Original Article

Incorporating multiple magnetic resonance diffusion models to differentiate low- and high-grade adult gliomas: a machine learning approach

Junqi Xu[1#^], Yan Ren[2#^], Xueying Zhao[1^], Xiaoqing Wang[3^], Xuchen Yu[1^], Zhenwei Yao[2^], Yan Zhou[3^], Xiaoyuan Feng[2^], Xiaohong Joe Zhou[4^], He Wang[1,5^]

[1]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China; [2]Radiology Department, Hua Shan Hospital, Fudan University, Shanghai, China; [3]Department of Radiology, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China; [4]Center for Magnetic Resonance Research, Departments of Radiology, Neurosurgery, and Bioengineering, University of Illinois at Chicago, Chicago, IL, USA; [5]Human Phenome Institute, Fudan University, Shanghai, China

Contributions: (I) Conception and design: J Xu, Y Ren, X Zhao, H Wang, XJ Zhou; (II) Administrative support: H Wang; (III) Provision of study materials or patients: Y Ren, X Wang, Z Yao, Y Zhou, X Feng, X Yu; (IV) Collection and assembly of data: J Xu, Y Ren, X Zhao, X Wang, Z Yao, Y Zhou, X Feng; (V) Data analysis and interpretation: J Xu, X Zhao, H Wang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: He Wang. Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, 220 Handan Road, Shanghai 200433, China. Email: hewang@fudan.edu.cn.

Background: Accurate grading of gliomas is a challenge in imaging diagnosis. This study aimed to evaluate the performance of a machine learning (ML) approach based on multiparametric diffusion-weighted imaging (DWI) in differentiating low- and high-grade adult gliomas.

Methods: A model was developed from an initial cohort containing 74 patients with pathology-confirmed gliomas, who underwent 3 tesla (3T) diffusion magnetic resonance imaging (MRI) with 21 b values. In all, 112 histogram features were extracted from 16 parameters derived from seven diffusion models [monoexponential, intravoxel incoherent motion (IVIM), diffusion kurtosis imaging (DKI), fractional order calculus (FROC), continuous-time random walk (CTRW), stretched-exponential, and statistical]. Feature selection and model training were performed using five randomly permuted five-fold cross-validations. An internal test set (15 cases of the primary dataset) and an external cohort (n=55) imaged on a different scanner were used to validate the model. The diagnostic performance of the model was compared with that of a single DWI model and DWI radiomics using accuracy, sensitivity, specificity, and the area under the curve (AUC).

Results: Seven significant multiparametric DWI features (two from the stretched-exponential and FROC models, and three from the CTRW model) were selected to construct the model. The multiparametric DWI model achieved the highest AUC (0.84, versus 0.71 for the single DWI model, $P<0.05$), an accuracy of 0.80 in the internal test, and both AUC and accuracy of 0.76 in the external test.

Conclusions: Our multiparametric DWI model differentiated low- (LGG) from high-grade glioma (HGG) with better generalization performance than the established single DWI model. This result suggests that the

^ ORCID: Junqi Xu, 0000-0002-9880-4191; Yan Ren, 0000-0001-5993-9248; Xueying Zhao, 0000-0001-5456-7609; Xiaoqing Wang, 0000-0002-4360-3253; Xuchen Yu, 0000-0002-0858-6582; Zhenwei Yao, 0000-0003-2390-6297; Yan Zhou, 0000-0001-9402-1109; Xiaoyuan Feng, 0000-0003-4525-7494; Xiaohong Joe Zhou, 0000-0003-0793-4925; He Wang, 0000-0002-2053-9439.

application of an ML approach with multiple DWI models is feasible for the preoperative grading of gliomas.

Keywords: Multiparametric diffusion-weighted imaging (DWI); machine learning (ML); glioma grading; magnetic resonance imaging (MRI)

Submitted Feb 16, 2022. Accepted for publication Aug 07, 2022.
doi: 10.21037/qims-22-145
View this article at: https://dx.doi.org/10.21037/qims-22-145

Introduction

Glioma is the most common neuroepithelial tumor of the cerebral nervous system and is classified into four grades by the World Health Organization (1,2). Low- (LGG) (grade II) and high-grade gliomas (HGG) (grades III and IV) differ in pathology and prognosis. In patients for whom an invasive procedure is considered feasible, the glioma grade is determined using stereotactic biopsy followed by histopathological analysis. However, the limitations of invasive procedures can lead to sampling errors, which can compromise the accuracy of diagnosis and the significant risks may be associated with the invasive procedure in some cases (3). Therefore, glioma grading through noninvasive medical imaging methods is needed to overcome these limitations.

Several previous studies have proposed grading gliomas based on quantitative parameters of magnetic resonance imaging (MRI) techniques, such as magnetic resonance (MR) spectroscopy, perfusion imaging, T2 mapping, and diffusion-weighted imaging (DWI). Of these methods, DWI is the most sensitive and has great potential for grading tasks (4-8). Many DWI models have been proposed over the past few years. One diffusion parameter, the apparent diffusion coefficient (ADC), is used to describe free diffusion with a monoexponential function, where the distribution of molecular displacements obeys a Gaussian law (9,10). However, different diffusion compartments may arise from the complex structure of tumor tissues (11-14). As a result, the diffusion displacement probability distribution can deviate substantially from Gaussian law (11). To overcome this dilemma, models incorporating multiple water diffusion components have been developed (11,15-18). For example, Le Bihan *et al.* (17) proposed the intravoxel incoherent motion (IVIM) model, which separates simple diffusion and microvascular perfusion in tissues. Bennett *et al.* (15) proposed the stretched-exponential model (SEM) and showed that signal attenuation is consistent with a multicompartmental theory of water diffusion in the brain. A statistical model (SM) to describe a considerable amount of diffusion-attenuated MR signals in biological systems has also been published (19). Diffusion kurtosis imaging (DKI) has previously been used to evaluate non-Gaussian water diffusion in bodily tissues (11,20), and in recent years, two advanced DWI models to measure tissue heterogeneity have also been proposed. Using a fractional order calculus (FROC) diffusion model has been shown to improve the accuracy of MR imaging in differentiating benign and malignant pediatric brain tumors and grading adult gliomas (9,21). Significant differences between malignant and benign pediatric tumors have also been observed using the continuous-time random walk (CTRW) model (22). However, these models have not been tested for reproducibility with other test sets, nor has the value of combining multiple DWI models in glioma grading been discussed (23). Therefore, it would be helpful to investigate whether combining multiple DWI models can improve their performance in grading gliomas.

Previous studies have utilized multimodality MRI radiomics with machine learning (ML) approaches to classify gliomas, demonstrating that DWI features might improve diagnostic accuracy (24,25). The results of one study showed that incorporating diffusion-weighted MRI into an ML-based radiomics model could improve the diagnosis of pseudoprogression in patients with glioblastoma (5). Another study also highlighted the potential of diffusion MR with radiomics analysis in the evaluation of glioma malignancy (26). Based on this prior work, we hypothesized that combining multiple DWI models with ML algorithms might improve the performance of DWIs in differentiating glioma grades.

In this study, information extracted from multiple diffusion models was combined and subjected to ML-based analysis to improve the performance of diffusion imaging in glioma grading. Further, the robustness of the proposed method was compared with that of the traditional single DWI model and DWI radiomics, and the results of

these comparisons are presented herein. We present the following article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-22-145/rc).

Methods

The current study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The institutional review boards of Hua Shan Hospital affiliated with Fudan University and Ren Ji Hospital affiliated with Shanghai Jiao Tong University approved this retrospective study and waived the requirement to obtain informed consent.

Core codes are available at https://github.com/Arroway-JQ/combined-DWI-modeling-and-tumor-classification.

Patients

This study recruited 147 adult patients with gliomas from Hua Shan Hospital affiliated with Fudan University between 2014 and 2015. A further 69 patients were recruited from Ren Ji Hospital affiliated with Shanghai Jiao Tong University between 2016 and 2019 as an external test cohort.

