Original Article

Development and validation of a robust MRI-based nomogram incorporating radiomics and deep features for preoperative glioma grading: a multi-center study

Salar Bijari[1#], Seyed Masoud Rezaeijo[2#], Sahar Sayfollahi[3], Ali Rahimnezhad[4], Sahel Heydarheydari[5,6]

[1]Department of Radiology, Faculty of Paramedical Sciences, Kurdistan University of Medical Sciences, Sanandaj, Iran; [2]Department of Medical Physics, Faculty of Medicine, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran; [3]Department of Neurosurgery, Faculty of Medical Sciences, Iran University of Medical Sciences, Tehran, Iran; [4]Student Research Committee, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran; [5]Department of Medical Imaging and Radiation Sciences, Faculty of Paramedicine, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran; [6]Cancer Research Center, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

Contributions: (I) Conception and design: S Bijari, S Heydarheydari; (II) Administrative support: SM Rezaeijo; (III) Provision of study materials or patients: S Sayfollahi; (IV) Collection and assembly of data: A Rahimnezhad; (V) Data analysis and interpretation: SM Rezaeijo, S Bijari; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Sahel Heydarheydari, PhD. Assistant Professor, Department of Medical Imaging and Radiation Sciences, Faculty of Paramedicine, Ahvaz Jundishapur University of Medical Sciences, Golestan Street, Ahvaz 61357-15794, Iran; Cancer Research Center, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. Email: sheydari7@gmail.com.

Background: Gliomas, the most common primary brain tumors, are classified into low-grade glioma (LGG) and high-grade glioma (HGG) based on aggressiveness. Accurate preoperative differentiation is vital for effective treatment and prognosis, but traditional methods like biopsy have limitations, such as sampling errors and procedural risks. This study introduces a comprehensive model that combines radiomics features (RFs) and deep features (DFs) from magnetic resonance imaging (MRI) scans, integrating clinical factors with advanced imaging features to enhance diagnostic precision for preoperative glioma grading.

Methods: In this retrospective multi-center study [2017–2022], 582 patients underwent preoperative contrast-enhanced T1-weighted (CE-T1w) and T2-weighted fluid-attenuated inversion recovery (T2w FLAIR) MRI. The dataset, divided into 407 training and 175 testing cases, included 340 LGGs and 242 HGGs. RFs and DFs were extracted from CE-T1w images, and radiomic scores (rad-score) and deep scores (deep-score) were calculated. Additionally, a clinical model based on demographics and MRI findings (CE-T1w and T2w FLAIR imaging) was developed. A nomogram model integrating rad-score, deep-score, and clinical factors was constructed using multivariate logistic regression analysis. Decision curve analysis (DCA) was employed to evaluate the nomogram's clinical utility in distinguishing between HGGs and LGGs.

Results: The study included 582 patients (mean age: 52±14 years; 57.91% male). No significant differences in age or sex were found between the training and testing groups (P>0.05). For RFs, 73.02% of the 215 extracted features were selected based on inter-class correlation coefficients (ICCs), while for DFs, 38.27% of the 15,680 extracted features were selected. Optimal penalization coefficients lambda (λ) for RFs and DFs were determined using a five-fold cross-validation and minimal criteria process. The resulting receiver operating characteristic-area under the curve (ROC-AUC) values were 0.93 [95% confidence interval (CI): 0.91–0.94] for the training set and 0.91 (95% CI: 0.89–0.93) for the testing set. The Hosmer-Lemeshow test yielded P values of 0.619 and 0.547 for the training and testing sets, respectively, indicating satisfactory calibration. The nomogram demonstrated the highest net benefit (NB) up to a threshold of 0.7, followed by DFs and RFs.

Conclusions: This study underscores the efficacy of integrating RFs and DFs alongside clinical data to accurately predict the pathological grading of HGGs and LGGs, offering a comprehensive approach for clinical decision-making.

Keywords: Glioma grading; magnetic resonance imaging-based nomogram (MRI-based nomogram); radiomics features (RFs); deep features (DFs)

Submitted Jul 29, 2024. Accepted for publication Nov 28, 2024. Published online Jan 22, 2025.
doi: 10.21037/qims-24-1543
View this article at: https://dx.doi.org/10.21037/qims-24-1543

Introduction

Glioma, the most prevalent type of primary tumor found in the central nervous system, resembles glial cells in its cellular structure. According to the World Health Organization (WHO) classification, gliomas are categorized into two main groups: low-grade gliomas (LGGs) and high-grade gliomas (HGGs) (1). In general, LGGs tend to exhibit less aggressive behavior in comparison to their HGGs counterparts. This distinction in aggressiveness is reflected not only in the rate of tumor growth but also in factors such as invasion into surrounding brain tissue and the likelihood of recurrence. Additionally, LGGs often present with fewer symptoms initially, allowing for a longer period before clinical intervention becomes necessary (2). However, despite their slower progression, LGGs can still pose significant challenges in terms of treatment and management, particularly due to their potential to transform into HGGs over time (3). Therefore, precisely identifying the grade of a glioma is essential for determining the most effective treatment strategy and patient prognosis, aiding in making more informed treatment decisions. While surgical pathology or biopsy are regarded as the definitive methods for glioma grading, they have certain limitations, including sampling errors and delays in diagnosis. Additionally, due to the location of the tumor, complete resection is sometimes not possible or can only be achieved with a significant risk of neurological deficits. Therefore, the development of a non-invasive and precise preoperative grading method is essential for improving treatment strategies and prognosis for glioma patients (4-6).

Magnetic resonance imaging (MRI), especially contrast-enhanced T1-weighted imaging (CE-T1w), is a routine clinical tool for characterizing gliomas based on their radiologic characteristics. However, the accuracy of glioma grading can be impacted by the inexperience of radiologists and variability between different observers (inter-observer variation) or even the same observer at different times (intra-observer variation). Quantitative features including radiomics features (RFs) and deep features (DFs) extracted from MRI scans can offer additional insights into the tumor's heterogeneity, aiding in clinical decision-making and improving glioma grading and prognosis. Through the use of mathematical algorithms, RFs and DFs enable the precise description of tumor phenotypes. RFs are generally classified into shape-based features and various statistical measures, including first-order, second-order, and higher-order statistics (7-9).

Although traditional radiomics software, such as standardized tools, aids in extracting RFs from regions of interest (ROIs), the integration of deep learning algorithms, like autoencoders, enables the direct extraction of DFs from images (4,5,10,11).

MRI-based RFs and DFs are currently advised in radiology for tasks including tumor grading, prognosis evaluation, and genetic status prediction (12-15).

Recently, machine learning techniques that use RFs, and DFs have become promising tools for glioma grading. These RFs and DFs models provide a non-invasive, reproducible, and cost-effective method for tumor research. They deliver high-dimensional features extracted from standard images, thereby improving diagnostic accuracy (4,16-18). Successful applications of RFs and DFs models have been particularly noted in classifying tumors, especially in the head and neck region. Previous studies have also demonstrated the effectiveness of MRI radiomics analysis in differentiating benign parotid tumors. However, current research primarily focuses on RFs and DFs for tumor characterization and lacks a comprehensive approach (7,19,20).

