Multi-modal magnetic resonance imaging-based grading analysis for gliomas by integrating radiomics and deep features

Zhenyuan Ning[1,2#], Jiaxiu Luo[1,2#], Qing Xiao[1,2], Longmei Cai[3], Yuting Chen[3], Xiaohui Yu[1,2], Jian Wang[1,2,3], Yu Zhang[1,2]

[1]School of Biomedical Engineering, Southern Medical University, Guangzhou, China; [2]Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China; [3]Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, China

Contributions: (I) Conception and design: Y Zhang, J Wang, Z Ning, J Luo; (II) Administrative support: Y Zhang; (III) Provision of study materials or patients: Z Ning, J Luo; (IV) Collection and assembly of data: J Wang; (V) Data analysis and interpretation: Z Ning, J Luo; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Yu Zhang, PhD. School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, 510515, China. Email: yuzhang@smu.edu.cn; Jian Wang, PhD. School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, 510515, China. Email: wangj_gz1981@163.com.

Background: To investigate the feasibility of integrating global radiomics and local deep features based on multi-modal magnetic resonance imaging (MRI) for developing a noninvasive glioma grading model.

Methods: In this study, 567 patients [211 patients with glioblastomas (GBMs) and 356 patients with low-grade gliomas (LGGs)] between May 2006 and September 2018, were enrolled and divided into training (n=186), validation (n=47), and testing cohorts (n=334), respectively. All patients underwent postcontrast enhanced T1-weighted and T2 fluid-attenuated inversion recovery MRI scanning. Radiomics and deep features (trained by 8,510 3D patches) were extracted to quantify the global and local information of gliomas, respectively. A kernel fusion-based support vector machine (SVM) classifier was used to integrate these multi-modal features for grading gliomas. The performance of the grading model was assessed using the area under receiver operating curve (AUC), sensitivity, specificity, Delong test, and *t*-test.

Results: The AUC, sensitivity, and specificity of the model based on combination of radiomics and deep features were 0.94 [95% confidence interval (CI): 0.85, 0.99], 86% (95% CI: 64%, 97%), and 92% (95% CI: 75%, 99%), respectively, for the validation cohort; and 0.88 (95% CI: 0.84, 0.91), 88% (95% CI: 80%, 93%), and 81% (95% CI: 76%, 86%), respectively, for the independent testing cohort from a local hospital. The developed model outperformed the models based only on either radiomics or deep features (Delong test, both of P<0.001), and was also comparable to the clinical radiologists.

Conclusions: This study demonstrated the feasibility of integrating multi-modal MRI radiomics and deep features to develop a promising noninvasive grading model for gliomas.

Keywords: Glioma grading; integrative analysis; radiomics; deep learning; artificial intelligence (AI)

Submitted Jun 03, 2020. Accepted for publication Nov 22, 2020.
doi: 10.21037/atm-20-4076
View this article at: http://dx.doi.org/10.21037/atm-20-4076

Introduction

Gliomas are the most common tumors of the central nervous system, accounting for 80% of all malignant tumors in the brain (1). In accordance with the World Health Organization criteria, gliomas are categorized into low-grade gliomas (LGGs) and glioblastomas (GBMs) in terms of histopathological findings (2,3). Preoperative glioma grading is important and meaningful for treatment decision and prognosis analysis (3-5). Histopathological diagnosis after the biopsy is the golden standard for glioma grading. However, its invasiveness may introduce discomfort to the patients (6-8). Accordingly, an accurate and noninvasive model is helpful for the preoperative grading of gliomas.

Radiomics provides an efficient and feasible analysis for constructing a noninvasive model based on high-throughput feature extraction. It has been used in various clinical tasks, such as disease detection, diagnosis, and prognosis analysis (9-11). Several studies have developed radiomics models for grading gliomas by extracting global radiomics features from entire regions of interest (ROIs) or volumes of interest (VOIs) on magnetic resonance imaging (MRI) sequences, such as contrast enhanced T1-weighted (T1ce) and T2 fluid-attenuated inversion recovery (T2 FLAIR) (12-14). However, these global radiomics features may lose local and glioma-specific information. Currently, deep learning-based methods have shown promising performance in medical image analysis (15-17) and have also been used for glioma grading (18,19). The advantage of deep learning-based approaches is that they can learn deep features automatically, instead of extracting hand-crafted radiomics features (20-22). To generate sufficient data for model training, many deep learning-based methods have utilized a patch-based strategy for glioma grading (23-25). Compared with the "global" radiomics features extracted from whole VOIs, the deep features extracted from patches can be regarded as the "local" features of the VOIs. Naturally, it raises the idea that whether the combination of "local" and "global" features in the glioma grading model will outperform the models based on individual "local" or "global" features.

In this work, we aimed to investigate the feasibility of integrating global radiomics and local deep features based on multi-modal MR images for developing a noninvasive glioma grading model. First, radiomics and deep features were extracted to quantify the global and local information of gliomas, respectively. Then, a kernel fusion-based support vector machine (SVM) classifier was used to integrate these multi-modal features for grading gliomas. The performance of the grading model was assessed using the area under receiver operating curve (AUC), sensitivity, specificity, Delong test, and *t*-test. The results showed that our proposed model outperformed the models based only on either radiomics or deep features, and was also comparable to the clinical radiologists. We present the following article in accordance with the TRIPOD reporting checklist (available at http://dx.doi.org/10.21037/atm-20-4076).

Methods

Patient cohorts

This retrospective study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board of Nanfang Hospital (Guangzhou, Guangdong, China; ID: NFEC-2020-251) and individual consent for this retrospective study was waived. Two datasets were collected in this study: an open dataset from The Cancer Imaging Archive, TCIA (http://www.cancerimagingarchive.net/) and the other from a local hospital. Totally, 567 patients with gliomas recorded from May 2006 to September 2018 were enrolled.