The inclusion criteria for the study were as follows: (I) three types of MRI images [T1-weighted images with enhancement (T1WI+C), T2-weighted fluid-attenuated inversion recovery (T2W-FLAIR), and DWI] were available for evaluation; (II) surgery was performed for a pathologic diagnosis after MR imaging and integrated clinical information was obtained; (III) the DWI scan was performed using the correct number of b-values (21 for the first dataset and 17 for the second). After these criteria had been applied, the primary dataset consisted of 74 patients (18–75 years old), including 15 patients as the internal test set. Of the patients in the primary dataset, 37 had LGG and 37 had HGG according to the World Health Organization classification (2). The external test set comprised 55 patients (14–78 years old), of whom 25 patients had LGGs, with the remaining patients having HGGs.

Acquisition of MRI scans

The MRI scans (T1WI+C, T2W-FLAIR, and DWI sequences) were performed on two 3.0 tesla scanners (MR750, Signa HDxt, GE Medical System, Milwaukee, WI, USA) using a standard eight-channel phased-array head coil. The DWI was acquired using a single-shot spin-echo planar imaging sequence with 21 and 17 b-values at Hua Shan Hospital and Ren Ji Hospital, respectively. Diffusion gradients were applied in all three orthogonal directions (x-, y-, and z-axes) to obtain a trace-weighted image to minimize the influence of diffusion anisotropy. Other core image acquisition parameters are shown in *Figure 1*.

Image preprocessing

The entire process for building the prediction model is shown in *Figure 2*.

The diffusion images were eddy-current corrected, and the skulls were removed through MRI tools using the Functional Magnetic Resonance Imaging of the Brain Software Library (FSL) (27). Subsequently, a median filter was used to smoothen and denoise the images. The diffusion-attenuated signals were acquired at the voxel level, and then the signal intensity was normalized to the signal intensity of the b0 image.

With reference to T1WI+C images, two radiologists placed regions of interest (ROIs) on the solid part of tumors on the b =0 DWI images, avoiding necrosis, edema, and hemorrhage. The ROIs were then propagated to each slice of the parameter maps. For the external test set, a single radiologist at the second hospital read the diffusion images of multiple b-values, and the diagnostic accuracy was 0.71 (39 of 55 cases were predicted the same as the ground truth) with an area under the receiver operating characteristic (ROC) curve (AUC) of 0.71 (95% confidence interval: 0.57–0.86).

Multi-DWI models

Based on the above theories, the present study applied seven diffusion models (ADC, IVIM, SEM, SM, DKI, FROC, and CTRW) using MATLAB (MathWorks, Inc., 2019b, Natick, MA, USA).

The monoexponential model is described as Eq. [1], where S denotes the signal intensity with diffusion sensitization, and S_0 denotes the signal intensity without sensitization.

$$S = S_0 \exp\left(-b \times ADC\right) \quad [1]$$

The IVIM model was fitted according to Eq. [2], where f is the perfusion fraction; D_f is the pseudodiffusion coefficient, which represents faster diffusion; and D_s is the actual diffusion coefficient, which represents slower diffusion (28).

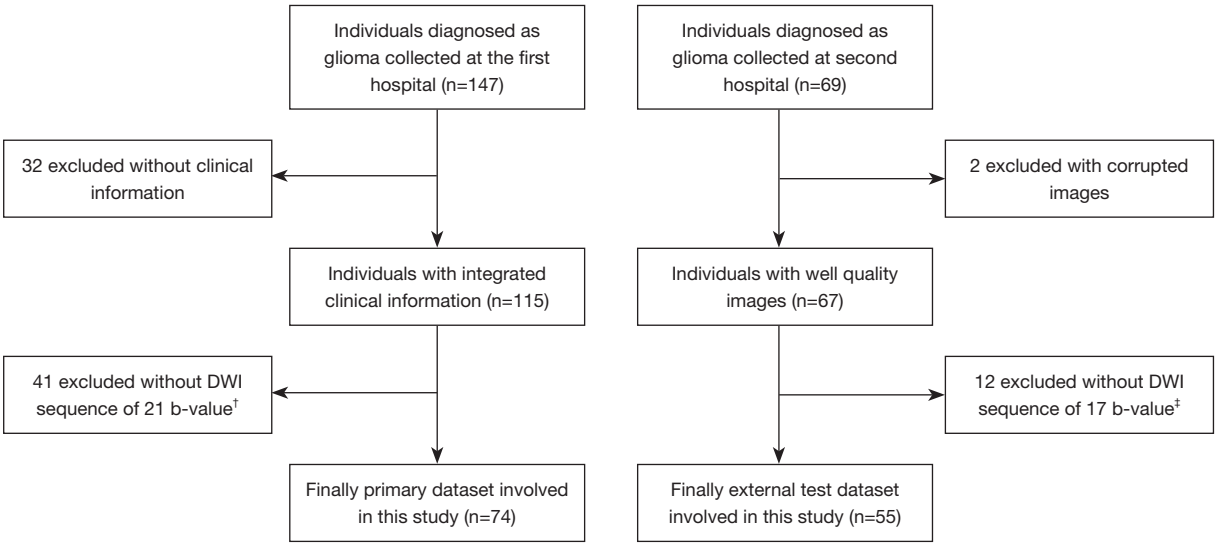

Figure 1 Flowchart of the study inclusion and exclusion process. †, 21 b-value: 0, 10, 20, 30, 50, 100, 150, 200, 300, 400, 500, 600, 800, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, and 4,500 s/mm^2. Other core image acquisition parameters for Hua Shan Hospital were as follows: partial Fourier, average times NEX =2 (=4 for b =3,500–4,500 s/mm^2), TR =5,000 ms, TE =90.6 ms, separation between two diffusion gradient lobes Δ =42.688 ms, duration of each diffusion gradient δ =29.404 ms, slice thickness =4 mm, acquisition matrix size =128×128 zeros-padded to 256×256, flip angle =90° and pixel size =1×1 mm^2; ‡, 17 b-value: 0, 20, 50, 80, 150, 200, 300, 500, 800, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, and 4,500 s/mm^2. Other core image acquisition parameters for Ren Ji Hospital were as follows: partial Fourier, acceleration =2, TR =3,000 ms, TE =105.8 ms, separation between two diffusion gradient lobes Δ =42.688 ms, duration of each diffusion gradient δ =29.404 ms, slice thickness =6 mm, acquisition matrix size =192×192, flip angle =90°, and pixel size =1×1 mm^2. DWI, diffusion-weighted imaging; TR, repetition time; TE, echo time.

$$S = S_0\left(f\exp\left(-bD_f\right)+\left(1-f\right)\exp\left(-bD_s\right)\right) \quad [2]$$

The SEM model is presented as Eq. [3], where $\alpha \in (0,1)$ represents the deviation of the signal attenuation (15), and *DDC* is the distributed diffusion coefficient.

$$S = S_0\exp\left(-\left(b\times DDC\right)^{\alpha}\right) \quad [3]$$

The SM model is described as Eq. [4], where σ is the distribution width and ADC_S is the position of the distribution maxima (19).

$$S = S_0\exp\left(-bADC_S+\frac{1}{2}\sigma^2b^2\right) \quad [4]$$

The DKI model was applied according to Eq. [5] using additional information on the diffusion kurtosis *K*.

$$S = S_0\exp\left(-bD_K+\frac{1}{6}b^2D_K^2K\right) \quad [5]$$

The FROC model is presented as Eq. [6], where δ is the diffusion gradient pulse width, Δ is the gradient lobe separation, β_f^* correlates with tissue heterogeneity, and μ is the microstructural quantity (21).

$$S = S_0\exp\left\{-D\mu^{2\left(\beta_f^*-1\right)}\left(\frac{b}{\Delta-\delta/3}\right)^{\beta_f^*}\left(\Delta-\frac{2\beta_f^*-1}{2\beta_f^*+1}\delta\right)\right\} \quad [6]$$

The CTRW model was written using the Mittag-Leffler function (MLF) as in Eq. [7], where D_c denotes the anomalous diffusion coefficient, α_c and β_c represent the diffusion heterogeneity of time and space (22), respectively.

$$S = S_0E_{\alpha_c}\left(-\left(bD_c\right)^{\beta_c}\right) \quad [7]$$

All DWI models were applied voxel by voxel with the R-squared (R^2) value recorded to evaluate the goodness of fit. The ADC, SM, and DKI models were calculated using polynomial fitting, and the others were fitted by applying the Levenberg-Marquardt algorithm (29). In total, 16 parameters were derived from the seven models (*ADC*, *f*, D_s, D_f, *DDC*, α, ADC_S, σ, D_K, *K*, *D*, μ, β_f^*, D_c, β_c, α_c).