In this study, we present an innovative approach that integrates the inter-class correlation coefficient (ICC) to ensure reproducibility, merges RFs and DFs, and develops distinct scores for both RFs and DFs to improve tumor

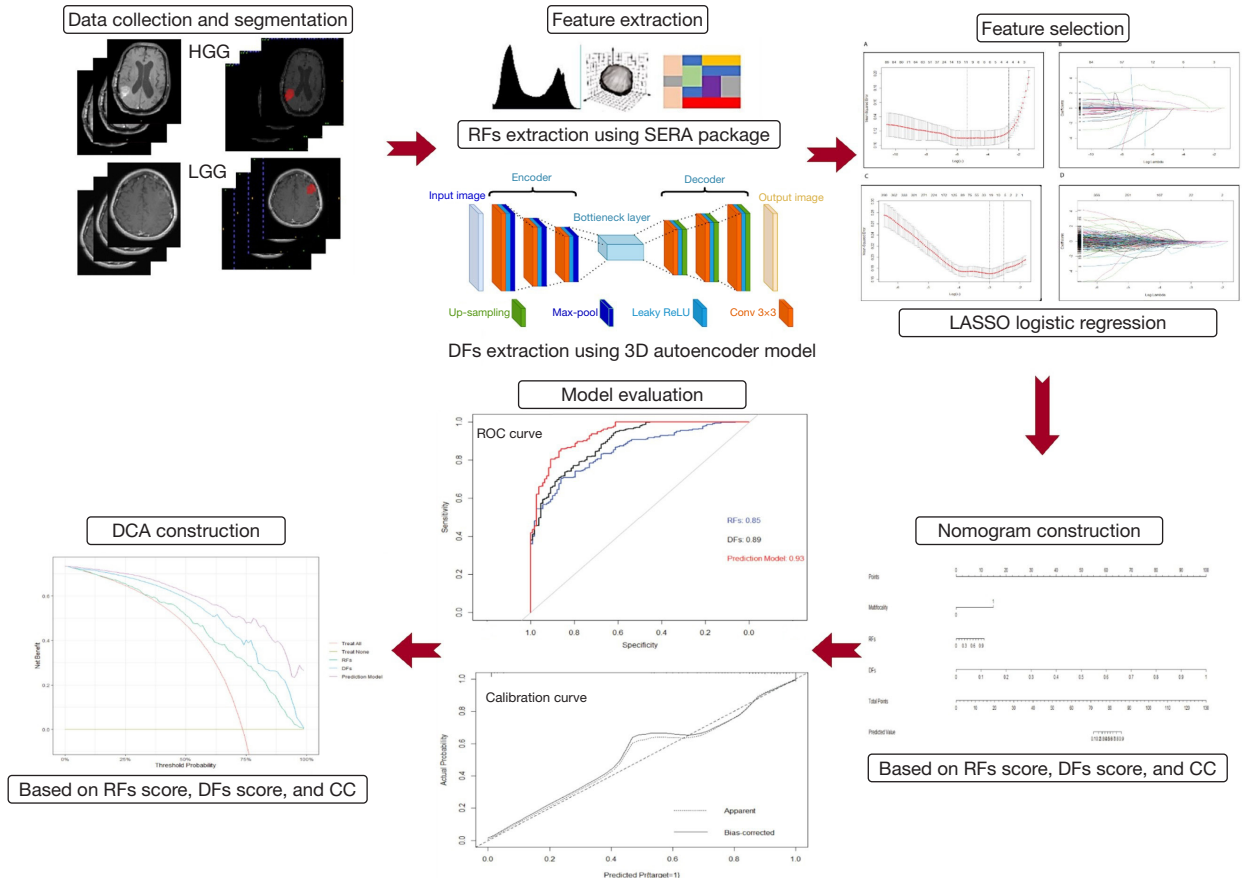

Figure 1 Proposed approach flowchart. HGG, high-grade glioma; LGG, low-grade glioma; SERA, Standardized Environment for Radiomics Analysis; RFs, radiomics features; DFs, deep features; LASSO, least absolute shrinkage and selection operator; DCA, decision curve analysis; CC, clinical characteristic; ROC, receiver operating characteristic.

characterization. Furthermore, we design a comprehensive nomogram that combines clinical parameters with RFs and DFs, supplemented by decision curve analysis (DCA) to assess diagnostic efficacy. This methodology seeks to establish a robust MRI-based nomogram for the preoperative differentiation between LGGs and HGGs, providing clinicians with a valuable tool for enhanced patient management through the extraction of a broader spectrum of statistical features. To the best of our knowledge, no existing study has employed this integrated methodology for distinguishing LGGs from HGGs.

Our primary contributions are as follows:

- ❖ Implementation of ICC to ensure feature reproducibility.
- ❖ Integrated utilization of RFs and DFs.
- ❖ Creation of RFs and DFs scores to advance tumor characterization.
- ❖ Development of a nomogram that incorporates both clinical factors and RFs/DFs.
- ❖ Application of DCA to evaluate the diagnostic performance.

Through these contributions, we aimed to develop and validate a robust MRI-based nomogram incorporating RFs and DFs for the preoperative differentiation of LGGs and HGGs, providing clinicians with a valuable tool for improved patient management. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1543/rc).

Methods

Dataset

Figure 1 presents a schematic overview of the main stages in the proposed methodology. The experimental

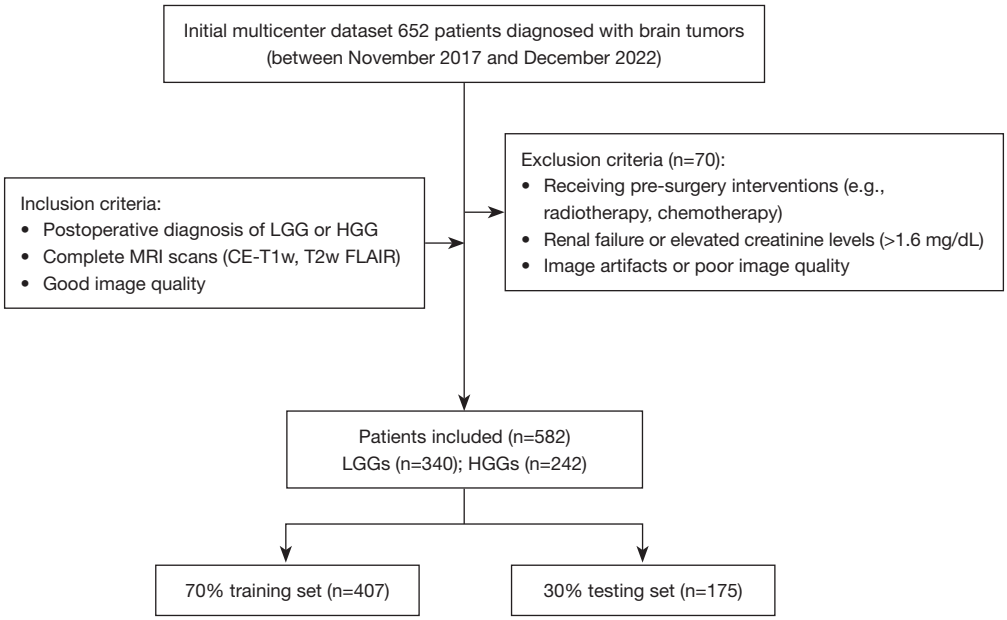

Figure 2 Inclusion flowchart. LGG, low-grade glioma; HGG, high-grade glioma; MRI, magnetic resonance imaging; CE-T1w, contrast-enhanced T1-weighted; T2w FLAIR, T2-weighted fluid-attenuated inversion recovery.

procedures for this retrospective study were approved by the institutional ethics committee of Ahvaz Jundishapur University of Medical Sciences (ethical approval No. IR.AJUMS.REC.1402.423). As the study was retrospective, the requirement for informed consent was waived by the committee. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

As shown in *Figure 2*, a retrospective analysis was conducted on a multicenter dataset that initially included 652 patients diagnosed with brain tumors. After excluding patients with incomplete data (n=70), 582 patients who underwent preoperative CE-T1w MRI and T2-weighted fluid-attenuated inversion recovery (T2w FLAIR) imaging between November 2017 and December 2022 were selected. This cohort included 242 HGGs cases (74 anaplastic astrocytoma and 168 glioblastoma) and 340 LGGs cases (48 pilocytic astrocytoma, 135 diffuse astrocytoma, and 157 oligodendroglioma. The patients were not included consecutively. Instead, they were selected based on predefined inclusion and exclusion criteria.

