrease in accuracy and sensitivity from the training to the external validation phase might be attributed to the model's exposure to a more diverse set of data, a common occurrence in ML models.[46] Contrary to expectations based on prior studies, the specificity was higher in external validation compared to internal validation, which may stem from the relatively smaller size of our external validation set.[15]

Moreover, the broader CI in the external validation phase, particularly for the AUC, indicates variability in the model's performance across different external datasets. This variability could be due to differences in data quality, distribution, or other external factors.[47] Such findings highlight the need for further investigation into the factors influencing the model's performance in diverse clinical settings.

Our research demonstrates the model's high efficacy in identifying positive cases with notable accuracy and sensitivity. However, the variability observed in the external validation phase underscores the challenges faced in ML when models are applied to diverse datasets.[36]This finding is crucial for medical diagnostics, as it highlights the need for models that are not only accurate but also adaptable to different patient populations and imaging conditions.

The inclusion criteria, based on the latest WHO classification, along with rigorous imaging data acquisition and tumor segmentation processes within PACS, have contributed to the model's accuracy. The use of advanced radiomic feature extraction, including shape, intensity, and texture characteristics, parallels the methodologies employed in similar studies.[48,49] The selection of these features using an elbow plot analysis has optimized the model's performance, as evidenced by its AUC scores in both the internal and external validation phases.

The integration of multi-modal MRI data, as employed in our study, is in line with the current best practices in ML for healthcare applications.[39,45]The use of advanced radiomics features and a U-Net Deep Learning segmentation tool reflects the state-of-the-art in medical image analysis, ensuring high fidelity in tumor delineation crucial for subsequent feature extraction.[41,44]

Future research should focus on expanding the model's applicability to other glioma subtypes and incorporating additional clinical variables for more comprehensive predictions. The integration of genomic data, alongside radiomic features, could further enhance the model's predictive capabilities.[50] Additionally, ongoing validation with larger and more varied external datasets will be essential to continually assess and refine the model's performance.

Limitations

Our study, with its retrospective nature, encounters several limitations. The heterogeneity of the external dataset, particularly in terms of imaging protocols and equipment, may have introduced variability impacting model performance. The sample size, although substantial, may not capture the full spectrum of clinical scenarios, limiting the generalizability of our findings. Additionally, the complexity of the model might challenge clinical interpretability, and the absence of longitudinal data limits insights into long-term efficacy. The study's focus on radiomic features, excluding broader clinical and genomic data, potentially restricts its comprehensive applicability. Finally, further validation in diverse global healthcare settings is necessary to confirm the model's universal effectiveness.

Conclusion

Our study contributes significantly to the field of neuro-oncology by providing a robust, externally validated ML model for predicting IDH mutation status in gliomas. Its successful validation across heterogeneous patient cohorts and clinical settings lays the groundwork for its potential clinical implementation, offering a promising tool for personalized patient management in glioma treatment. The implications of this research extend to potentially improving patient outcomes and informing treatment strategies, marking a pivotal step towards more tailored and effective neuro-oncological care.

Supplementary material

Supplementary material is available online at *Neuro-Oncology Advances* (https://academic.oup.com/noa).

Keywords:

gliomas | machine learning | MRI | neuro-oncology

Funding

JL received funding from the doctoral scholarship of Heinrich-Heine-University. American Society of Neuroradiology Fellow Award 2018 (MA). This publication was made possible by KL2 TR001862 (MA) from the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH. Arman Avesta is a PhD Student in the Investigative Medicine Program at Yale which is supported by CTSA Grant Number UL1 TR001863 from the National Center for Advancing Translational Science, a component of the National Institutes of Health (NIH). Arman holds securities in Hyperfine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. MingDe Lin is an employee and stockholder of Visage Imaging, Inc., and unrelated to this work, receives funding from NIH/NCI R01 CA206180 and is a board member of Tau Beta Pi Engineering Honor Society.

Conflict of interest statement

MingDe Lin is an employee and stockholder of Visage Imaging, Inc., and unrelated to this work, receives funding from NIH/NCI R01 CA206180 and is a board member of Tau Beta Pi Engineering Honor Society. Khaled Bousabarah is an employee of Visage Imaging, GmbH. Michael Sabel is a consultant for Novocure and Codman.