The open dataset consisted of 233 patients, including 106 patients with GBMs and 127 patients with LGGs. We randomly and equiprobably divided the dataset into the training cohort (n=186) and the validation cohort (n=47) with the ratio of 4:1, to train and select the parameters of model, respectively. For evaluating the developed model, an independent cohort from Nanfang Hospital was recruited as the external testing cohort, which comprised 334 patients, including 105 patients with GBMs and 229 patients with LGGs. All patients were pathologically confirmed as gliomas. The inclusion criteria were as follows: preoperative MR image data; available T2 FLAIR and T1ce images; high image quality without significant head motion or artifacts; and available histological grading information.

MRI acquisition and VOI segmentation

All patients underwent multi-institutional routine clinically preoperative MRI scanning, including T1ce and T2 FLAIR. For the TCIA cohort, images were acquired by using the magnetic field of 1.5T and 3T MRI systems from multiple institutions. Imaging parameters were as follows: repetition time and echo time, 5–11,000 and 0–155 msec, respectively; slice thickness, 2.5–6 mm; percentage phase field of view, 70–100%; flip angle, 90° or 150°; matrix size, 256×256 or 512×512.

For the independent testing cohort, MR images were acquired from one of three MR scanners: a 1.5T MR scanner (Achieva, Philips Healthcare, Best, The Netherlands), with a repetition time of 214 msec, echo time of 4.6 msec, slice thickness of 6 mm, flip angle of 80°, percentage phase field of view of 82%, and matrix of 512×512; a 3T MR scanner (Signa, GE Healthcare, Milwaukee, Wis, USA), with a repetition time of 600 msec, echo time of 17 msec, slice thickness of 5 mm, flip angle of

90°, percentage phase field of view of 75%, and matrix of 256×256; a 1.5T MR scanner (Avanto, Siemens Healthcare, Erlangen, Germany) with a repetition time of 663 msec, echo time of 17 msec, slice thickness of 5 mm, flip angle of 90°, percentage phase field of view of 100%, and matrix of 512×512.

The VOIs (including the whole glioma, peritumoral edema, and necrotic regions) were manually delineated on T2 FLAIR images by using ITK-SNAP 3.6 (ITK-SNAP 3.x Team, www.itksnap.org) by a radiologist with 10 years of experience. The contours of VOIs were copied to T1ce images that were aligned with T2 FLAIR images via rigid registration (*Figure 1A*).

Extraction of global radiomics features

As shown in *Figure 1B*, radiomics features were extracted based on entire VOIs, including non-texture and texture features (26). Non-texture features included the size, solidity, volume, and eccentricity of VOIs. Since different scan parameters and irrelevant information in the images would influence the texture feature extraction, several image preprocessing operators were performed to normalize all MR images, including wavelet band-pass filtering [weighted ratio: (1/2, 2/3, 1, 3/2, 2)], isotropic resampling [resampling size: (in-pR, 1, 2, 3, 4, and 5)], and gray level quantization {algorithm: (Equal, Lloyd); number of gray level: [8, 16, 32, 64]} (26) (see details in Appendix 1). Then the first-order (global), second-order (gray-level co-occurrence matrix), and high-order (gray-level run-length matrix, gray-level zone size matrix, and neighborhood gray-tone difference matrix) features were extracted.

Extraction of local deep features

We developed a convolutional neural network (CNN) model to extract deep features and used the pathologically confirmed grade as a reference standard (*Figure 1C*). The input of the CNN was non-overlapping 3D patches with a size of 24×24×24. Since gliomas are of various sizes, the choice of 24×24×24 is to ensure that the relatively small gliomas can also generate sufficient patches to train the network. Data augmentation was performed by image random rotation, translation, and zooming. Finally, a total of 9,940 3D patches were extracted from the TCIA cohort, in which 8,510 and 1,430 patches were extracted in training and validation cohorts, respectively. The designed CNN structure contained three convolutional blocks and each of the first two blocks was followed by an average pooling layer (see details in Figure S1). For each convolutional layer, kernel size was chosen as 2×2×2 with a stride of 1 and a padding size of 1, which could capture highly relevant edge information and involve detailed local textures. The kernel number of convolutional layers in three convolutional blocks were set to 32, 64, and 64, respectively. To prevent overfitting, a dropout operator with a rate of 0.2 or 0.3 was plugged into the 2×2×2 average pooling layer and the last convolutional layer. The crucial parameters of the CNN structure were experimentally tuned by internal validation cohort and will be discussed later. At the end of the network, two fully connected layers with a sigmoid activation function were used to grade gliomas. In the training step, the network was optimized using RMSprop optimizer (27) and the weights and bias were updated by the minimal batch which contained 16 patches. A predefined number of epochs was 250, and the training would been stopped when the network showed no significant performance improvement on the internal validation cohort. The learning rate was experimentally set to 0.00001.

For deep feature extraction, the feature maps output by the last convolutional layer was reshaped to a vector as a deep feature vector. Consequently, the proposed CNN could extract a feature vector for each patch. In clinical application, the objects of study were patients rather than patches. Meanwhile, the patches from the same patient might be classified by the network into different categories and confuse decision-making. To overcome this problem, we employed an average pooling strategy (28) to integrate the deep features of patches sampled from the same patient. This average pooling performed an average operation on each corresponding element in the feature vectors of all patches from the same patient, and obtained a final deep feature vector for each patient.