Inclusion criteria

The inclusion criteria for research subjects were: (I) postoperative diagnosis of LGG or HGG; and (II) complete MRI scans (CE-T1w, T2w FLAIR) of good image quality.

Exclusion criteria

Exclusion criteria included: (I) patients who received interventions such as radiotherapy or chemotherapy before surgery; (II) image artifacts or overlaps leading to poor or damaged image quality that hindered diagnosis; and (III) patients with renal failure or elevated creatinine levels (>1.6 mg/dL).

Demographic, clinical and MRI characteristics

This dataset included both demographic and clinical variables, such as age, sex, and tumor subtype classifications (*Table 1*). Tumor subtypes were categorized according to the WHO classification, with LGGs comprising pilocytic astrocytoma, diffuse astrocytoma, and oligodendroglioma, and HGGs including anaplastic astrocytoma and glioblastoma. In addition to demographic data, clinical and MRI features were collected for each patient, including necrosis-like, multifocality, hemorrhage, enhancing margins, tumor length, location, and volume, as summarized in *Table 2*. All MRI scans were reviewed independently by two experienced radiologists to confirm these features, and any disagreements were resolved through consultation with a third radiologist to reach a consensus. These characteristics

Table 1 Demographic and clinical characteristics of patients with LGG and HGG

Demographic and clinical features	Total (n=582)	LGGs (n=340)	HGGs (n=242)	P value
Age (years)	52±14	48.18±26	49.83±24	>0.05
Gender				>0.05
Male	337 (57.90)	205 (60.29)	132 (54.54)	
Female	245 (42.09)	135 (39.70)	110 (45.45)	
Subtypes				>0.05
Pilocytic astrocytoma	48 (8.24)	48 (14.11)	0	
Diffuse astrocytoma	135 (23.19)	135 (39.70)	0	
Oligodendroglioma	157 (26.97)	157 (46.17)	0	
Anaplastic astrocytoma	74 (12.71)	0	74 (30.57)	
Glioblastoma	168 (28.86)	0	168 (69.42)	

Data are presented as mean ± standard deviation or n (%). LGG, low-grade glioma; HGG, high-grade glioma.

Table 2 MRI characteristics of patients with LGG and HGG

MRI features	Total (n=582)	LGGs (n=340)	HGGs (n=242)	P value
Necrosis-like				0.228
Without	126 (21.64)	80 (23.52)	46 (19.00)	
With	456 (78.35)	260 (76.47)	196 (80.99)	
Multifocality				<0.001
Without	530 (91.06)	340 (100.00)	190 (78.51)	
With	52 (8.93)	0	52 (21.48)	
Hemorrhage				0.454
Without	511 (87.80)	290 (85.29)	200 (82.64)	
With	71 (12.19)	50 (14.70)	42 (17.36)	
Enhancing margin				<0.001
Well-defined	270 (46.39)	140 (41.17)	130 (53.71)	
Poorly-defined	312 (53.60)	200 (58.82)	112 (46.28)	
Length				0.901
>25 mm	306 (52.57)	180 (52.94)	126 (52.06)	
≤25 mm	276 (47.42)	160 (47.05)	116 (47.93)	
Location				0.248
Other	402 (69.08)	228 (67.05)	174 (71.90)	
Frontal lobe	180 (30.92)	112 (32.94)	68 (28.09)	
Subventricular zone				0.12
Without	162 (27.83)	110 (32.35)	63 (26.03)	
With	420 (72.16)	230 (67.64)	179 (73.96)	
Volume				0.589
<50 cm^3	197 (33.84)	80 (23.52)	117 (48.34)	
≥50 cm^3	385 (66.15)	260 (76.47)	125 (51.65)	
Intratumoral vascular				0.299
Without	22 (3.78)	10 (2.94)	12 (4.95)	
With	560 (96.21)	330 (97.05)	230 (95.04)	

Data are presented as n (%). MRI, magnetic resonance imaging; LGG, low-grade glioma; HGG, high-grade glioma.

were analyzed to assess their association with tumor malignancy and subtype, providing valuable insights into the distinction between LGG and HGG cases.

Imaging protocol

Gadobutrol (Gadovist; Bayer Healthcare, Berlin, Germany) was administered at the standard dose of 0.1 mmol/kg of body weight for contrast enhancement, with postcontrast images obtained promptly after administration. MRI was performed using a 1.5T scanner (Avanto; Siemens Medical Solution, Erlangen, Germany) for all patients.

Feature extraction

For clinical characterization, we utilized both CE-T1w and T2w FLAIR images. Two radiologists with 6 and 14 years of experience independently conducted the MRI feature analysis. Blinded to clinical-pathological data, they assessed these images for MRI features such as tumor regions, maximum diameter, and the presence of cysts, hemorrhage, and necrosis. They consulted a third radiologist in cases of discrepancy. However, RFs and DFs were extracted only from CE-T1w images, as this modality is routinely available and more accessible.

Before RFs extraction, all CE-T1w MRI images were resampled to a voxel size of 1 mm × 1 mm × 1 mm and converted to grayscale with a bandwidth of 25. Resampling CE-T1w images to a uniform voxel size enhances spatial consistency, which can improve feature reproducibility by reducing variations due to scanner resolution and patient positioning. Converting images to grayscale further reduces noise and minimizes color variance, contributing to more stable feature extraction. These preprocessing steps have shown to positively impact prediction accuracy, highlighting their significance in the overall model performance.

Using the Standardized Environment for Radiomics Analysis (SERA) package, a total of 215 quantitative RFs were extracted from each CE-T1w MRI sequence. Among these, 79 were first-order features, while the remaining 136 included 3D features such as morphology (Morph), local intensity (LOC), statistics (STAT), intensity histogram (IH), intensity volume histogram (IVH), co-occurrence matrix (CM), run length matrix (RLM), size zone matrix (SZM), distance zone matrix (DZM), neighbourhood grey tone difference matrix (NGT), and neighbouring grey level dependence matrix (NGL) features.

For DFs extraction, we utilized a 3D autoencoder neural network architecture, as detailed in our previously published papers (4,6,18). As shown in *Figure 1*, the autoencoder consisted of an encoder network that maps input images to a latent representation and a decoder network that reconstructs the original images from this representation. The encoder followed a standard convolutional architecture, comprising three 3×3 convolutional layers, each followed by a Leaky ReLU activation and max-pooling. The decoder path included three 3×3 convolutional layers, Leaky ReLU activation, and up-sampling. Training was performed by minimizing binary cross-entropy loss using the Adam optimization algorithm. The autoencoder processed MRI images to yield 15,680 features from the bottleneck layer. Training was carried out with carefully selected parameters over 20 epochs, using a batch size of 8. A learning rate of 0.001 was chosen to balance convergence speed and optimization stability. The dataset comprised all available 3D images to ensure robust model training and evaluation. The dataset was split before feature extraction, so that DFs were extracted from both training and testing images using the trained 3D autoencoder. This comprehensive approach improves dataset representation and enhances model generalizability.