Feature extraction

The primary dataset was randomly stratified into training (n=59) and test (n=15) sets at a ratio of 8 to 2. Five

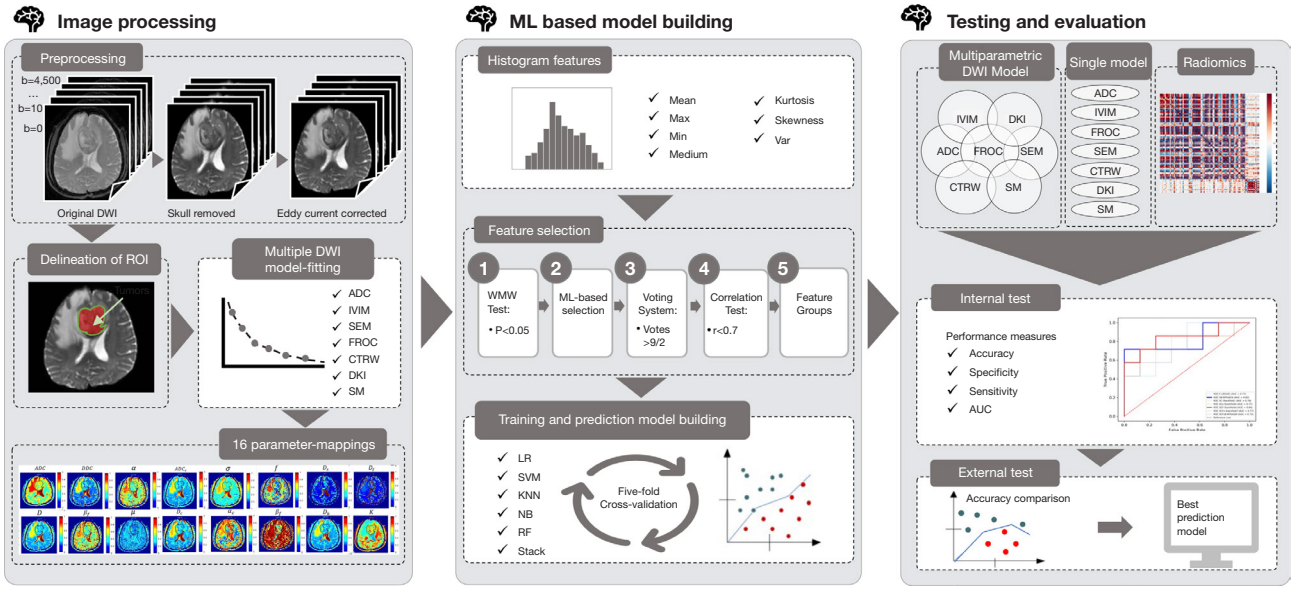

Figure 2 Flow chart for all the procedures to predict LGG and HGG in this study. The first stage is image processing, including DWI model building and parameter mapping. The second stage is ML-based model building. In this part, seven histogram features were extracted and selected using a five-step procedure: a two-sided Wilcoxon-Mann-Whitney U-test, ML feature selection, a voting system, a correlation test, and feature combination. Prediction models were trained and selected in the primary dataset. The third stage included validation and evaluation of our proposed model and traditional DWI models in the internal and external cohorts. DWI, diffusion-weighted imaging; ROI, region of interest; ADC, apparent diffusion coefficient; IVIM, intravoxel incoherent motion; SEM, stretched-exponential model; FROC, fractional order calculus; CTRW, continuous-time random walk; DKI, diffusion kurtosis imaging; SM, statistical model; ML, machine learning; LR, logistic regression; SVM, support-vector machine; KNN, K-nearest neighbors; NB, naïve Bayes; RF, random forests; AUC, area under the curve; LGG, low-grade glioma; HGG, high-grade glioma.

randomly permuted five-fold cross-validations were used to evaluate our method. Balance of the training and test data was considered. Thus, the ratio of HGGs and LGGs was maintained between the training, validation, and test sets. An external test data set (n=55) from another medical center was also included for another verification.

The DWI signals of ROIs were filtered by ranking their R^2 value from curve fitting, retaining only the top 95% of voxels for each tumor. The mean, maximum, minimum, median, kurtosis, skewness, and variance values were calculated for each parameter and each patient in the primary and external datasets. In all, our model extracted 112 (16×7) features from each case.

The DWI radiomic features were extracted using the PyRadiomics package in Python software (v. 3.6, Python Software Foundation, Wilmington, DE, USA, https://www.python.org/) (30). Feature scaling was performed on both the primary and external datasets using Z-score transformation (31). In this study, 7,076 and 6,100 features were extracted from each case in the primary dataset and the external test set, respectively (Appendix 1).

Feature selection

All work in this section was accomplished using an open ML library and scikit-learn (ver. 0.22) in Python (32). The clinical information for the training and test datasets is shown in *Table 1*.

Feature selection was performed on five-fold cross-validation sets. A rigorous five-step process (*Figure 2*) was implemented to select significant features and avoid collinearity. A two-sided Wilcoxon-Mann-Whitney U-test was performed to identify features which were significantly different (P<0.05) between HGGs and LGGs for subsequent analysis. Then, the most significant features (votes >9/2) were selected using nine ML methods: logistic regression, support-vector machine, K-nearest neighbors, random forest for a single feature, random forest for all features, naïve Bayes, stacking, recursive-feature elimination, and Least Absolute Shrinkage and Selection

Table 1 Clinical information of patients in the training and test datasets

Data	LGG	HGG	Age, years, mean ± SD	Male	Female
Training set (n=59)	29	30	46±15	40	19
Internal test set (n=15)	8	7	43±14 (P=0.45)‡	9	6
Primary set (n=74)	37	37	45±14	49	25
External test set (n=55)	25	30	49±15 (P=0.26)‡	31	24

‡, P value no significant different in patient age between the training and internal test sets or the training and external test sets. LGG, low-grade glioma; HGG, high-grade glioma; SD, standard deviation.

Operator (see Figure S1). After that, the retained features were subjected to correlation tests to eliminate potentially collinear features (r>0.70), which introduce redundant information to the prediction model (Table S1). Finally, the features were combined into subgroups based on their corresponding DWI models (Table S2).

The feature subgroups based on the single DWI model consisted of the mean value of each parameter without undergoing ML selection (see Appendix 2, which describes the feature selection for the two other DWI methods).

Training and estimator selection

Six estimators (logistic regression, support-vector machine, K-nearest neighbors, random forest, naïve Bayes, and stacking) were used to construct the classification models and to learn how best to combine the predictions from the above base machine as the new features. They were reclassified with logistic regression as a metaclassifier. Stacking is an ensemble ML algorithm that uses meta-learning. The benefit of stacking is that it can harness the capabilities of a range of well-performing models by using their output as input and ultimately achieve a better predictive performance than any single model in the ensemble (33). For each estimator, a grid search was conducted for automatic parameter tuning.

Prediction models were first trained in the five-fold cross-validation set (34). Then, the final prediction models for the three classification methods were selected based on the highest AUC in the internal test set (see Table S3 for integrated training and internal testing results).

Testing and comparisons

Internal and external test sets were included to evaluate the performances of the single DWI, DWI radiomics, and ML-based multiparametric DWI prediction models. To determine the models' accuracy, the AUCs were calculated as evaluation indices. The cutoff values that provided the best sensitivity and specificity were determined according to the maximum value of the Youden index (35). Differences between the three models were compared. The AUCs of the three models were compared using the DeLong test (36-38).

Results

Multimodel DWI fitting

The DWI signal attenuation curves were fitted based on the theoretical bases of the seven diffusion models. *Figure 3* shows the maps of the 16 parameters obtained compared to b0 images of an LGG and an HGG.

The SEM, FROC, and CTRW models outperformed the other models (Table S4), with R^2_{mean} 0.9959, 0.9788, and 0.9801, respectively. The R^2 values of the SM and DKI models were similar, while those of ADC and IVIM models were relatively lower than the other models (<0.96).