Feature reduction

To reduce redundancy among features and suppress overfitting, relief algorithm (29) was used to select the features with the best distinguishing power. According to relief algorithm, *k* features ranked by the importance were employed for classification (see Appendix 2). The relief algorithm was simultaneously applied to the four feature sets (i.e., T1ce radiomics, T1ce deep, T2 FLAIR radiomics, and T2 FLAIR deep features), and parameter *k* was determined in terms of the average value of the four AUCs for grading glioma. In addition, we also compared two classical feature

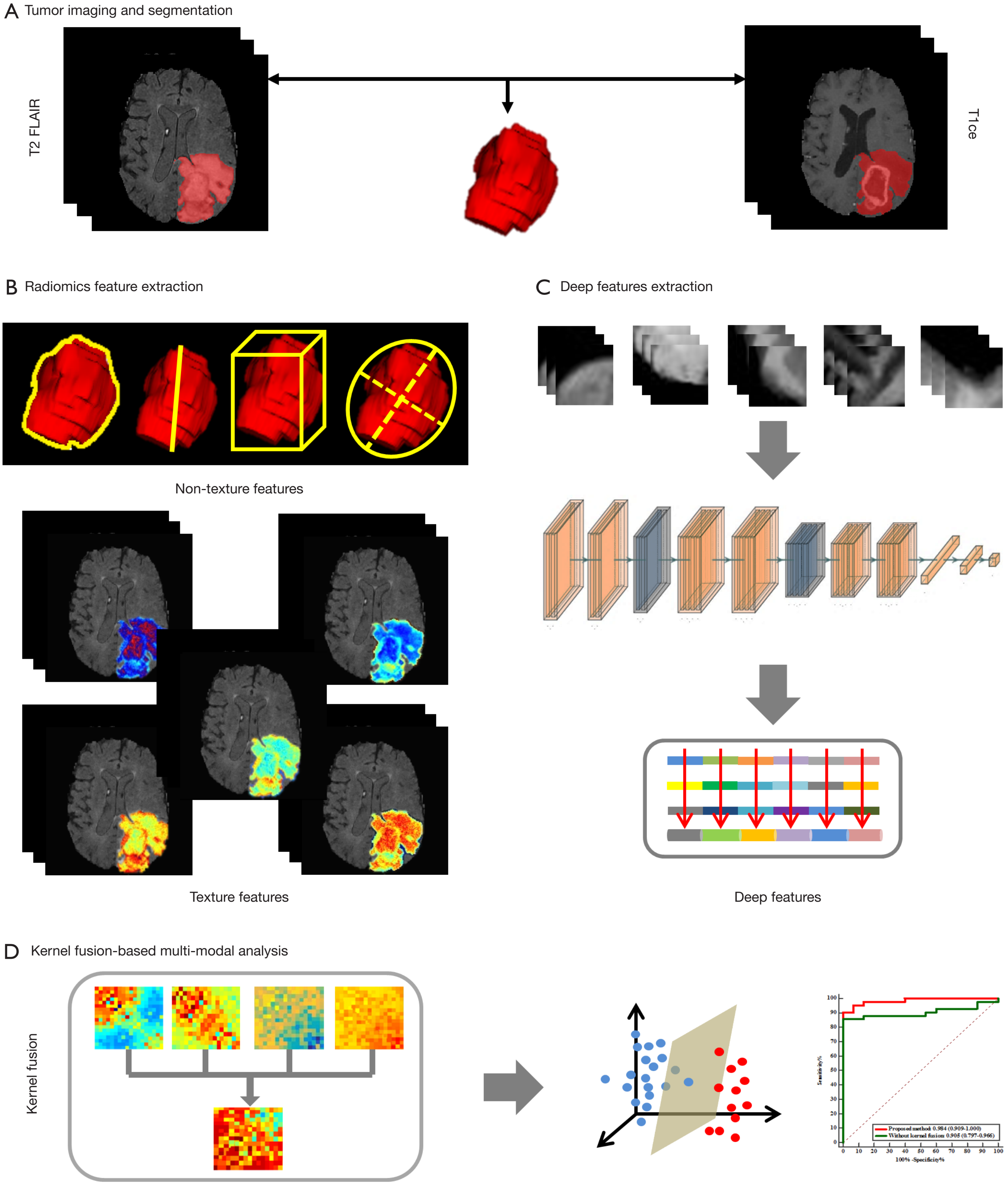
A Tumor imaging and segmentation
T2 FLAIR
T1ce
B Radiomics feature extraction
Non-texture features
Texture features
C Deep features extraction
Deep features
D Kernel fusion-based multi-modal analysis
Kernel fusion
100
90
80
70
60
50
40
30
20
10
0
Sensitivity%
0 10 20 30 40 50 60 70 80 90 100
100% -Specificity%
Proposed method: 0.984 (0.909-1.000)
Without kernel fusion: 0.905 (0.797-0.966)

Figure 1 The flowchart of the proposed integrative framework. It included four steps, namely, (A) tumor imaging and segmentation, (B) radiomics feature extraction, (C) deep feature extraction, and (D) kernel fusion-based multi-modal analysis.

selection methods, i.e., minimum redundancy maximum correlation (mRMR) and forest-based feature selection (FF) (30,31).

Multi-modal feature integrative analysis

Since features extracted from different modalities by use of different methodologies may contain different information, simply concatenating them might not make full use of the various information to build a high-precise grading model. Therefore, a kernel fusion-based SVM (32) was used to take full advantage of these features. The kernel fusion-based SVM constructed a grading model by integrating the four feature spaces into an adaptive feature kernel space with mixed weights $\omega=\{\omega_1,\omega_2,\omega_3,\omega_4|\omega_1+\omega_2+\omega_3+\omega_4=1\}$, where subscripts 1, 2, 3, and 4 denote the radiomics features on T2 FLAIR and T1ce and deep features on T2 FLAIR and T1ce, respectively. Particularly, different features are usually fitted to different kernels due to the diverse data structures. The choice of the kernel type is important and depends on the data distribution and specific application. Hence, several common kernel types were compared and the best type was selected on the basis of performance on validation cohort (see Table S1). The 10-fold cross-validation was implemented to select optimal parameters of SVM during training. Finally, combining four specific feature sets, a kernel-fusion based SVM classifier was built for glioma grading and the model was evaluated on the external testing cohort.