Authorship statement

J.L. conceptualized the study framework, collected/ analyzed data, and prepared the primary manuscript. N.A. contributed to data analysis, and the conception of the study, and prepared the primary manuscript. L.J., M.vR., and K.W. contributed to data analysis, data collection, and manuscript revisions. N.T., S.M., and G.C.P. contributed to data collection and manuscript revisions. A.A. and S.A. contributed to data analysis and manuscript revisions. D.R., A.O., A.N., S.B., K.B., and M.L. contributed to manuscript revisions. M.S. contributed to the study framework and manuscript revisions. M.A. was the primary supervisor of the project and contributed to the study framework, data collection, and manuscript revision.

Affiliations

Department of Neurosurgery, Heinrich-Heine University, Dusseldorf, Germany (J.L.); College of Medicine, Alfaisal University, Riyadh, Saudi Arabia (N.A.); DKFZ Division of Translational Neurooncology at the WTZ, German Cancer Consortium, DKTK Partner Site, University Hospital Essen, Essen, Germany (L.J.); University of Leipzig, Leipzig, Germany

(M.R.); Department of Diagnostic and Interventional Radiology, Medical Faculty, University Dusseldorf, Dusseldorf, Germany (N.T.); University of Leipzig, Leipzig, Germany (K.W.); University of Ulm, Ulm, Germany (S.M.); University of Göttingen, Göttingen, Germany (G.C.P.); Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts, USA (A.A.); Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, Connecticut, USA (D.R.); Department of Neurology and Yale Cancer Center, Yale School of Medicine, New Haven, Connecticut, USA (A.O.); Department of Radiology, Perelman School of Medicine, Hospital of University of Pennsylvania, University of Pennsylvania, Philadelphia, Pennsylvania, USA (A.N.); Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA (S.B.); Visage Imaging, Inc., Berlin, Germany (K.B.); Visage Imaging, Inc., San Diego, California, USA (M.D.L.); Department of Therapeutic Radiology, Yale School of Medicine, New Haven, Connecticut, USA (S.A.); Department of Neurosurgery, Heinrich-Heine-University, Duesseldorf, Germany (M.S.); Department of Radiology, Children's Hospital of Philadelphia (CHOP), Philadelphia, Pennsylvania, USA (M.A.)