Significant features

After five-step feature selection, only ten features were selected as significant. The subgroups of these features and combination descriptions are shown in *Table 2* and Table S2. Based on DWI radiomics analysis, five sequential texture features and four wavelet transformations were selected after a similar feature selection process. For the construction of the single DWI prediction model, only the mean values of parameters were chosen (see Table S5, which demonstrates the feature combinations of the two established DWI methods).

Optimal prediction models

Figure 4A shows the ROC curves of cross-validations of

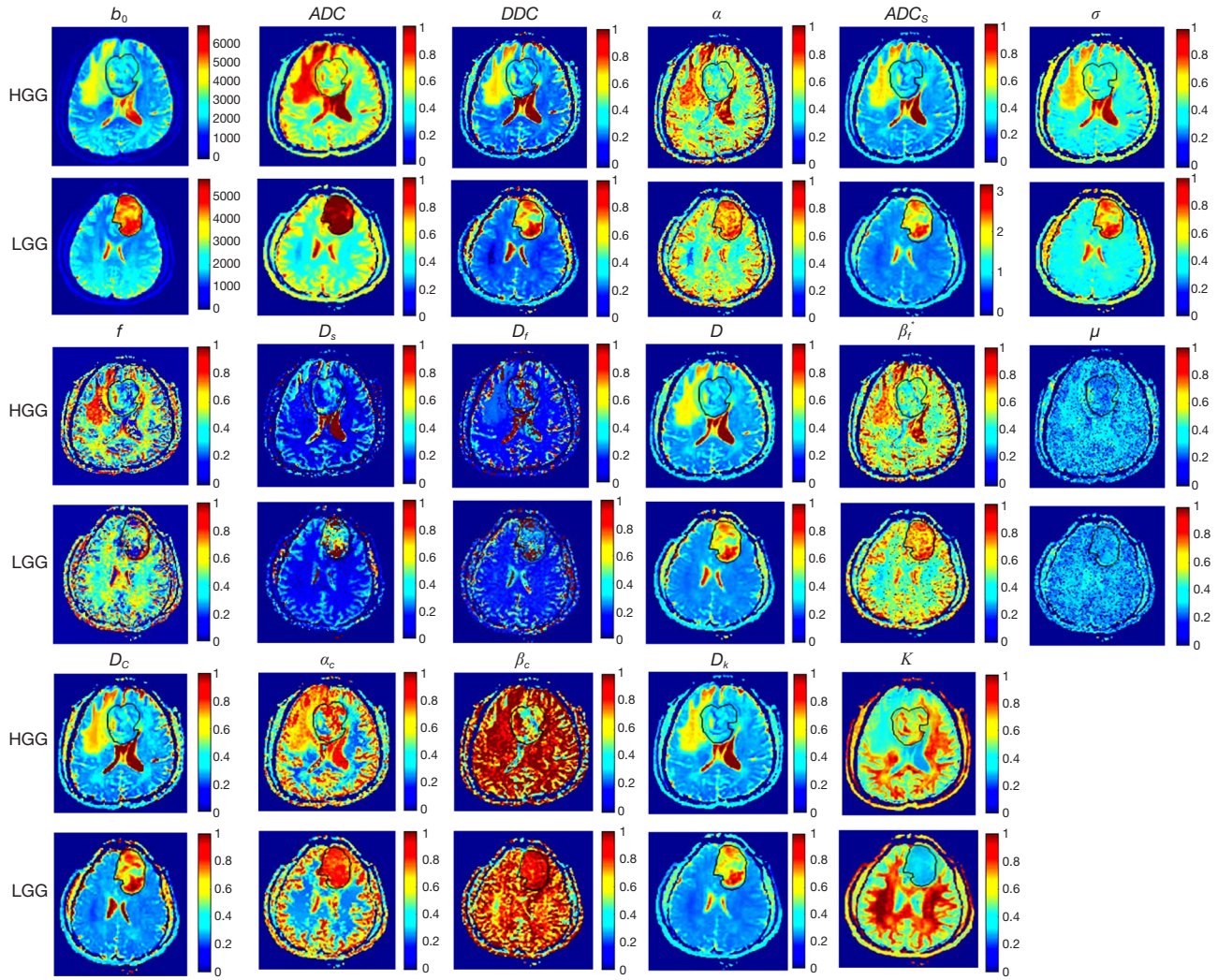

Figure 3 Sixteen parameter maps for HGG and LGG cases. The images in sequence are: b_0 map; *ADC* map; *DDC* map and α map for SEM; ADC_S map and σ map for SM; f map, D_s map, and D_f map for IVIM model; D map, β_f^* map, and μ map for FROC model; D_C map, α_c map, and β_c map for CTRW model; and D_K map and K map for DKI model. ADC, apparent diffusion coefficient; HGG, high-grade glioma; LGG, low-grade glioma; SEM, stretched-exponential model; SM, statistical model; IVIM, intravoxel incoherent motion; FROC, fractional order calculus; CTRW, continuous-time random walk; DKI, diffusion kurtosis imaging.

the combined SEM, CTRW, and FROC (SCF) model containing seven features (SEM model with *DDC*_min and α_skewness, CTRW model with α_c_kurtosis, α_c_variance and β_c_variance, and FROC model with β_f^* min and μ_skewness as features). Then, the fourth-fold stack estimator for SCF was chosen as the final prediction model.

As shown in *Figure 4B*, the SCF model had the highest AUC (0.84) (sensitivity =0.86 and specificity =0.75) in the internal test set. *Figure 4C* shows that the SEM model had the highest AUC value (0.71) among the single DWI models. Furthermore, the subgroup TOP6 had the best performance (AUC =0.84) among the DWI radiomics features, as shown in *Figure 4D* and Table S6. Both the SCF model and the radiomics model significantly improved the predictive performance of the single DWI model ($P=8.60\times10^{-4}$, 1.90×10^{-4} for DeLong test, respectively).

Compared with the established methods, our method performed better in the external cohort. As shown in *Table 3*, in the external cohort, the SCF model showed both a higher accuracy and AUC value than the SEM

Table 2 Subgroups of selected 10 features in the multiparametric DWI model

S (SEM)
*DDC*_min
α_skewness
C (CTRW)
α_c_kurtosis
α_c_variance
β_c_variance
F (FROC)
β^*_f_min
μ_skewness
S (SM)
σ_skewness
I (IVIM)
D_s_mean
D_s_min

DWI, diffusion-weighted imaging; SEM, stretched-exponential model; CTRW, continuous-time random walk; FROC, fractional order calculus; SM, statistical model; IVIM, intravoxel incoherent motion.

(DDC_mean and α_mean in SEM model) and TOP6 models (accuracy =0.76, 0.53, and 0.67, respectively). *Table 4* shows that the SCF model performed significantly better in classifying the external test set than did the SEM model (AUC =0.76 and 0.53, respectively, P=0.02 for the DeLong test). The AUC of the SCF model was higher than that of the DWI radiomics model, but the difference was not significant (AUC =0.72, P=0.61 for the DeLong test) (*Figure 5*).

Discussion

In this study, a multiparametric DWI model to differentiate LGGs and HGGs was proposed. We used images with multiple high b-values to extract higher-order features from 16 parameters derived from seven DWI models proposed in previous studies (7,15,17,19-22). Features were selected by using ML algorithms and statistical analyses. We found that the SCF prediction model performed best in both the primary dataset and the external test set. The robustness of our method was evaluated in the external test set and compared with that of other methods, and the proposed method was found to have advantages over the two established DWI methods.

Multiparametric DWI model

Based on different approaches to diffusion imaging, seven DWI models were incorporated into our model. As shown in *Figure 3*, the *ADC*, *DDC*, *Dc*, *D*, ADC_S, and D_K maps share similar areas of contrast, reflecting similar water diffusion distribution in the tissues, whereas the α, β^*_f, α_c, and β_c maps show tissue heterogeneity; these findings are consistent with results from previous studies (21,22). The SEM and CTRW models reflected microstructure characteristics better with high b values than other models and thus, had better fitting quality for signal attenuation (R^2=0.9959, 0.9801, respectively, see Table S4). In line with the findings of Niendorf *et al.* (18), the monoexponential model had the worst performance of all the models, with the fitted curve noticeably deviating from the original curve in the high b-value region.