Radiologists reading

Three radiologists with 10, 8, and 5 years of clinical experience in radiology predicted the glioma grading on the basis of the following information: unprocessed T1ce images, T2 FLAIR images, patient age, and gender. All radiologists were blinded to the reference standard, the prediction results of the proposed model, and other radiologists' predictions. The comparison between radiologists and the proposed model was performed in terms of AUC, sensitivity, and specificity.

Statistical analysis and implementation tool

All radiomics feature extraction algorithms, feature selection methods, and kernel fusion-based SVM algorithms were implemented using MATLAB 2016b (Mathworks, Natick, USA). The deep feature extraction was implemented on Python 3.6 (Python Software Foundation, Wilmington, Delaware, USA) based on the Keras package with the TensorFlow library as the backend. Descriptive demographic statistics were summarized as mean ± standard deviation, and different groups were compared using Student's *t*-test. AUC, sensitivity, and specificity were used to assess the performance of all models. The Youden index was used to determine the optimal sensitivity and specificity. The comparisons of AUCs were performed by Delong test. The model is available at https://github.com/zhang-de-lab/zhang-lab?from=singlemessage.

Results

Baseline characters

The grading model was developed with training cohorts (n=186) from the TCIA cohort. The rest of the TCIA cohort and the database from a local hospital were used as the internal validation cohort (n=47) and external testing cohort (n=334), respectively, to evaluate the model. We ensured that the three cohorts were independent during the study. The baseline characters of the enrolled cohorts are summarized in *Table 1*, and inclusion and exclusion criteria are defined in *Figure 2*.

Critical parameter setting of the proposed model

Tuning of CNN architecture

Since deep features learned by CNN may be affected by the architecture of CNN, we performed a sequence of experiments to validate the effectiveness of the proposed architecture, including kernel size, stride and the type of activation function and pooling (see Table S2). The best architecture (kernel size: 2×2×2, stride: 1, type of activation function: Relu, type of pooling: Average) yielded the highest AUCs of 0.81 [95% confidence interval (CI): 0.78, 0.83] and 0.82 (95% CI: 0.80, 0.84) for T2 FLAIR and T1ce, respectively, in the validation cohort.

Determination of feature dimension

For feature reduction, exhaustive experiments were conducted to select the *k* (from 1 to 30) discriminant features from the four obtained feature sets (i.e., T2 FLAIR radiomics, T1ce radiomics, T2 FLAIR deep, and T1ce deep features) on the basis of the average performance of the grading models on the validation cohort. As shown in *Figure 3*, *k*=19 achieved the optimal average AUC of 0.86

Table 1 Clinical characteristics of patients on the TCIA cohort and independent testing cohort

Characteristic	TCIA cohort			Independent testing cohort		
	GBM	LGG	P value	GBM	LGG	P value
Patients	106/233 (45.5)	127/233(54.5)		105/334 (31.4)	229/334 (68.6)	
Age (year)	58.7±13.6	46.0±13.8	<0.001*	44.3±13.3	37.5±11.6	<0.001*
Gender			0.04*			<0.001*
Woman	41/106 (38.7)	63/127 (49.6)		35/105 (33.3)	107/229 (46.7)	
Man	65/106 (61.3)	64/127 (50.4)		70/105 (66.7)	122/229 (53.3)	
Tumor grade			<0.001*			<0.001*
WHO II	0 (0.0)	63 (49.6)		0 (0.0)	112 (48.9)	
WHO III	0 (0.0)	64 (50.4)		0 (0.0)	117 (51.1)	
WHO IV	106 (100.0)	0 (0.0)		105 (100.0)	0 (0.0)	
1p19q codeletion			<0.001*			0.70
Codeletion	0/106 (0.0)	36/127 (28.3)		5/105 (4.8)	14/229 (6.1)	
Wild type	106/106 (100.0)	91/127 (71.7)		12/105 (11.4)	31/229 (13.5)	
Unknown	0/106 (0.0)	0/127 (0.0)		88/105 (83.8)	184/229 (80.4)	
IDH mutation			0.64			<0.001*
Mutation	8/106 (7.5)	103/127 (81.1)		18/105 (17.1)	110/229 (48.0)	
Wild type	98/106 (92.5)	24/127 (18.9)		46/105 (43.8)	48/229 (21.0)	
Unknown	0/106 (0.0)	0/127 (0.0)		41/105 (39.1)	71/229 (31.0)	
Histology			<0.001*			<0.001*
Astrocytoma	0/106 (0.0)	36/127 (28.3)		0/105 (0.0)	82/229 (35.8)	
Oligoastrocytoma	0/106 (0.0)	32/127 (25.2)		0/105 (0.0)	83/229 (36.2)	
Oligodendroglioma	0/106 (0.0)	59/127 (46.5)		0/105 (0.0)	44/229 (19.2)	
Glioblastoma	106/106 (100.0)	0/127 (0.0)		105/105 (100.0)	0/229 (0.0)	
Unknown	0/106 (0.0)	0/127 (0.0)		0/105 (0.0)	20/229 (8.8)	

Data in parentheses are percentages. * indicates significant difference.

among the 30 options. Finally, 19 features were selected (see Tables S3-S5).