Evaluation of ICC for RFs and DFs

The radiomics signature was evaluated by calculating ICC using a subset of 150 randomly selected MRI images, which included 98 HGGs and 52 LGGs. Two radiologists, with at least 5 years of experience in neuroimaging, independently delineated the ROIs. Any discrepancies between their delineations were resolved through consultation with a third radiologist to reach a consensus. To assess the consistency of the extracted features between the two readers, ICC was calculated. An ICC value greater than 0.75 was considered indicative of good agreement, following the guidelines by Koo and Li [2016] (21), where ICC values less than 0.5 indicate poor agreement, values between 0.5 and 0.75 indicate moderate agreement, values between 0.75 and 0.9 indicate good agreement, and values greater than 0.9 indicate excellent agreement.

Feature selection

A 7:3 ratio was used to randomly divide the acquired patient characteristics into a training set (n=407) and a testing set (n=175). Before feature selection, z-score normalization was applied to standardize the data and eliminate unit limitations. Z-score normalization standardized

the features to a common scale, promoting stable convergence of the logistic regression models. Without normalization, the training process showed instability, which negatively impacted performance on the test set. Applying normalization enhanced optimization efficiency, contributing to more reliable and robust predictions.

The analysis focused on selecting stable and reproducible RFs, ensuring an ICC greater than 0.75 for reliability. To identify optimal features and eliminate irrelevant or redundant ones, the least absolute shrinkage and selection operator (LASSO) method was applied through tenfold cross-validation. Using Fisher's exact test and the Chi-squared test, we assessed the significance of clinical features, identifying enhancing margin and multifocality as significant qualitative factors. Subsequently, a multivariate logistic regression analysis ranked these features by their predictive importance, ultimately identifying multifocality, RFs score, and DFs score as independent risk predictors for HGG patients. Logistic regression prediction models incorporating these factors were tested for generalizability on the independent test set. For each patient, a rad-score was calculated based on both DFs and RFs, weighted by the LASSO coefficients, to create a comprehensive risk prediction model included in the nomogram.

Construction of a nomogram and evaluation of model performance

A nomogram was developed for both the training and testing sets using multiple logistic regression, incorporating clinically significant factors, rad-score, and deep score derived from the most effective RFs and DFs signatures. The model's goodness of fit was assessed using the Hosmer and Lemeshow test, and calibration curves were generated. The area under the curve (AUC) was calculated for both the training and test sets to evaluate the diagnostic performance of the nomogram based on RFs and DFs factors. Additionally, DCA was used to assess the clinical utility of the nomogram in distinguishing between HGGs and LGGs.

To illustrate the practical application of the nomogram, consider a patient with an RFs score of 2.5, a DFs score of 1.8, and clinical multifocality present. By locating these values on their respective axes of the nomogram and determining the corresponding points (e.g., 50 points for RFs, 40 points for DFs, and 20 points for multifocality), we sum these to obtain a total of 110 points. This total is then used to find the probability of HGG by drawing a line downward from 110 on the "Total Points" axis to the "Probability of HGG" axis, which might indicate a 75% chance. This example demonstrates how clinicians can estimate the probability of a patient having HGG using the nomogram based on the RFs score, DFs score, and clinical features (22,23).

The preprocessing steps (e.g., resampling, grayscale conversion, and normalization) contributed significantly to the model's performance. Comparative analysis with non-preprocessed data showed a notable decrease in the model's diagnostic accuracy. This confirms the critical role of preprocessing in ensuring the reproducibility and reliability of the extracted features and the overall model predictions.

Statistical analysis

Statistical analysis was performed using R software, version 4.2.0. Univariate analysis compared clinical factors between groups. Qualitative data were analyzed with Fisher's exact test or the Chi-squared test, while quantitative data were compared using the *t*-test or Mann-Whitney *U* test.

Nomogram development and calibration plots were created using the "rms" package, and the Hosmer-Lemeshow test was conducted with the "generalhoslem" package. LASSO regression analysis was carried out using the "glmnet" package. Receiver operating characteristic (ROC) curves were plotted with the "pROC" package, and the Delong test was used to estimate differences in AUC values among the models. DCA was performed using the "dca.R" package. Statistical significance was defined as $P<0.05$.

Results

Clinicoradiological characteristics

The participants had an average age of 52±14 years, with 337 (57.91%) being male. Of these, 340 (58.41%) were diagnosed with LGGs, 242 (41.58%) with HGGs. CE-T1w images showed enhancing tumors in 320 patients (54.99%), while 262 patients (45.01%) had no enhancement. There were no statistically significant differences in patient age or sex between the two groups categorized by glioma malignancy. In terms of CE-T1w imaging characteristics, HGGs demonstrated a significantly higher incidence of positive multifocality compared to LGGs ($P<0.001$). Additionally, HGGs were more frequently associated with an enhancing margin ($P<0.001$). The demographic, clinical,

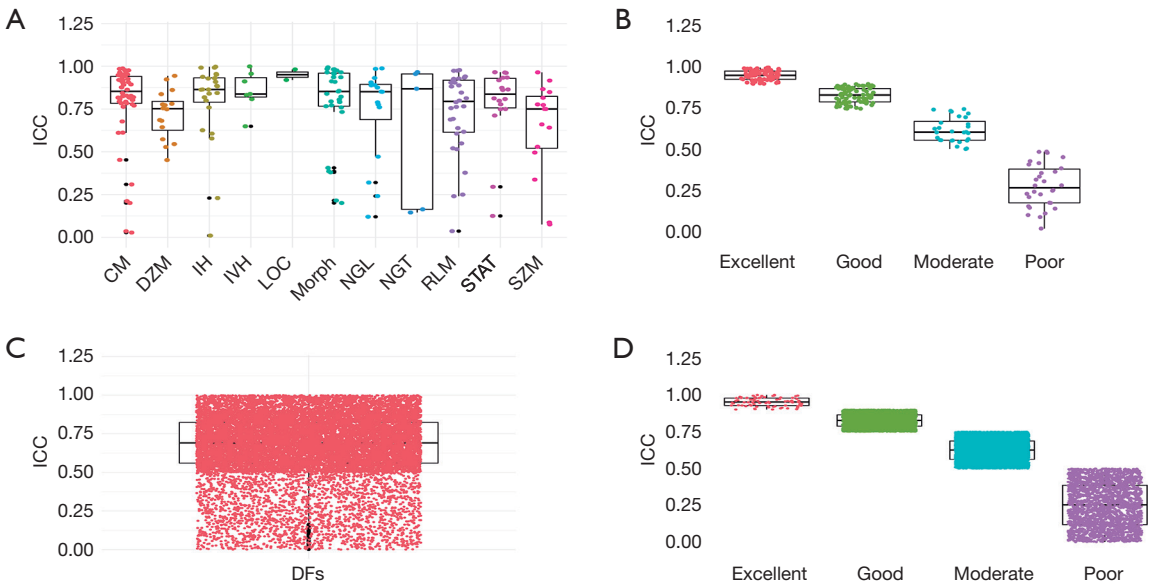

Figure 3 Detailed breakdown of RFs and DFs based on ICC values. (A) Feature categories. (B) Features into poor, moderate, good, and excellent reliability based on ICC. (C,D) The distribution of DFs across reliability levels, with poor, moderate, good, and excellent based on ICC. ICC, inter-class correlation coefficient; CM, co-occurrence matrix; DZM, distance zone matrix; IH, intensity histogram; IVH, intensity volume histogram; LOC, local intensity; Morph, morphology; NGL, neighbouring grey level dependence matrix; NGT, neighbourhood grey tone difference matrix; RLM, run length matrix; STAT, statistics; SZM, size zone matrix; RFs, radiomics features; DFs, deep features.

and MRI features of patients in both the training and testing sets are summarized in *Tables 1,2*.