References

1. Hanif F, Muzaffar K, Perveen K, Malhi SM, Simjee Sh U. Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical presentation and treatment. *Asian Pac J Cancer Prev.* 2017;18(1):3–9.
2. Mesfin FB, Karsonovich T, Al-Dhahir MA. *Gliomas.* In: *StatPearls [Internet].* Treasure Island, FL: StatPearls Publishing; 2024.
3. Mohammed S, Dinesan M, Ajayakumar T. Survival and quality of life analysis in glioblastoma multiforme with adjuvant chemoradiotherapy: A retrospective study. *Rep Pract Oncol Radiother.* 2022;27(6):1026–1036.
4. Molinaro AM, Taylor JW, Wiencke JK, Wrensch MR. Genetic and molecular epidemiology of adult diffuse glioma. *Nat Rev Neurol.* 2019;15(7):405–417.
5. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro Oncol.* 2021;23(8):1231–1251.
6. Peng H, Huo J, Li B, et al. Predicting isocitrate dehydrogenase (IDH) mutation status in gliomas using multiparameter MRI radiomics features. *J Magn Reson Imaging.* 2021;53(5):1399–1407.
7. Wen PY, Kesari S. Malignant gliomas in adults. *N Engl J Med.* 2008;359(5):492–507.
8. Giantini-Larsen AM, Abou-Mrad Z, Yu KK, et al. Treatment and outcomes of IDH1-mutant gliomas in elderly patients. *J Neurosurg.* 2023;140(2):367–376.
9. Dang L, White DW, Gross S, et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature.* 2009;462(7274):739–744.
10. Zhang CB, Bao ZS, Wang HJ, et al. Correlation of IDH1/2 mutation with clinicopathologic factors and prognosis in anaplastic gliomas: A report of 203 patients from China. *J Cancer Res Clin Oncol.* 2014;140(1):45–51.
11. Han S, Liu Y, Cai SJ, et al. IDH mutation in glioma: Molecular mechanisms and potential therapeutic targets. *Br J Cancer.* 2020;122(11):1580–1589.
12. 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.
13. Parker NR, Khong P, Parkinson JF, Howell VM, Wheeler HR. Molecular heterogeneity in glioblastoma: Potential clinical implications. *Front Oncol.* 2015;5:55.
14. Friedmann-Morvinski. Glioblastoma heterogeneity and cancer cell plasticity. 2014.
15. Lost J, Verma T, Jekel L, et al. Systematic literature review of machine learning algorithms using pretherapy radiologic imaging for glioma molecular subtype prediction. *AJNR Am J Neuroradiol.* 2023;44(10):1126–1134.
16. Miller JJ, Gonzalez Castro LN, McBrayer S, et al. Isocitrate dehydrogenase (IDH) mutant gliomas: A Society for Neuro-Oncology (SNO) consensus review on diagnosis, management, and future directions. *Neuro Oncol.* 2023;25(1):4–25.
17. Cryan JB, Haidar S, Ramkissoon LA, et al. Clinical multiplexed exome sequencing distinguishes adult oligodendroglial neoplasms from astrocytic and mixed lineage gliomas. *Oncotarget.* 2014;5(18):8083–8092.
18. Hegi ME, Diserens A-C, Gorlia T, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med.* 2005;352(10):997–1003.
19. Gutman DA, Dunn WD, Jr, Grossmann P, et al. Somatic mutations associated with MRI-derived volumetric features in glioblastoma. *Neuroradiology.* 2015;57(12):1227–1237.
20. Fathi Kazerooni A, Bakas S, Saligheh Rad H, Davatzikos C. Imaging signatures of glioblastoma molecular characteristics: A radiogenomics review. *J Magn Reson Imaging.* 2020;52(1):54–69.
21. Johnson DR, Guerin JB, Giannini C, et al. 2016 updates to the WHO brain tumor classification system: What the radiologist needs to know. *Radiographics.* 2017;37(7):2164–2180.
22. 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.
23. Chow D, Chang P, Weinberg BD, et al. Imaging genetic heterogeneity in glioblastoma and other glial tumors: Review of current methods and future directions. *AJR Am J Roentgenol.* 2018;210(1):30–38.
24. Suh CH, Kim HS, Jung SC, Choi CG, Kim SJ. Imaging prediction of isocitrate dehydrogenase (IDH) mutation in patients with glioma: A systemic review and meta-analysis. *Eur Radiol.* 2019;29(2):745–758.
25. Niklas Tillmanns AEL, Cassinelli G, Merkaj S, et al. Identifying clinically applicable machine learning algorithms for glioma segmentation: Recent advances and discoveries. *Neurooncol. Adv.* 2022;4(1):vdac093.
26. Bakas S, Shukla G, Akbari H, et al. Overall survival prediction in glioblastoma patients using structural magnetic resonance imaging (MRI): Advanced radiomic features may compensate for lack of advanced MRI modalities. *J Med Imaging (Bellingham).* 2020;7(3):031505.
27. Jekel L, Brim WR, von Reppert M, et al. Machine Learning applications for differentiation of glioma from brain metastasis—A systematic review. *Cancers.* 2022;14(6):1369.
28. Cassinelli Petersen GI, Shatalov J, Verma T, et al. Machine learning in differentiating gliomas from primary CNS lymphomas: A systematic review, reporting quality, and risk of bias assessment. *AJNR Am J Neuroradiol.* 2022;43(4):526–533.
29. Bahar RC, Merkaj S, Cassinelli Petersen GI, et al. Machine learning models for classifying high- and low-grade gliomas: A systematic review and quality of reporting analysis. *Front Oncol.* 2022;12:856231.
30. Jian A, Jang K, Manuguerra M, et al. Machine learning for the prediction of molecular markers in glioma on magnetic resonance imaging: A systematic review and meta-analysis. *Neurosurgery.* 2021;89(1):31–44.
31. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): A step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging.* 2023;14(1):75.
32. Aboian M, Bousabarah K, Kazarian E, et al. Clinical implementation of artificial intelligence in neuroradiology with development of a novel

workflow-efficient picture archiving and communication system-based automated brain tumor segmentation and radiomic feature extraction. Original Research. *Front Neurosci.* 2022;16:16:860208.

33. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 2017;77(21):e104–e107.
34. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, California, USA. doi: 10.1145/2939672.2939785
35. Capper D, Weißert S, Balss J, et al. Characterization of R132H mutation-specific IDH1 antibody binding in brain tumors. *Brain Pathol.* 2010;20(1):245–254.
36. Qin Y, Alaa A, Floto A, Schaar MV. External validity of machine learning-based prognostic scores for cystic fibrosis: A retrospective study using the UK and Canadian registries. *PLOS Digit Health.* 2023;2(1):e0000179.
37. Youssef A, Pencina M, Thakur A, et al. External validation of AI models in health should be replaced with recurring local validation. *Nat Med.* 2023;29(11):2686–2687.
38. Du P, Liu X, Wu X, et al. Predicting histopathological grading of adult gliomas based on preoperative conventional multimodal MRI Radiomics: A Machine learning model. *Brain Sci.* 2023;13(6):912.
39. de Hond AAH, Kant IMJ, Fornasa M, et al. Predicting readmission or death after discharge from the ICU: External validation and retraining of a machine learning model. *Crit Care Med.* 2023;51(2):291–300.
40. Chen TL-W, Buddhiraju A, Seo HH, et al. Internal and external validation of the generalizability of machine learning algorithms in predicting non-home discharge disposition following primary total knee joint arthroplasty. *J Arthroplasty.* 2023;38(10):1973–1981.
41. Kunze KN, Kaidi A, Madjarova S, et al. External validation of a machine learning algorithm for predicting clinically meaningful functional improvement after arthroscopic hip preservation surgery. *Am J Sports Med.* 2022;50(13):3593–3599.
42. Mari T, Asgard O, Henderson J, et al. External validation of binary machine learning models for pain intensity perception classification from EEG in healthy individuals. *Sci Rep.* 2023;13(1):242.
43. Li G, Li L, Li Y, et al. An MRI radiomics approach to predict survival and tumour-infiltrating macrophages in gliomas. *Brain.* 2022;145(3):1151–1161.
44. Luo AL, Ravi A, Arvisais-Anhalt S, et al. Development and internal validation of an interpretable machine learning model to predict readmissions in a United States healthcare system. *Informatics.* 2023;10(2):33.
45. Huang C-Y, Güiza F, Wouters P, et al. Development and validation of the creatinine clearance predictor machine learning models in critically ill adults. *Crit Care.* 2023;27(1):272.
46. Liu Y, Zheng Z, Wang Z, et al. Using radiomics based on multicenter magnetic resonance images to predict isocitrate dehydrogenase mutation status of gliomas. *Quant Imaging Med Surg.* 2023;13(4):2143–2155.
47. Dal Bo M, Polano M, Ius T, et al. Machine learning to improve interpretability of clinical, radiological and panel-based genomic data of glioma grade 4 patients undergoing surgical resection. *J Transl Med.* 2023;21(1):450.
48. Kumar A, Jha AK, Agarwal JP, et al. Machine-learning-based radiomics for classifying glioma grade from magnetic resonance images of the brain. *J Person Med.* 2023;13(6):920.
49. Kim B-H, Lee H, Choi KS, et al. Validation of MRI-based models to predict MGMT promoter methylation in gliomas: BraTS 2021 radiogenomics challenge. *Cancers.* 2022;14(19):4827.
50. Xu J, Ren Y, Zhao X, et al. Incorporating multiple magnetic resonance diffusion models to differentiate low- and high-grade adult gliomas: A machine learning approach. *Quant Imaging Med Surg.* 2022;12(11):5171–5183.