Weight setting and kernel type selection in kernel fusion-based SVM

The grid method was used to identify the optimal assembly of four weights. The four weights with an interval of 0.05 from 0 to 1 were set and satisfied the condition of $\omega = \{\omega_1, \omega_2, \omega_3, \omega_4 | \omega_1 + \omega_2 + \omega_3 + \omega_4 = 1\}$. Among six kernel types, chi-square kernel yielded the highest AUC of 0.94 (95% CI: 0.85, 0.99) with the best coefficient combination of 0.35, 0.15, 0.15, and 0.35 (see Table S6). Furthermore, to validate the effectiveness of kernel fusion, we compared it with direct concatenating integration. The performance of kernel fusion with AUC of 0.94 (95% CI: 0.85, 0.99) and 0.88 (95% CI: 0.84, 0.91) was superior to that of direct concatenating integration with AUC of 0.91 (95% CI: 0.79, 0.97) and 0.84 (95% CI: 0.79, 0.88) on the internal validation and external testing cohort (Delong test: P<0.05), respectively. This result indicated that kernel fusion was an effective strategy for multi-modal and multi-feature analysis.

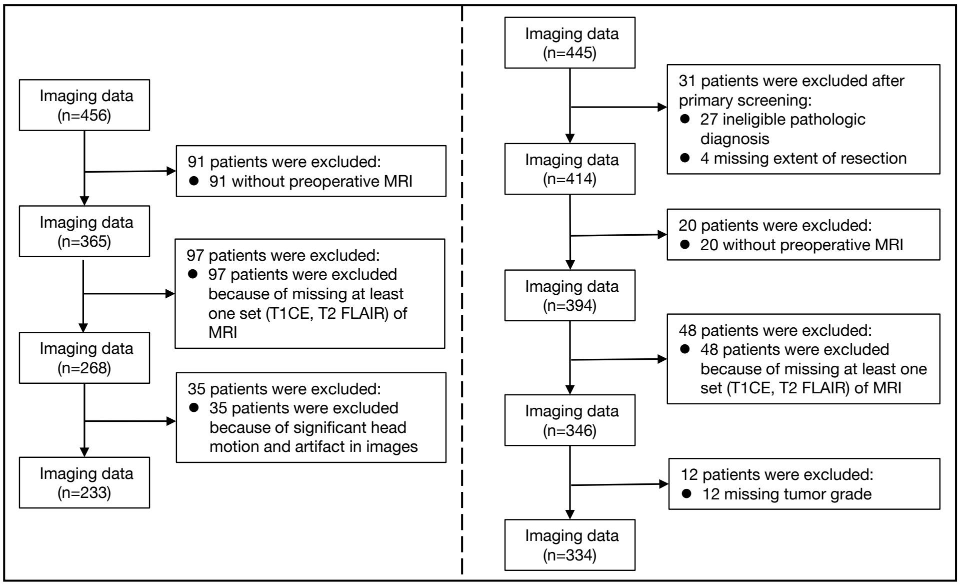
Imaging data (n=456)
91 patients were excluded:
● 91 without preoperative MRI
Imaging data (n=365)
97 patients were excluded:
● 97 patients were excluded because of missing at least one set (T1CE, T2 FLAIR) of MRI
Imaging data (n=268)
35 patients were excluded:
● 35 patients were excluded because of significant head motion and artifact in images
Imaging data (n=233)
Imaging data (n=445)
31 patients were excluded after primary screening:
● 27 ineligible pathologic diagnosis
● 4 missing extent of resection
Imaging data (n=414)
20 patients were excluded:
● 20 without preoperative MRI
Imaging data (n=394)
48 patients were excluded:
● 48 patients were excluded because of missing at least one set (T1CE, T2 FLAIR) of MRI
Imaging data (n=346)
12 patients were excluded:
● 12 missing tumor grade
Imaging data (n=334)

Figure 2 Patient inclusion and exclusion criteria on internal validation cohort and external testing cohort.

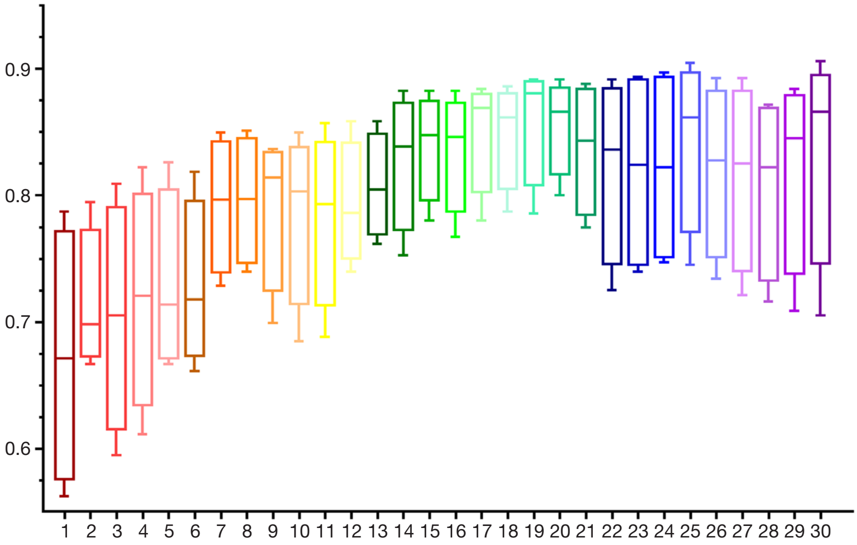

Figure 3 Average AUC of models on internal validation cohort with different feature dimensions from 1 to 30. The *k*=19 achieved the best performance with AUC of 0.86.