Analyzing RFs and DFs based on ICC

In this section, we assess the reliability and consistency of RFs and DFs by classifying these features based on their ICC values. Our main objective is to use ICC as a measure of reliability for both RFs and DFs. We categorized the features into four distinct groups according to their ICC values.

Figure 3 provides a detailed breakdown of the features in each reliability category. In *Figure 3A*, the x-axis labels are CM, DZM, IH, IVH, LOC, Morph, NGL, NGT, RLM, STAT, SZM, each representing a specific feature category. *Figure 3B* shows the x-axis labeled as poor, moderate, good, and excellent, with these categories containing 30, 28, 83, and 71 features, respectively. Overall, 73.02% of all features were selected based on their ICC values. *Figure 3C,3D* provides a detailed breakdown of the features associated with each reliability level of DFs. In *Figure 3C*, the x-axis categorizes DFs into specific reliability levels. *Figure 3D* labels the x-axis with poor, moderate, good, and excellent categories. These categories include 2,183, 7,496, 4,253, and 1,748 features, respectively, representing 38.271% of all features selected based on ICC.

Development of RFs score and DFs score

The feature extraction process showed high inter-observer reproducibility, with inter-observer ICCs ranging from 0.75 to 1.00. From each magnetic resonance (MR) image in the training set, a combined total of 157 RFs and 6,001 DFs were extracted. Using LASSO algorithms, 13 features with non-zero coefficients were identified for LGGs and 26 for HGGs from the RF and DF datasets, respectively (*Figure 4A-4D*).

Figure 4 illustrates the selection of brain MRI features using the LASSO logistic regression model in the training set, with *Figure 4A,4C* representing the feature selection process for RFs and DFs, respectively. A five-fold cross-validation and minimal criteria process were employed to determine the optimal penalization coefficient lambda (λ) in the LASSO model. The vertical line indicates the optimal λ values, where the model achieves its best fit to the data. For RFs and DFs, the optimal λ values of 0.04204534 and

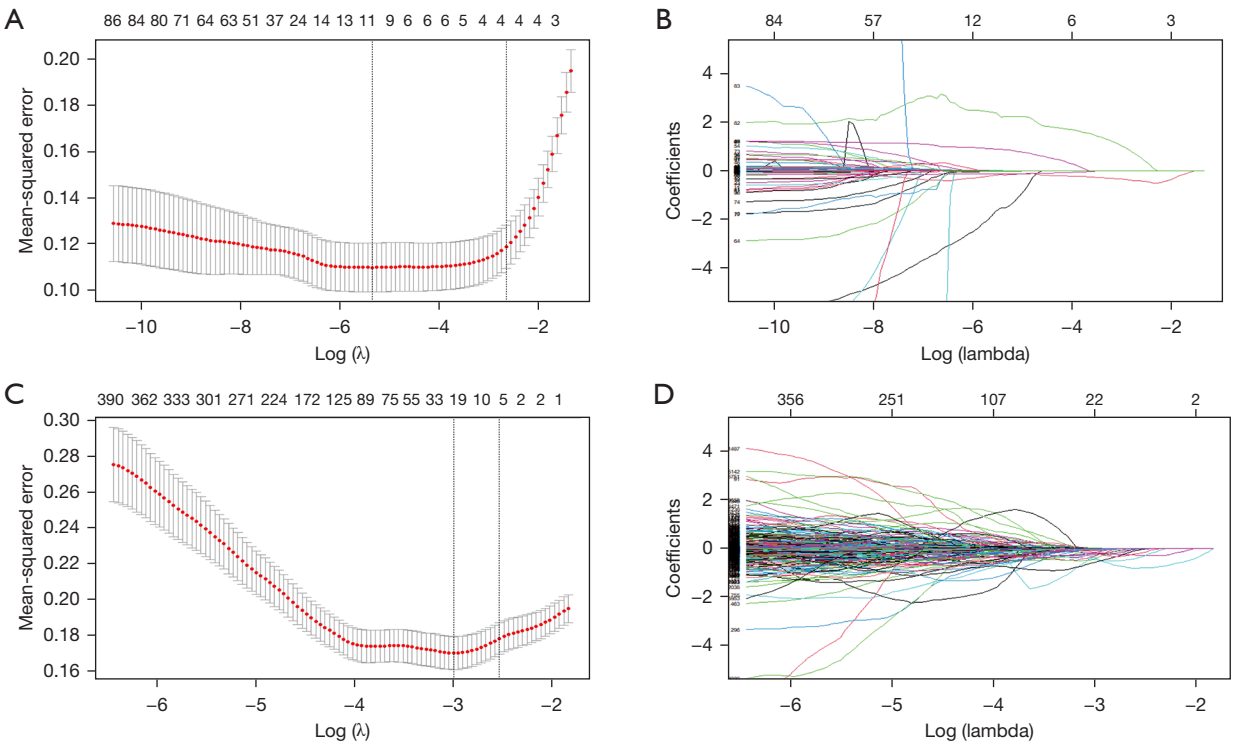

Figure 4 Feature selection using LASSO regression analysis. (A,C) The binomial deviance curves illustrate the partial likelihood deviance as a function of log (lambda), identifying the optimal lambda for RFs and DFs; (B,D) the LASSO coefficient profiles of non-zero coefficients are plotted against the log (lambda) sequence for RFs and DFs, highlighting the selected features. LASSO, least absolute shrinkage and selection operator; RFs, radiomics features; DFs, deep features.

0.01589755 [–log (λ) =3.16 and 4.14, respectively] were selected. *Figure 4B,4D* illustrate the LASSO coefficient profiles of the RFs and DFs, respectively. A vertical line is drawn at the value selected using five-fold cross-validation, resulting in 12 and 54 nonzero coefficients for RFs and DFs, respectively.

Constructing a nomogram based on RFs score, DFs score, and clinical characteristics

Multivariate logistic regression analysis identified three significant factors—RFs score, DFs score, and clinical characteristics (including multifocality)—as independent risk predictors for HGG patients. Nomograms incorporating these predictors were developed (*Figure 5*). The AUC for the optimal model was computed using data from both the training and testing sets, resulting in values of 0.93 [95% confidence interval (CI): 0.91–0.94] for the training set and 0.91 (95% CI: 0.89–0.93) for the testing set (*Figure 6*).

The calibration curves (*Figure 7*) visually represent the agreement between predicted probabilities and actual outcomes for the nomogram in both the training and test sets. These curves help assess the model's calibration performance across various predicted probabilities. Additionally, the Hosmer-Lemeshow test was performed to evaluate the nomogram's goodness-of-fit, yielding P values of 0.619 for the training set and 0.547 for the test set, indicating satisfactory calibration.