In this study, the ML approach was used, and features from multiple DWI models were combined. Due to the incorporation of multiple features into our model, the problem of overfitting had to be considered. To mitigate the risk of overfitting, we adopted two preventive measures. The first was the use of a rigorous five-step feature selection procedure with a reduced number of features. The second was the use of independent internal and external test data sets.

Comparisons with established DWI methods

Instead of focusing on multimodality radiomics, this work focused on investigating and improving the diagnostic potency of DWI models. Our study demonstrated that multimodel DWI was useful, with our results being comparable to those of another multimodality radiomics study, in which the AUC of the external cohort was 0.75 (34). We agree that advanced MRI sequences like diffusion imaging can provide meticulous radiologic information about glioma and may be suitable for a predictive model (5). Many previous studies have demonstrated the resolving ability of single DWI models. For example, one study reported the value of DKI with radiomics in grading gliomas (26). However, further investigation using a larger sample and an external test set is still needed. Compared with models in two previous

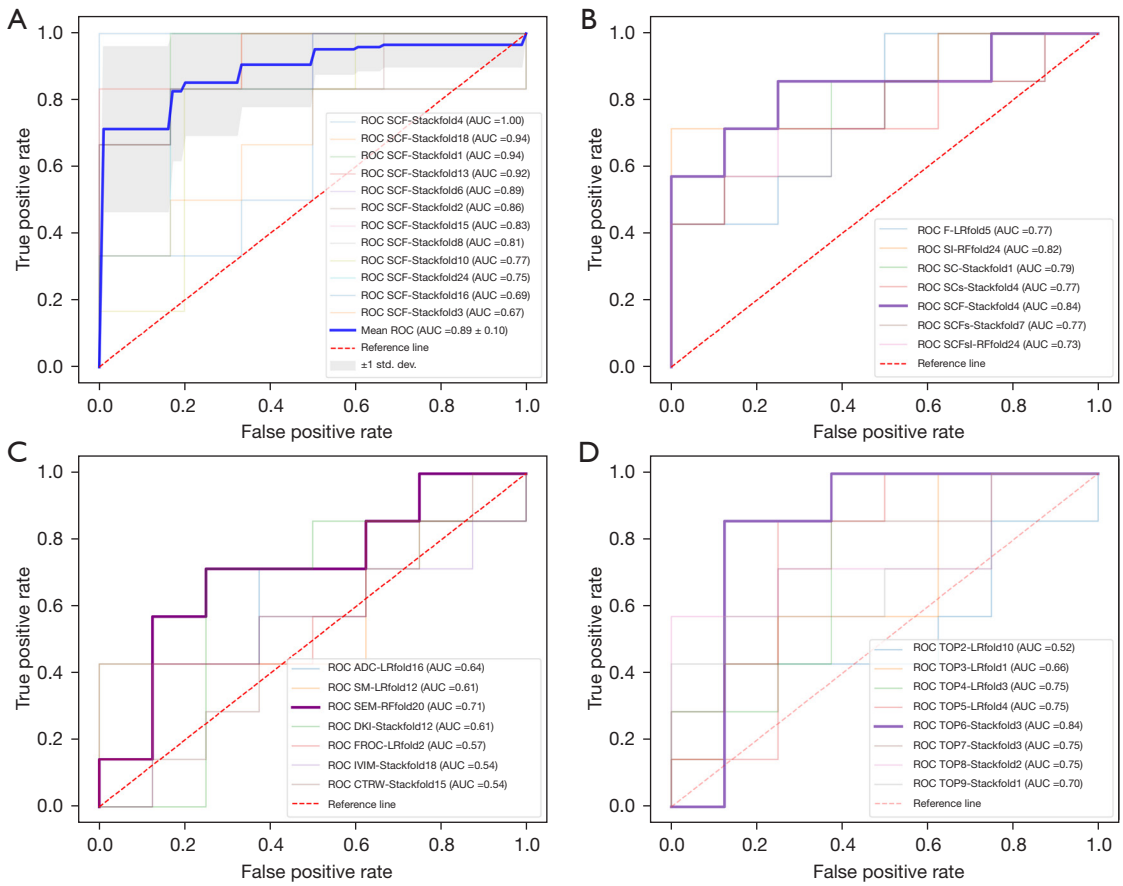

Figure 4 ROC curves of different models in the internal test and external validation sets. (A) The ROC curve of five randomly permuted five-fold cross-validation sets using SCF as input and stack as the estimator. (B) The ROC curve of feature combinations based on the multiparametric DWI model and ML selection in the internal test set. (C) The ROC curve of feature combinations based on the single DWI models in the internal test set. (D) The ROC curve of feature combinations based on DWI radiomics in the internal test set. ROC, receiver operating characteristic; SCF, SEM, CTRW, and FROC models; AUC, area under the curve; F, FROC model; SI, SEM and IVIM models; SC, SEM and CTRW models; SCs, SEM, CTRW, and statistic models; LR, logistic regression; RF, random forest; ADC, apparent diffusion coefficient; SM, statistical model; SEM, stretched-exponential model; DKI, diffusion kurtosis imaging; FROC, fractional order calculus; IVIM, intravoxel incoherent motion; CTRW, continuous-time random walk; DWI, diffusion-weighted imaging; ML, machine learning.

Table 3 The predictive accuracy of the proposed model, the single DWI model, and DWI radiomics model in the internal and external test sets

Model	Feature combination	Feature-num	Prediction estimator	Train CV-mean accuracy	Train CV-mean AUC	Internal test accuracy	External test accuracy
Multiparametric DWI	SCF[†]	7	Stackfold4	0.91	0.89	0.80	0.76
Single DWI	SEM[‡]	2	RFfold20	0.79	0.79	0.73	0.53
DWI radiomics	TOP6[§]	6	Stackfold3	0.96	0.98	0.80	0.67

[†], "SCF" means SEM with *DDC*_min and α_skewness, CTRW model with α_c_kurtosis, α_c_variance and β_c_variance, FROC model with β^*_f min and μ_skewness; [‡], "SEM" denotes stretched-exponential model, i.e., the mean values of σ, *DDC* in SEM model; [§], "TOP6" means radiomics features: kurtosis of the minor axis length calculated from all b-value images, maximum of HHL calculated based on b = 3,500 s/mm^2 images, skewness of the minor axis length calculated from all b-value images, kurtosis of HHL calculated based on b = 0 s/mm^2 images, kurtosis of the interquartile range of HHH calculated from all b-value images, kurtosis of HHH calculated based on b = 4,000 s/mm^2 images. "H" and "L" denote high-pass and low-pass filters, respectively. DWI, diffusion-weighted imaging; CV, cross-validation sets; SEM, stretched-exponential model; CTRW, continuous-time random walk; FROC, fractional order calculus.

Table 4 The AUC, sensitivity, and specificity of the proposed model, the single DWI model, and DWI radiomics model in internal and external test sets

Model	Internal test set			External test set		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Multiparametric DWI	0.84	0.86	0.75	0.76	0.80	0.68
Single DWI	0.71	0.71	0.75	0.53 (P=0.02)†	0.43	0.60
DWI radiomics	0.84	0.86	0.88	0.72 (P=0.61)‡	0.50	0.92

†, DeLong test between the multiparametric DWI and single DWI models with P<0.05; ‡, DeLong test between the multiparametric DWI and DWI radiomics models with P>0.1. DWI, diffusion-weighted imaging; AUC, area under the curve.