Comparison of different models

In this work, two excellent methods (i.e., radiomics and deep learning) were separately implemented on two MRI sequences (i.e., T2 FLAIR and T1ce) to obtain four specific feature sets. To validate the effectiveness of integrating radiomics and deep learning models, we compared our proposed model with these models based on an individual approach and single MRI sequence. As shown in *Table 2*, the best grading performance was obtained by the proposed integrative model with the AUC of 0.94 (95% CI: 0.85, 0.99) on internal validation cohort and 0.88 (95% CI: 0.84, 0.91) on external testing cohort. The performances were inferior when only a single method with a single MRI sequence was used. For internal validation cohort, the AUC was 0.88 (95% CI: 0.75, 0.95) for radiomics on T2 FLAIR,

Table 2 Comparison of the proposed method with the models based on different methodologies and modalities

Method	Validation AUC	Sensitivity (%)	Specificity (%)	Testing AUC	Sensitivity (%)	Specificity (%)
R_T2 FLAIR	0.88 (0.75, 0.95)	76 (16/21)	85 (22/26)	0.81 (0.77, 0.85)	80 (84/105)	71 (163/229)
R_T1ce	0.87 (0.74, 0.95)	81 (17/21)	77 (20/26)	0.80 (0.75, 0.84)	71 (75/105)	79 (180/229)
R_T2 FLAIR + T1ce	0.92 (0.80, 0.98)	86 (18/21)	85 (22/26)	0.85 (0.81, 0.89)	85 (89/105)	76 (175/229)
D_T2 FLAIR	0.86 (0.73, 0.95)	81 (17/21)	77 (20/26)	0.79 (0.75, 0.84)	75 (79/105)	76 (175/229)
D_T1ce	0.88 (0.75, 0.95)	71 (15/21)	88 (23/26)	0.80 (0.76, 0.84)	80 (84/105)	69 (158/229)
D_T2 FLAIR + T1ce	0.91 (0.79, 0.97)	81 (17/21)	88 (23/26)	0.84 (0.80, 0.88)	82 (86/105)	77 (176/229)
R + D_ T2 FLAIR	0.90 (0.77, 0.97)	81 (17/21)	85 (22/26)	0.82 (0.78, 0.86)	78 (82/105)	76 (173/229)
R + D_T1ce	0.90 (0.80, 0.98)	86 (18/21)	81 (21/26)	0.82 (0.77, 0.86)	81 (85/105)	75 (172/229)
Proposed method	0.94 (0.85, 0.99)	86 (18/21)	92 (24/26)	0.88 (0.85, 0.92)	88 (92/105)	81 (186/229)

Note. R_: Radiomics, D_: Deep learning, R + D_: Radiomics + Deep learning

Table 3 Performance of models using different feature selection methods and classifiers

Method	Validation AUC	Sensitivity (%)	Specificity (%)	Testing AUC	Sensitivity (%)	Specificity (%)
mRMR + SVM	0.92 (0.80, 0.98)	86 (18/21)	85 (22/26)	0.85 (0.81, 0.89)	87 (91/105)	74 (169/229)
FF + SVM	0.90 (0.77, 0.97)	86 (18/21)	88 (23/26)	0.84 (0.80, 0.88)	83 (87/105)	75 (171/229)
Relief + RF	0.92 (0.80, 0.98)	81 (17/21)	92 (24/26)	0.86 (0.82, 0.90)	85 (89/105)	81 (186/229)
Relief + SVM	0.94 (0.85, 0.99)	86 (18/21)	92 (24/26)	0.88 (0.84, 0.91)	88 (92/105)	81 (186/229)

0.87 (95% CI: 0.74, 0.95) for radiomics on T1ce, 0.86 (95% CI: 0.73, 0.95) for deep learning on T2 FLAIR, and 0.88 (95% CI: 0.75, 0.95) for deep learning on T1ce. For external testing cohort, the AUC was 0.81 (95% CI: 0.77, 0.85), 0.80 (95% CI: 0.75, 0.84), 0.79 (95% CI: 0.75, 0.84), and 0.80 (95% CI: 0.76, 0.84), respectively. By contrast, the results were improved whether conducting each single approach on multi-modal images or combining two approaches on a single sequence. The AUC for radiomics on multi-modal images, deep learning on multi-modal images, combined methods on T2 FLAIR, and combined methods on T1ce was 0.92 (95% CI: 0.80, 0.98), 0.91 (95% CI: 0.79, 0.97), 0.90 (95% CI: 0.77, 0.97), and 0.90 (95% CI: 0.80, 0.98) for internal validation cohort; and 0.85 (95% CI: 0.81, 0.89), 0.84 (95% CI: 0.80, 0.88), 0.82 (95% CI: 0.78, 0.86), and 0.82 (95% CI: 0.78, 0.86) for external testing cohort, respectively.

Some conclusions could be drawn: (I) the combination of radiomics and deep learning method achieved the better performance compared with individual radiomics or deep learning method whether single or multiple modal MR images were used; (II) the models based on multiple modal MR images were superior to these models based on individual T1ce or T2 FLAIR MR images. In addition, the comparison results of the proposed integrative model with other models with different feature selection methods and classifiers are listed in *Table 3*, from which we can see that relief algorithm and SVM showed the better results when compared with other methods (Delong test: P<0.05).

Radiologists reading

Table 4 shows the sensitivity and specificity of three radiologists and the proposed model. For internal validation and external testing cohorts performed by the proposed model, the sensitivity at the optimal threshold of the Youden index was 86% and 88%, respectively, while the specificity was 92% and 81%, respectively. The sensitivity of the radiologists ranged between 80% and 90%, while the specificity was between 72% and 85%. *Figure 4A,B* show the ROC of our proposed model on internal validation and external testing cohort, respectively. For comparison, the points representing the sensitivity and specificity of the three radiologists for grading glioma are also shown

Table 4 Comparison between three radiologists and the proposed method for gliomas grading

Radiologist	Validation cohort		Testing cohort	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Reader 1	90 (19/21)	77 (20/26)	84 (88/105)	75 (172/229)
Reader 2	86 (18/21)	85 (22/26)	81 (85/105)	78 (178/229)
Reader 3	86 (18/21)	73 (19/26)	80 (84/105)	72 (165/229)
Proposed method	86 (18/21)	92 (24/26)	88 (92/105)	81 (186/229)