Constructing and clinical use of DCA plot

DCA provides a comprehensive evaluation of model performance by accounting for the clinical implications of true positives and false positives across various decision thresholds. It is a valuable tool for assessing the practical utility of diagnostic or predictive models in clinical settings. By incorporating information on the benefits and harms associated with different thresholds, DCA helps researchers make informed decisions about

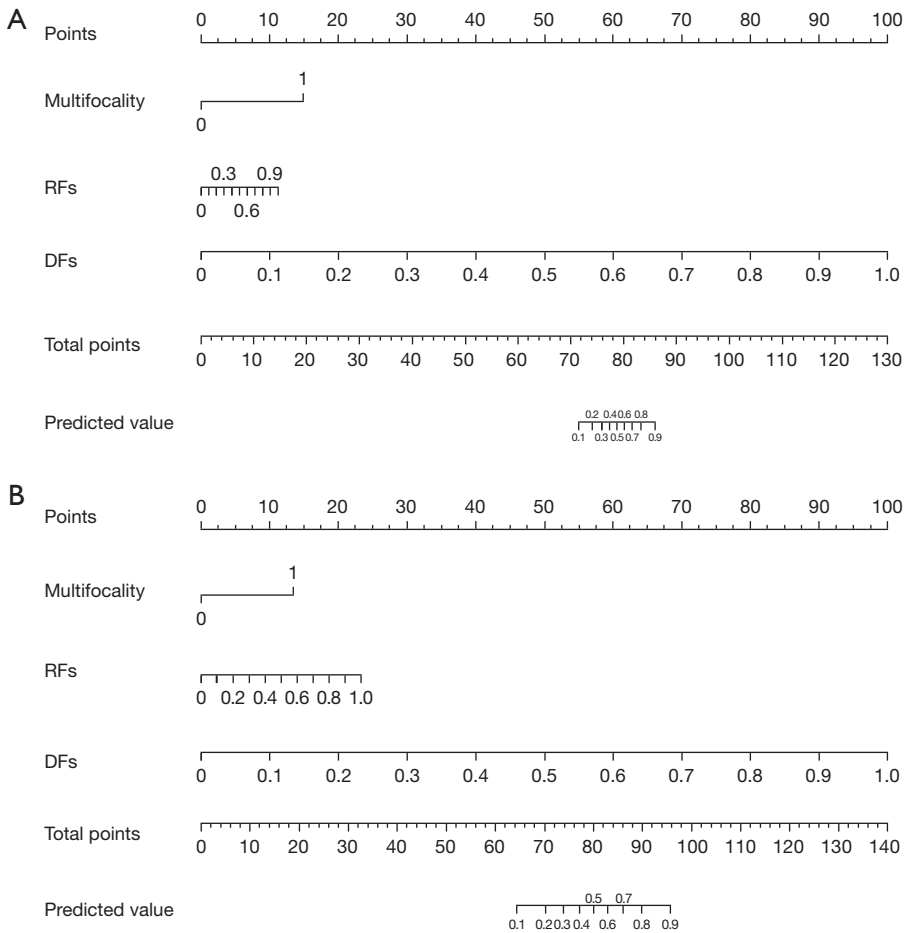

Figure 5 Nomogram for RFs and DFs, and clinical scores in discriminating between high- and low-grade glioma in the training (A) and testing (B) sets. RFs, radiomics features; DFs, deep features.

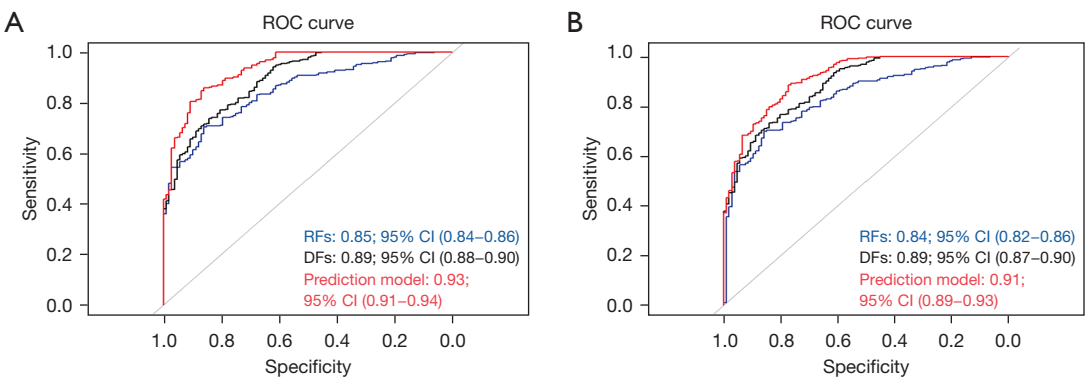

Figure 6 ROC curves for the prediction model in the training set (A) and the testing set (B). ROC, receiver operating characteristic; RFs, radiomics features; DFs, deep features; CI, confidence interval.

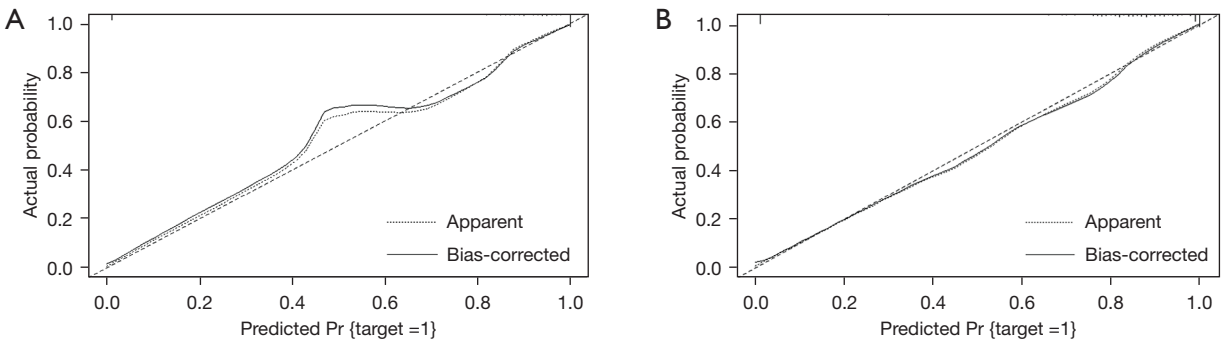

Figure 7 Calibration curves of the nomogram in the training set (A) and testing set (B).