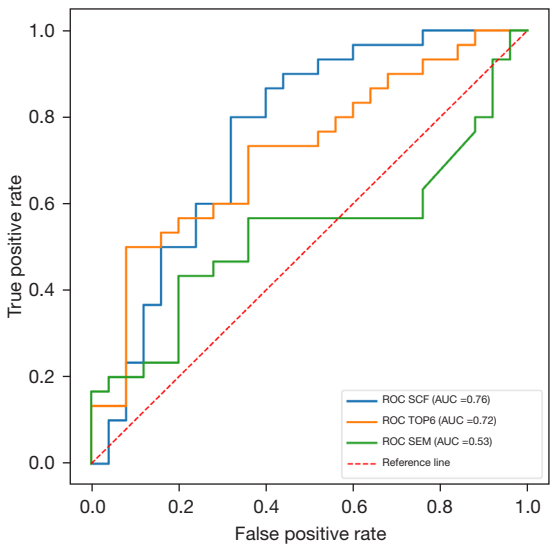

Figure 5 ROC curves of the multiparametric DWI model, the single DWI model, and the DWI radiomics model in the external test set. ROC, receiver operating characteristic; SCF, SEM, CTRW, and FROC models; AUC, area under the curve; SEM, stretched-exponential model; DWI, diffusion-weighted imaging; CTRW, continuous-time random walk; FROC, fractional order calculus.

studies focusing on diffusion MRI in glioma grading (accuracy =0.80, 0.82) (9,39), our model performed better in cross-validation sets (accuracy =0.91), and neither of these studies included an independent test set. Furthermore, the repeatability of our multiparametric DWI model was demonstrated in both the internal and external test sets. In this study, the AUCs of the multiparametric DWI model and the radiomics model were significantly higher (P<0.05) than that of the single DWI model (AUC =0.73) in the internal test set. As shown in *Table 3*, the single DWI model showed a sharp decrease in performance in the external test set (AUC =0.53), while our multiparametric DWI model showed a superior performance in the external test set (AUC =0.76, P=0.02). These results indicate that measuring the mean value of parameters within ROIs in tissues based on the single DWI model might fail to sufficiently capture the tumor complexity; thus, this method would have limited applicability to other datasets.

Although the SCF (our method) and TOP6 (the DWI radiomics method) models had comparable accuracy (0.80, P>1) in the internal test set, the accuracy of the SCF model (0.76) was much higher than that of the TOP6 model (0.67) in the external test set. The AUC values of the SCF model in the external test set were decreased compared to those in the internal test set on account of the decreasing of specificity in the external test set. But, the sensitivity remained high at 0.80. The accuracy and AUC value of the SCF model were higher than those of the DWI radiomics model in the external test set; however the AUC showed no significant difference between the two models. Also, compared to that in the internal test set, the sensitivity of the DWI radiomics model in the external test set decreased substantially to 0.5, while the specificity remained high at 0.92. Our findings demonstrate that quantitative analysis using our multiparametric DWI model may be more generalizable than signal analysis of images (the DWI radiomics model).

Our study has several limitations. First, differences in scanning parameters, such as b-values and the echo time (TE), between the training and external sets may have introduced biases which impacted the accuracy of the external test set results. In our study, the accuracy of the SCF model decreased by 0.04 in the external test set relative to the internal test set, although we used feature scaling (32), applied a regularization algorithm for feature selection, and used cross-validation to evaluate the model's generalization error and to select the estimator (40,41).

Subsequent studies should be conducted by augmenting the number of samples, and a standard methodology of normalization between different cohorts also needs further investigation. Furthermore, the DWI data in this work were collected in three orthogonal directions, which did not meet the requirements for computing some of the direction-dependent matrices, such as $k_{\perp}$ and $k_{||}$, in the DKI model. The limited diffusion directions could have affected the isotropic K, which may be one of the reasons for the poor performance of the DKI model. Also, multiple diffusion directions, if clinically available, make it possible to analyze other diffusion models, such as neurite orientation dispersion and density imaging, diffusion basis spectrum imaging, and constrained diffusional variance decomposition models (42-44).

Conclusions

In conclusion, our multiparametric DWI model with an ML algorithm was found to be feasible and valuable for predicting LGGs and HGGs. Multiple DWI parameters can provide abundant critical information for clinical diagnosis. Compared to that of the single DWI model, the performance of the SCF model in glioma classification was significantly improved (P<0.05), with our model achieving higher accuracy and AUC values in both the internal (accuracy =0.80, AUC =0.84) and external (accuracy =0.76, AUC =0.76) test sets.

In summary, our method is credible and robust for differentiating LGGs and HGGs in adults. The promising results of this study will pave the way for further research combining other diffusion models and involving larger patient groups.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China (No. 81971583 to HW), Shanghai Natural Science Foundation (No. 20ZR1406400 to HW), Science and Technology Support Project for Medicine sponsored by Science and Technology Commission of Shanghai Municipality (No. 18411967300 to HW), and Shanghai Municipal Science and Technology Major Project (Nos. 2017SHZDZX01 and 2018SHZDZX01 to HW).

Footnote

Reporting Checklist: The authors completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-22-145/rc

Conflicts of Interest: All authors completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-145/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 current study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The institutional review boards of Hua Shan Hospital affiliated with Fudan University and Ren Ji Hospital affiliated with Shanghai Jiao Tong University approved this retrospective study, and the requirement to obtain informed consent was waived.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P, Ellison DW. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol 2016;131:803-20.
3. Jackson RJ, Fuller GN, Abi-Said D, Lang FF, Gokaslan

ZL, Shi WM, Wildrick DM, Sawaya R. Limitations of stereotactic biopsy in the initial management of gliomas. Neuro Oncol 2001;3:193-200.
4. Provenzale JM, Mukundan S, Barboriak DP. Diffusion-weighted and perfusion MR imaging for brain tumor characterization and assessment of treatment response. Radiology 2006;239:632-49.
5. Kim JY, Park JE, Jo Y, Shim WH, Nam SJ, Kim JH, Yoo RE, Choi SH, Kim HS. Incorporating diffusion- and perfusion-weighted MRI into a radiomics model improves diagnostic performance for pseudoprogression in glioblastoma patients. Neuro Oncol 2019;21:404-14.
6. Kono K, Inoue Y, Nakayama K, Shakudo M, Morino M, Ohata K, Wakasa K, Yamada R. The role of diffusion-weighted imaging in patients with brain tumors. AJNR Am J Neuroradiol 2001;22:1081-8.
7. Bulakbasi N, Guvenc I, Onguru O, Erdogan E, Tayfun C, Ucoz T. The added value of the apparent diffusion coefficient calculation to magnetic resonance imaging in the differentiation and grading of malignant brain tumors. J Comput Assist Tomogr 2004;28:735-46.
8. Gu W, Fang S, Hou X, Ma D, Li S. Exploring diagnostic performance of T2 mapping in diffuse glioma grading. Quant Imaging Med Surg 2021;11:2943-54.
9. Sui Y, Xiong Y, Jiang J, Karaman MM, Xie KL, Zhu W, Zhou XJ. Differentiation of Low- and High-Grade Gliomas Using High b-Value Diffusion Imaging with a Non-Gaussian Diffusion Model. AJNR Am J Neuroradiol 2016;37:1643-9.
10. Le Bihan D. The‘’wet min”: water and functional neuroimaging. Phys Med Biol 2007;52:R57-90.
11. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magn Reson Med 2005;53:1432-40.
12. Le Bihan D. Intravoxel incoherent motion perfusion MR imaging: a wake-up call. Radiology 2008;249:748-52.
13. Wáng YXJ. Mutual constraining of slow component and fast component measures: some observations in liver IVIM imaging. Quant Imaging Med Surg 2021;11:2879-87.
14. Wáng YXJ. A reduction of perfusion can lead to an artificial elevation of slow diffusion measure: examples in acute brain ischemia MRI intravoxel incoherent motion studies. Ann Transl Med 2021;9:895.
15. Bennett KM, Schmainda KM, Bennett RT, Rowe DB, Lu H, Hyde JS. Characterization of continuously distributed cortical water diffusion rates with a stretched-exponential model. Magn Reson Med 2003;50:727-34.
16. Le Bihan D. Looking into the functional architecture of the brain with diffusion MRI. Nat Rev Neurosci 2003;4:469-80.
17. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology 1988;168:497-505.
18. Niendorf T, Dijkhuizen RM, Norris DG, van Lookeren Campagne M, Nicolay K. Biexponential diffusion attenuation in various states of brain tissue: implications for diffusion-weighted imaging. Magn Reson Med 1996;36:847-57.
19. Yablonskiy DA, Bretthorst GL, Ackerman JJ. Statistical model for diffusion attenuated MR signal. Magn Reson Med 2003;50:664-9.
20. Wang X, Gao W, Li F, Shi W, Li H, Zeng Q. Diffusion kurtosis imaging as an imaging biomarker for predicting prognosis of the patients with high-grade gliomas. Magn Reson Imaging 2019;63:131-6.
21. Sui Y, Wang H, Liu G, Damen FW, Wanamaker C, Li Y, Zhou XJ. Differentiation of Low- and High-Grade Pediatric Brain Tumors with High b-Value Diffusion-weighted MR Imaging and a Fractional Order Calculus Model. Radiology 2015;277:489-96.
22. Karaman MM, Sui Y, Wang H, Magin RL, Li Y, Zhou XJ. Differentiating low- and high-grade pediatric brain tumors using a continuous-time random-walk diffusion model at high b-values. Magn Reson Med 2016;76:1149-57.
23. Bai Y, Lin Y, Tian J, Shi D, Cheng J, Haacke EM, Hong X, Ma B, Zhou J, Wang M. Grading of Gliomas by Using Monoexponential, Biexponential, and Stretched Exponential Diffusion-weighted MR Imaging and Diffusion Kurtosis MR Imaging. Radiology 2016;278:496-504.
24. Qin JB, Liu Z, Zhang H, Shen C, Wang XC, Tan Y, Wang S, Wu XF, Tian J. Grading of Gliomas by Using Radiomic Features on Multiple Magnetic Resonance Imaging (MRI) Sequences. Med Sci Monit 2017;23:2168-78.
25. Su C, Jiang J, Zhang S, Shi J, Xu K, Shen N, Zhang J, Li L, Zhao L, Zhang J, Qin Y, Liu Y, Zhu W. Radiomics based on multicontrast MRI can precisely differentiate among glioma subtypes and predict tumour-proliferative behaviour. Eur Radiol 2019;29:1986-96.
26. Takahashi S, Takahashi W, Tanaka S, Haga A, Nakamoto T, Suzuki Y, Mukasa A, Takayanagi S, Kitagawa Y, Hana T, Nejo T, Nomura M, Nakagawa K, Saito N. Radiomics Analysis for Glioma Malignancy Evaluation Using Diffusion Kurtosis and Tensor Imaging. Int J Radiat Oncol