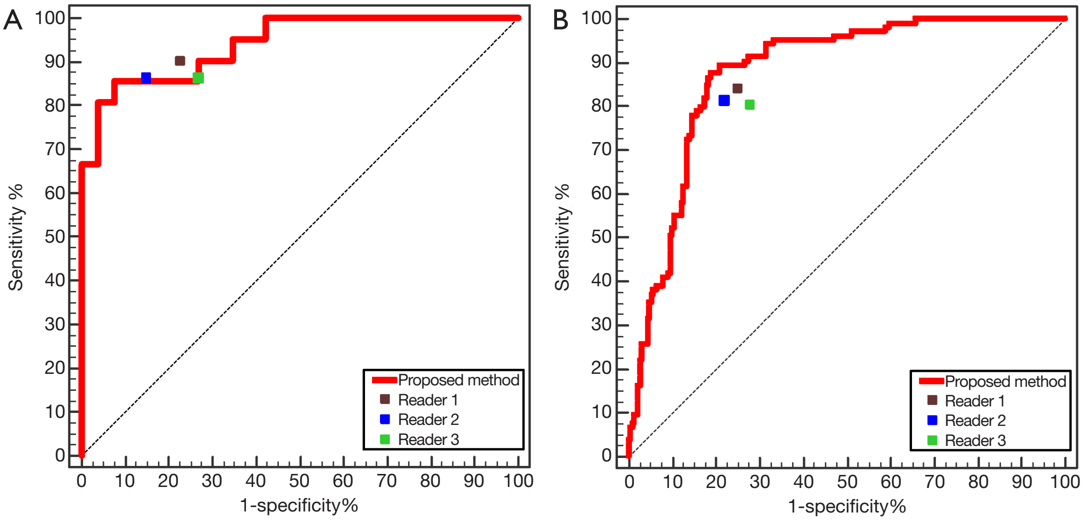

Figure 4 ROC curve of the proposed method on internal validation cohort (A) and external testing cohort (B). Three points representing sensitivity and specificity of three readers were plotted.

in *Figure 4*. These points were close to the ROC of the proposed model.

Discussion

In this study, we presented a multi-modal MRI-based grading analysis by combining the radiomics and deep learning technologies. The proposed grading model took full advantage of the global radiomics and local deep features from multi-modal MR images and achieved outstanding results in the internal validation and external testing cohorts. This finding demonstrated the feasibility of integrating global radiomics and local deep features to develop a model for glioma grading.

Most previous studies only focused on individual radiomics or deep learning approaches to conduct a quantitative analysis of glioma grading (5,14,33). However, few researches explored whether combining radiomics and deep learning approaches can improve the performance of a grading model for gliomas. According to the comparison results between the proposed method and two individual approaches based on single and multiple modal MR images, it could be found that (I) the integration of radiomics and deep learning model outperformed any single method whether single or multiple modal MR images were used; (II) multi-modal MR images could provide more information than single modal MR images regardless of the methodologies. Meanwhile, the sensitivity and specificity of the proposed model were comparable to those of clinical radiologists, which further emphasized the promising preliminary results of the proposed method for glioma grading. In addition, for radiologists, the sensitivity on the testing cohort were lower than that of the validation cohort, which might be caused by more testing samples, and affected by the experience, and some subjective factors (such as fatigue and attention). Moreover, the proposed model could provide robust sensitivity in glioma grading. Meanwhile, the lower specificity of the proposed model on the testing cohort needs to be improved, as the high number of false-positive gliomas would require the clinical radiologists to verify the actual presence of disease, which would increase the overall time of image interpretation. A potential reason of lower specificity was that cohorts used in our studies were from different centers, in which different fields of view, spatial resolutions, section thickness, and intersection gaps of the different sequences were acquired during the MRI examination.

One of the challenges of this study is the high feature dimension, which is inclined to trap in overfitting; and the modeling with high-dimension features is also time-consuming. In our work, a total of 10,324 radiomics and 13,824 deep features were extracted for singular modality. Therefore, the following three strategies were used to select discriminant features and suppress overfitting in our work. (I) The proposed CNN was equipped with pooling and dropout operators, which prevented the redundancy of features and improved robustness. (II) Patch pooling was used to integrate features from the patch level to the patient level rather than directly connecting all deep features. This procedure could also address the number of inconsistency caused by the patch strategy. (III) The relief algorithm was conducted on four feature sets, and the reduced dimension remained the same to fit the subsequent analysis. The second challenge is taking full advantage of multi-modal information and multiple types of features. A kernel fusion-based SVM was introduced to weigh different features instead of direct concatenation. In this way, features extracted from different modal MRIs by different technologies were effectively integrated, and the classification results also demonstrated the fusion strategy was helpful to analysis of multi-modal features.

Our study has several deficiencies. First, tumor segmentation was still a manual process, which was time-consuming and depended on the experience of the radiologists. Second, only T1ce and T2 FALIR sequences were used in our study; however, more than two modal MRI data can be collected for further analysis. Lastly, this work was a retrospective study, and a prospective cohort is required to further evaluate the performance of the glioma grading model.

In conclusion, an effective integrative strategy that combined two popular technologies, i.e., radiomics and deep learning, was proposed for grading gliomas. The approach adopted the kernel fusion method to build a discriminative SVM classifier based on postcontrast enhanced T1-weighted and T2 fluid-attenuated inversion recovery sequences. Furthermore, an independent external testing cohort was used to assess the generalization performance of the proposed grading model. The promising results demonstrated the feasibility of integrating radiomics and deep learning based on multi-modal magnetic resonance images for grading gliomas.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China [grantno.61671230, 61971213]; the Science and Technology Program of Guangdong Province [grant no. 2017A020211012]; the Guangdong Provincial Key Laboratory of Medical Image Processing [grant no.2014B030301042]; and the Natural Science Foundation of Guangdong Province [grant no. 2019A1515010417].