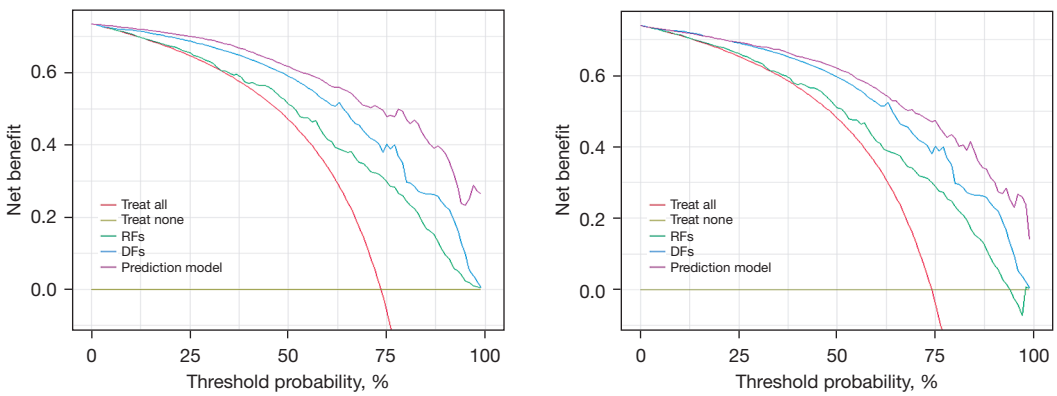

Figure 8 DCA of the nomogram based on RFs and DFs, and clinical characteristics. RFs, radiomics features; DFs, deep features; DCA, decision curve analysis.

model adoption. The net benefit (NB) is calculated using the following formula:

$$\text{Net Benefit} = \frac{\text{True positives} - \text{false positives}}{\text{Number of patients} \times \text{threshold}} \quad [1]$$

Additionally, DCA helps identify the threshold at which the model achieves the optimal balance between sensitivity and specificity, thereby maximizing its clinical utility. This detailed approach allows clinicians to tailor decision-making to individual patient needs, ensuring that interventions are well-targeted and resources are used efficiently. Ultimately, DCA facilitates the translation of predictive models from research into real-world clinical practice, contributing to improved patient outcomes and healthcare delivery.

The results of the DCA curve analysis for various models in both the training and testing datasets are shown in *Figure 8*. The nomogram curve demonstrates the highest NB up to a threshold of 0.7, followed by the DFs and RFs curves, with the DCA curve showing the lowest benefit.

Discussion

Imaging features have consistently been crucial for distinguishing between LGGs and HGGs, although reaching a consensus on this differentiation remains challenging. LGGs typically present as a focal area with minimal or no contrast enhancement, indicating limited blood-brain barrier (BBB) disruption and reduced contrast leakage. In contrast, HGGs often exhibit moderate to strong contrast enhancement on gadolinium-enhanced T1 sequences, reflecting enhanced microvascularity and significant BBB disruption. The presence of necrosis is also a crucial diagnostic feature for HGGs, further complicating the differentiation process.

In this retrospective, multi-center study, we developed and validated MRI-based nomograms that integrate RFs, DFs, and clinical characteristics extracted from

MR images for preoperative glioma grading. Through multivariate logistic regression analysis, we identified three significant predictors—RFs score, DFs score, and clinical characteristics—as independent risk factors for HGGs. These predictors were incorporated into nomograms, which were subsequently evaluated using ROC-AUC metrics for both training and testing sets, achieving high levels of accuracy with ROC-AUC values of 0.93 (95% CI: 0.91–0.94) for the training set and 0.91 (95% CI: 0.89–0.93) for the testing set.

Our approach aligns with the study by Ding *et al.* (24), who assessed the predictive efficacy of a deep learning radiomics model for glioma grading. Their model, which integrated RFs and VGG16 deep learning features, demonstrated improved differentiation between HGGs and LGGs, achieving an AUC of 0.847 in the training cohort and 0.898 in the test cohort. This comparison underscores the value of combining MR image features, as our study similarly shows that integrating RFs and DFs enhances diagnostic accuracy beyond the use of either feature set alone.

Similarly, Kobayashi *et al.* (25), extracted a standardized set of feature vectors to capture various tumor imaging phenotypes and developed a logistic regression classifier using deep radiomics, achieving an accuracy of 90%. While our study utilized a larger sample size (n=407 for training and n=175 for testing) and advanced these findings by thoroughly evaluating model performance on both datasets. This approach resulted in superior ROC-AUC values of 0.93 (95% CI: 0.91–0.94) for the training set and 0.91 (95% CI: 0.89–0.93) for the testing set. We focused on a standard T1-weighted MRI sequence with gadolinium injection for extracting RFs and DFs and employed innovative pooling and modeling techniques, which contributed to our robust outcomes. Unlike numerous studies that investigate various advanced imaging sequences, our study concentrated on a standard sequence, enhancing the practicality and accessibility of our methodology. To further validate our approach, we applied the bootstrap technique to both the training and test datasets, thereby reinforcing the model's robustness and its ability to generalize effectively across different clinical settings.

Banerjee *et al.* (26) conducted an extensive examination of the efficacy of deep convolutional neural networks in classifying brain tumors using multi-sequence MR images. They introduced ConvNet models trained from the ground up, utilizing MRI patches, slices, and multi-planar volumetric slices, and demonstrated superior accuracy, particularly when trained on multi-planar volumetric datasets. While their study highlights the potential of ConvNets, our approach differs by combining radiomics and deep learning techniques specifically for glioma grading. This combination resulted in notable performance metrics in terms of accuracy and model robustness, supported by our use of the bootstrap technique to ensure the model's reliability and generalizability. Additionally, our study incorporates unique methodologies, such as the utilization of RFs and DFs, as well as the construction of a nomogram for enhanced tumor characterization.

Our study introduces several significant and novel contributions to the field of glioma management. One of the primary goals of this research is to facilitate personalized treatment decisions for each patient. We achieved this by ensuring feature reproducibility through ICC and by combining RFs and DFs for enhanced tumor characterization. Additionally, we developed a comprehensive nomogram that integrates these features with clinical factors, providing a well-rounded tool for glioma grading. To further validate our model, we employed DCA, which is particularly valuable in clinical prediction. DCA measures the NB of a model by balancing true positives against false positives across various threshold probabilities, offering a more transparent and clinically meaningful assessment than traditional metrics (25). In our study, DCA was crucial in demonstrating the practical value of integrating RFs and DFs, confirming that our model significantly improves decision-making in preoperative glioma management. This approach not only enhances predictive accuracy but also equips clinicians with a reliable tool for making informed, patient-specific treatment decisions, ultimately leading to better clinical outcomes.

Although our study employed a substantial dataset—comprising 407 patients in the training set and 175 in the testing set—the generalizability of the results may be limited by the sample size and the specific characteristics of the studied population. Our retrospective design was also restricted to using only CE-T1w and T2w FLAIR images due to protocol limitations, which may have impacted the detection of certain features, such as hemorrhage and necrosis. While we identified necrosis-like and hemorrhage through a combination of CE-T1w and FLAIR images, more advanced imaging techniques, such as susceptibility-weighted imaging (SWI), could have provided better detection. Incorporating additional modalities like T1 mapping could further enhance diagnostic precision by offering quantitative insights into tissue properties and complementing standard

imaging methods (27). Future research with larger and more diverse cohorts, using a broader range of imaging sequences, is necessary to validate the robustness of our findings across different clinical settings. Moreover, while our study achieved promising results in terms of accuracy and model robustness, external validation using independent cohorts and real-world clinical data is essential to confirm the clinical utility of our predictive model. Further studies are also needed to evaluate the practical impact of implementing our nomogram and decision support tool on clinical decision-making and patient outcomes.