Biol Phys 2019;105:784-91.
27. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. Neuroimage 2012;62:782-90.
28. Federau C, Meuli R, "Brien K, Maeder P, Hagmann P. Perfusion measurement in brain gliomas with intravoxel incoherent motion MRI. AJNR Am J Neuroradiol 2014;35:256-62.
29. Press WH, T Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes: the art of scientific computing. 3rd edition. Cambridge: Cambridge University Press, 2007.
30. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77:e104-7.
31. Loizou CP, Pantziaris M, Seimenis I, Pattichis CS. Brain MR Image Normalization in Texture Analysis of Multiple Sclerosis. 2009 9th Internatianl Conference onInformation Technology and Applications in Biomedicine. Larnaca: IEEE, 2009:1-5.
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel T, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825-30.
33. Wolpert DH. Stacked generalization. Neural Networks 1992;5:241-59.
34. Nakamoto T, Takahashi W, Haga A, Takahashi S, Kiryu S, Nawa K, Ohta T, Ozaki S, Nozawa Y, Tanaka S, Mukasa A, Nakagawa K. Prediction of malignant glioma grades using contrast-enhanced T1-weighted and T2-weighted magnetic resonance images based on a radiomic analysis. Sci Rep 2019;9:19411.
35. Kallner A. Formulas. In: Laboratory Statistics. 2nd edition. Amsterdam: Elsevier, 2018:1-140.
36. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.
37. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. 2nd edition. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2011.
38. Sun X, Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. IEEE Signal Processing Letters 2014;21:1389-93.
39. Inano R, Oishi N, Kunieda T, Arakawa Y, Yamao Y, Shibata S, Kikuchi T, Fukuyama H, Miyamoto S. Voxel-based clustered imaging by multiparameter diffusion tensor images for glioma grading. Neuroimage Clin 2014;5:396-407.
40. Bishop CM. Pattern Recognition and Machine Learning. International Edition. Kolkata: Springer India, 2013.
41. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
42. Wen Q, Kelley DA, Banerjee S, Lupo JM, Chang SM, Xu D, Hess CP, Nelson SJ. Clinically feasible NODDI characterization of glioma using multiband EPI at 7 T. Neuroimage Clin 2015;9:291-9.
43. Lampinen B, Szczepankiewicz F, Mårtensson J, van Westen D, Sundgren PC, Nilsson M. Neurite density imaging versus imaging of microscopic anisotropy in diffusion MRI: A model comparison using spherical tensor encoding. Neuroimage 2017;147:517-31.
44. Wang Y, Wang Q, Haldar JP, Yeh FC, Xie M, Sun P, Tu TW, Trinkaus K, Klein RS, Cross AH, Song SK. Quantification of increased cellularity during inflammatory demyelination. Brain 2011;134:3590-601.

Cite this article as: Xu J, Ren Y, Zhao X, Wang X, Yu X, Yao Z, Zhou Y, Feng X, Zhou XJ, Wang H. 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. doi:10.21037/qims-22-145

Appendix 1

Radiomics feature extraction

In this study, 18 first order features, 22 gray-level co-occurrence matrix (GLCM) features, 14 neighboring gray-level dependence matrix (NGLDM), 16 gray-level run length matrix (GLRLM) features, 16 gray-level size zone matrix (GLSZM) features and 14 shape/size features were adopted as the radiomic features (45). A three-dimensional (3D) Coiflet wavelet transform was applied to the DWI images in order to extract the first order features in frequency decomposed images. The frequency components were HHH, HHL, HLH, HLL, LHH, LHL, LLH, and LLL, where "H" and "L" denote high-pass and low-pass filters, respectively (34). To characterize the textural changes on DWI images over different diffusion gradient (different b values), we measured 8 new sequential features from the 21 b values for each texture feature, including mean, max, min, median, variance, kurtosis, skewness, energy. Therefore, a total of 7076 features were extracted on primary dataset for each tumor with 21 b values and 6,100 features for external testing dataset for each tumor with 17 b values. It should be noticed that the number of radiomics features were different between two datasets. This was due to the different number of b values. However, this problem has been solved during the feature selection procedure by choosing the features from the DWI images of which b values were equal between two datasets.

Appendix 2

Feature Selection procedure

All work in this part was accomplished using an open ML library scikit-learn (ver. 0.22), in Python (32). The whole dataset was split into training set (80 percent) and testing set (20 percent). The external test set included 55 cases on five-fold cross-validation sets. Radiomics features with b values that were not included in the external test were excluded. A five-step rigorous selection process has been implemented both on combined DWI-model features and radiomics features:

Step I WMW U-test
All features of the training data were tested by a non-parametric WMW U-test with a significant setting of $P<0.05$.

Step II ML methods
On one way, a learning model-based single feature sequencing approach was involved. The idea of this approach was to use Logistic Regression (LR), Support-vector Machine (SVM), K-nearest neighbors (KNN), Random Forests (RF), Naïve Bayes (NB) and Stacking Methods (Stacks) separately as a learning estimator to build a predictive model for each individual feature filtered by step 1.
On the other way, the top features were selected according to scores derived from Lasso, RF and Recursive feature elimination (RFE) methods. Grid search was used on these estimators to define the hyper-parameters of Lasso and RF. Features ranking in Lasso were determined by the final coefficient, and in RF were sorted by their importance. Recursive feature elimination (RFE) model was also applied for selection. RFE creates a model from all features, and then eliminates the least important features in turn by measuring the contribution of each feature in a given model (26). In this study, RideCV was utilized as an underlying function to stabilize it.
In total, 9 ML based selection methods with 5-fold cross-validation performed were used, each providing the top 20 features of this work.

Step III Voting system
A voting system was proposed to find the common features selected by 9 methods mentioned above. We only reserved features with votes >9/2 and 19 features were left for multiparametric DWI based on ML and 14 features left for DWI radiomics.

https://dx.doi.org/10.21037/qims-22-145

Step IV Correlation test

The final decision was made by calculating the Pearson correlation coefficient r and eliminate features with r>0.7 according to the rank. After this, 10 and 9 features were selected for multiparametric DWI and radiomics methods, respectively.

Step V Combination and Grouping

Features were combined in accordance with DWI models, where in this study 10 features belong to 5 different DWI models. Based on combination mathematics, there are $2^5-1=31$ types of DWI model combinations. The description of these combinations is simply combining their capitals, such as SC for SEM-model & CTRW-model, SFs for SEM-model & FROC-model & SM model and the like.