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at http://dx.doi.org/10.21037/atm-20-4076

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-4076). 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. This retrospective study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board of Nanfang Hospital (Guangzhou, Guangdong, China; ID: NFEC-2020-251) and individual consent for this retrospective study was waived.

References

1. Ostrom QT, Gittleman H, Liao P, et al. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2007-2011. Neuro Oncol 2014;16:iv1-iv63.
2. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A summary. Acta Neuropathol 2016;131:803-20.
3. Arvinda HR, Kesavadas C, Sarma PS, et al. Glioma grading: sensitivity, specificity, positive and negative predictive values of diffusion and perfusion imaging. J Neurooncol 2009;94:87-96.
4. Whittle IR. The dilemma of low grade glioma. J Neurol Neurosurg Psychiatry 2004;75:ii31-6.
5. Tian Q, Yan LF, Zhang X, et al. Radiomics strategy for glioma grading using texture features from multiparametric MRI: Radiomics Approach for Glioma Grading. J Magn Reson Imaging 2018;48:1518-28.
6. Jackson RJ, Fuller GN, Abi-Said D, et al. Limitations of stereotactic biopsy in the initial management of gliomas. Neuro Oncol 2001;3:193-200.
7. Glantz MJ, Burger PC, Herndon JE, et al. Influence of the type of surgery on the histologic diagnosis in patients with anaplastic gliomas. Neurology 1991;41:1741-4.
8. Field M, Witham TF, Flickinger JC, et al. Comprehensive assessment of hemorrhage risks and outcomes after stereotactic brain biopsy. J Neurosurg 2001;94:545-51.
9. Huang YQ. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. J Clin Oncol 2016;34:2157-64.
10. Kim M, Jung SY, Park JE, et al. Diffusion- and perfusion-weighted MRI radiomics model may predict isocitrate dehydrogenase (IDH) mutation and tumor aggressiveness in diffuse lower grade glioma. Eur Radiol 2020;30:2142-51.
11. Yang X, He J, Wang J, et al. CT-based radiomics signature for differentiating solitary granulomatous nodules from solid lung adenocarcinoma. Lung Cancer 2018;125:109-14.
12. Hsieh KL, Lo CM, Hsiao CJ. Computer-aided grading of gliomas based on local and global MRI features. Comput Methods Programs Biomed 2017;139:31-8.
13. Lu CF, Hsu FT, Hsieh LC, et al. Machine Learning-Based Radiomics for Molecular Subtyping of Gliomas. Clin Cancer Res 2018;24:4429-36.
14. Su C, Jiang J, Zhang S, et al. Radiomics based on multicontrast MRI can precisely differentiate among glioma subtypes and predict tumour-proliferative behaviour. Eur Radiol 2019;29:1986-96.
15. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199-210.
16. Chen PJ, Lin MC, Lai MJ, et al. Accurate classification of diminutive colorectal polyps using computer aided analysis. Gastroenterology 2018;154:568-75.
17. Wang K, Lu X, Zhou H, et al. Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. Gut 2019;68:729-41.
18. Decuyper M, Bonte S, Van Holen R. Binary Glioma Grading: Radiomics versus Pre-trained CNN Features. In: Frangi A, Schnabel J, Davatzikos C, et al. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Cham: Springer, 2018:498-505.
19. Huang P, Li D, Jiao Z, et al. CoCa-GAN: Common-Feature-Learning-Based Context-Aware Generative Adversarial Network for Glioma Grading. In: Shen Dm Liu T, Peters TM, et al. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Cham: Springer, 2019.
20. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
21. Chaudhary K, Poirion OB, Lu L, et al. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res 2018;24:1248-59.
22. Wang S, Zhou M, Liu Z, et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. Med Image Anal 2017;40:172-83.
23. Sajjad M, Khan S, Muhammad K, et al. Multi-Grade Brain Tumor Classification using Deep CNN with Extensive Data Augmentation. J Comput Sci 2018;30:174-82.
24. Banerjee S, Mitra S, Masulli F, et al. Deep Radiomics for Brain Tumor Detection and Classification from Multi-Sequence MRI. arXiv:1903.09240v1 [preprint]. 2019 [cited 2019 Mar 21]. Available online: https://arxiv.org/abs/1903.09240
25. Decuyper M, Holen VR. Fully Automatic Binary Glioma Grading based on Pre-Therapy MRI using 3D

Convolutional Neural Networks. arXiv:1908.01506 [preprint]. 2019 [cited 2019 Aug 5]. Available online: https://arxiv.org/abs/1908.01506v1

26. Vallières M, Freeman CR, Skamene S, et al. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol 2015;60:5471-96.
27. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA 2012:4:26-31.
28. Ning Z, Luo J, Li Y, et al. Pattern Classification for Gastrointestinal Stromal Tumors by Integration of Radiomics and Deep Convolutional Features. IEEE J Biomed Health Inform 2019;23:1181-91.
29. Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. Available online: https://www.aaai.org/Papers/AAAI/1992/AAAI92-020.pdf
30. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226-38.
31. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507-17.
32. Luo J, Ning Z, Zhang S, et al. Bag of deep features for preoperative prediction of sentinel lymph node metastasis in breast cancer. Phys Med Biol 2018;63:245014.
33. Yang Y, Yan LF, Zhang X, et al. Glioma Grading on Conventional MR Images: A Deep Learning Study With Transfer Learning. Front Neurosci 2018;12:804.

Cite this article as: Ning Z, Luo J, Xiao Q, Cai L, Chen Y, Yu X, Wang J, Zhang Y. Multi-modal magnetic resonance imaging-based grading analysis for gliomas by integrating radiomics and deep features. Ann Transl Med 2021;9(4):298. doi: 10.21037/atm-20-4076