Conclusions

In this retrospective, multi-center study, we developed and validated MRI-based nomograms that integrate RFs, DFs, and clinical characteristics extracted from MR images for preoperative glioma grading. This approach underscores the effectiveness of combining imaging and clinical features to accurately predict the pathological grading of HGGs and LGGs. The integration of these multifaceted data points offers a comprehensive method for precise grading. Furthermore, the resulting nomogram serves as a robust and intuitive tool, facilitating clinical decision-making and supporting personalized treatment strategies in glioma classification.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-1543/rc

Funding: This work was supported by the Ahvaz Jundishapur University of Medical Sciences (grant No. CRC-0223).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1543/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The experimental procedures for this retrospective study were approved by the institutional ethics committee of Ahvaz Jundishapur University of Medical Sciences (ethical approval No. IR.AJUMS.REC.1402.423). As the study was retrospective, the requirement for informed consent was waived by the committee. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Saluja S, Chandra M. Brain tumour classification into high grade & low-grade gliomas: a comparative study. Turk Online J Qual Inq 2021;12:8838-47.
2. Forst DA, Nahed BV, Loeffler JS, Batchelor TT. Low-grade gliomas. Oncologist 2014;19:403-13.
3. Fouke SJ, Benzinger T, Gibson D, Ryken TC, Kalkanis SN, Olson JJ. The role of imaging in the management of adults with diffuse low grade glioma: A systematic review and evidence-based clinical practice guideline. J Neurooncol 2015;125:457-79.
4. Salmanpour MR, Hosseinzadeh M, Sanati N, Jouzdani AF, Gorji A, Mahboubisarighieh A, Maghsudi M, Rezaeijo SM, Moore S, Bonnie L, Uribe C, Ho C, Rahmim A. Tensor Deep versus Radiomics Features: Lung Cancer Outcome Prediction using Hybrid Machine Learning Systems. Journal of Nuclear Medicine 2023;64:1174.
5. Gorji A, Jouzdani AF, Sanati N, Hosseinzadeh M, Mahboubisarighieh A, Rezaeijo SM, Maghsudi M, Moore S, Bonnie L, Uribe C, Ho C, Rahmim A, Salmanpour MR. PET-CT Fusion Based Outcome Prediction in Lung Cancer using Deep and Handcrafted Radiomics Features and Machine Learning. Journal of Nuclear Medicine 2023;64:1196.
6. Salmanpour MR, Rezaeijo SM, Hosseinzadeh M, Rahmim A. Deep versus Handcrafted Tensor Radiomics Features: Prediction of Survival in Head and Neck Cancer Using Machine Learning and Fusion Techniques. Diagnostics (Basel) 2023;13:1696.

7. Brancato V, Cerrone M, Lavitrano M, Salvatore M, Cavaliere C. A Systematic Review of the Current Status and Quality of Radiomics for Glioma Differential Diagnosis. Cancers (Basel) 2022;14:2731.
8. Shboul ZA, Alam M, Vidyaratne L, Pei L, Elbakary MI, Iftekharuddin KM. Feature-Guided Deep Radiomics for Glioblastoma Patient Survival Prediction. Front Neurosci 2019;13:966.
9. Li G, Li L, Li Y, Qian Z, Wu F, He Y, Jiang H, Li R, Wang D, Zhai Y, Wang Z, Jiang T, Zhang J, Zhang W. An MRI radiomics approach to predict survival and tumour-infiltrating macrophages in gliomas. Brain 2022;145:1151-61.
10. Bello M, Nápoles G, Sánchez R, Bello R, Vanhoof K. Deep neural network to extract high-level features and labels in multi-label classification problems. Neurocomputing 2020;413:259-70.
11. Kim T, Behdinan K. Advances in machine learning and deep learning applications towards wafer map defect recognition and classification: a review. J Intell Manuf 2023;34:3215-47.
12. Bogowicz M, Riesterer O, Bundschuh RA, Veit-Haibach P, Hüllner M, Studer G, Stieb S, Glatz S, Pruschy M, Guckenberger M, Tanadini-Lang S. Stability of radiomic features in CT perfusion maps. Phys Med Biol 2016;61:8736-49.
13. Gang GJ, Deshpande R, Stayman JW. Standardization of histogram- and gray-level co-occurrence matrices-based radiomics in the presence of blur and noise. Phys Med Biol 2021;66:074004.
14. Lu X, Li M, Zhang H, Hua S, Meng F, Yang H, Li X, Cao D. A novel radiomic nomogram for predicting epidermal growth factor receptor mutation in peripheral lung adenocarcinoma. Phys Med Biol 2020;65:055012.
15. Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol 2016;61:R150-66.
16. Rezaeijo SM, Chegeni N, Baghaei Naeini F, Makris D, Bakas S. Within-Modality Synthesis and Novel Radiomic Evaluation of Brain MRI Scans. Cancers (Basel) 2023;15:3565.
17. Salmanpour MR, Hosseinzadeh M, Akbari A, Borazjani K, Mojallal K, Askari D, Hajianfar G, Rezaeijo SM, Ghaemi MM, Nabizadeh AH, Arman Rahmim A. Prediction of TNM stage in head and neck cancer using hybrid machine learning systems and radiomics features. In: Medical Imaging 2022: Computer-Aided Diagnosis. Vol 12033. SPIE; 2022:662-7.
18. Khanfari H, Mehranfar S, Cheki M, Mohammadi Sadr M, Moniri S, Heydarheydari S, Rezaeijo SM. Exploring the efficacy of multi-flavored feature extraction with radiomics and deep features for prostate cancer grading on mpMRI. BMC Med Imaging 2023;23:195.
19. Defeudis A, De Mattia C, Rizzetto F, Calderoni F, Mazzetti S, Torresin A, Vanzulli A, Regge D, Giannini V. Standardization of CT radiomics features for multi-center analysis: impact of software settings and parameters. Phys Med Biol 2020;65:195012.
20. Teng X, Zhang J, Zwanenburg A, Sun J, Huang Y, Lam S, Zhang Y, Li B, Zhou T, Xiao H, Liu C, Li W, Han X, Ma Z, Li T, Cai J. Building reliable radiomic models using image perturbation. Sci Rep 2022;12:10035.
21. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15:155-63.
22. Bijari S, Jahanbakhshi A, Abdolmaleki P. Non-invasive radiomics nomogram model for determining the low and high-grade glioma base on MRI images. International Journal of Radiation Research 2023;21:275-80.
23. Yao Y, Fu Y, Zhou G, Wang X, Li L, Mao Y, Wang J, Tan Z, Jiang M, Yi X, Chen BT. Nomogram incorporating preoperative MRI-VASARI features for differentiating intracranial extraventricular ependymoma from glioblastoma. Quant Imaging Med Surg 2024;14:2255-66.
24. Ding J, Zhao R, Qiu Q, Chen J, Duan J, Cao X, Yin Y. Developing and validating a deep learning and radiomic model for glioma grading using multiplanar reconstructed magnetic resonance contrast-enhanced T1-weighted imaging: a robust, multi-institutional study. Quant Imaging Med Surg 2022;12:1517-28.
25. Kobayashi K, Miyake M, Takahashi M, Hamamoto R. Observing deep radiomics for the classification of glioma grades. Sci Rep 2021;11:10942.
26. Banerjee S, Mitra S, Masulli F, Rovetta S. Glioma classification using deep radiomics. SN Comput Sci 2020;1:209.
27. Müller SJ, Khadhraoui E, Voit D, Riedel CH, Frahm J, Ernst M. First clinical application of a novel T1 mapping of the whole brain. Neuroradiol J 2022;35:684-91.

Cite this article as: Bijari S, Rezaeijo SM, Sayfollahi S, Rahimnezhad A, Heydarheydari S. Development and validation of a robust MRI-based nomogram incorporating radiomics and deep features for preoperative glioma grading: a multi-center study. Quant Imaging Med Surg 2025;15(2):1125-1138. doi: 10.21037/qims-24-1543