In this study, we also aimed to compare with two traditional classification methods. One of traditional method is based on a single DWI model only measuring the average for each parameter shown in Eq. [1] to Eq. [7]. And the other method depends on DWI radiomics features. Therefore, we also trained and tested our estimators on these conventional combinations, which were also grouped following rules stipulated above. The selection procedure for radiomics features followed the same rules except the step 5, we only chose the top 1 to 9 features according to the rank in step 4 on the training set instead. (See *Table S5*, which shows feature combinations of two traditional methods).

References

45. Beylkin G, Coifman R, Rokhlin V. Fast wavelet transforms and numerical algorithms I. Communications on pure and applied mathematics 1991;44:141-83.

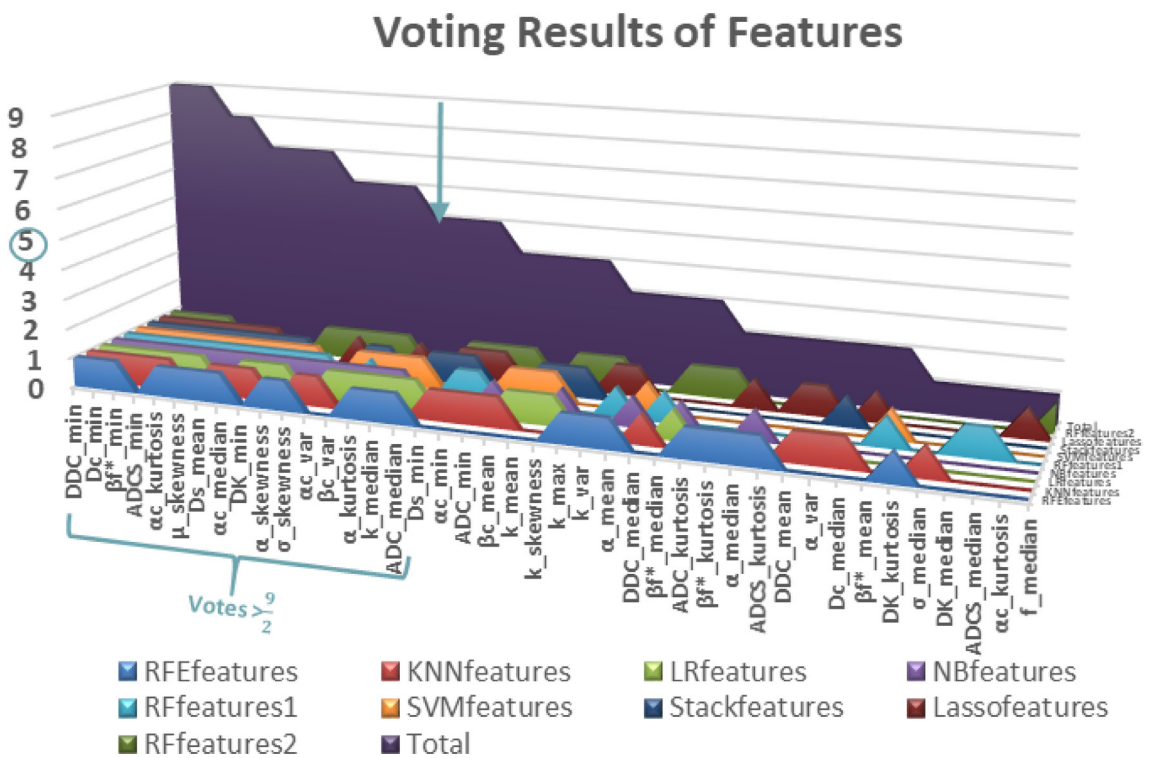

Figure S1 The votes of features selected by 9 ML selection methods (step 3).

https://dx.doi.org/10.21037/qims-22-145

Table S1 Correlation test results

	Pearson correlation matrix																		
	DDC_min	**Dc_min**	**βf*_min**	**αc_kurtosis**	**ADCS_min**	**μ_skewness**	**DK_min**	**α_skewness**	**σ_skewness**	**Ds_mean**	**αc_median**	**αc_var**	**k_median**	**Ds_min**	**αc_mean**	**αc_skewness**	**βc_var**	**α_kurtosis**	**ADC_median**
DDC_min	1	0.8810807	0.5478665	0.3947538	0.8414184	0.253429131	0.8414184	0.321240678	0.490974359	0.104029	0.4855554	0.6610031	0.587563	0.1599478	0.5824501	0.362759941	0.435051	0.182240221	0.536257366
Dc_min	0.8810807	1	0.6570947	0.3494493	0.9132883	0.297697103	0.9132883	0.343171208	0.474311275	0.3551752	0.38070135	0.5948554	0.5674203	0.175077	0.4618118	0.329001535	0.6096523	0.203332977	0.549041176
βf*_min	0.5478665	0.6570947	1	0.2486806	0.6338648	0.129205695	0.6338648	0.175528681	0.488938643	0.2721702	0.265729055	0.3497738	0.4577853	0.0395106	0.2915615	0.274598424	0.5820147	0.032292412	0.46968089
αc_kurtosis	0.3947538	0.3494493	0.2486806	1	0.3678287	0.207394817	0.3678287	0.551987294	0.474235638	0.0065027	0.794399189	0.67698	0.7874129	0.0754365	0.8276838	0.958688564	0.2806428	0.35463391	0.808941161
ADCS_min	0.8414184	0.9132883	0.6338648	0.3678287	1	0.226604814	1	0.292803324	0.455293802	0.3410058	0.375838049	0.6220602	0.5458609	0.1107667	0.463452	0.355409608	0.571527	0.152665357	0.535378527
μ_skewness	0.2534291	0.2976971	0.1292057	0.2073948	0.2266048	1	0.2266048	0.220910547	0.179502319	0.1274533	0.433178985	0.1532537	0.2613249	0.1605273	0.3676551	0.268078053	0.1973063	0.180574349	0.159125801
DK_min	0.8414184	0.9132883	0.6338648	0.3678287	1	0.226604814	1	0.292803324	0.455293802	0.3410058	0.375838049	0.6220602	0.5458609	0.1107667	0.463452	0.355409608	0.571527	0.152665357	0.535378527
α_skewness	0.3212407	0.3431712	0.1755287	0.5519873	0.2928033	0.220910547	0.2928033	1	0.252266711	0.0840269	0.489442806	0.4645366	0.5831461	0.0862979	0.520765	0.559983311	0.101712	0.855274669	0.584011623
σ_skewness	0.4909744	0.4743113	0.4889386	0.4742356	0.4552938	0.179502319	0.4552938	0.252266711	1	0.1555649	0.440414111	0.5304539	0.6002326	0.1253808	0.498797	0.489186612	0.3806164	0.061558816	0.622770263
Ds_mean	0.104029	0.3551752	0.2721702	0.0065027	0.3410058	0.127453316	0.3410058	0.084026859	0.155564925	1	0.126437327	0.2679803	0.1216034	0.0605586	0.0347546	0.00268296	0.7827822	0.061850281	0.101579967
αc_median	0.4855554	0.3807013	0.2657291	0.7943992	0.375838	0.433178985	0.375838	0.489442806	0.440414111	0.1264373	1	0.6523146	0.8254703	0.1773121	0.9639402	0.864013804	0.1804959	0.237974246	0.777779164
αc_var	0.6610031	0.5948554	0.3497738	0.67698	0.6220602	0.153253722	0.6220602	0.464536572	0.530453916	0.2679803	0.652314574	1	0.785901	0.0844488	0.8195331	0.65314769	0.5056715	0.313395889	0.741350071
k_median	0.587563	0.5674203	0.4577853	0.7874129	0.5458609	0.261324905	0.5458609	0.583146072	0.600232625	0.1216034	0.8254703	0.785901	1	0.0079255	0.8647078	0.822287263	0.5385137	0.314511353	0.963861831
Ds_min	0.1599478	0.175077	0.0395106	0.0754365	0.1107667	0.160527294	0.1107667	0.086297872	0.125380784	0.0605586	0.177312116	0.0844488	0.0079255	1	0.1193631	0.100624986	0.0931557	0.021338513	0.016813582
αc_mean	0.5824501	0.4618118	0.2915615	0.8276838	0.463452	0.367655063	0.463452	0.520765006	0.498797003	0.0347546	0.963940177	0.8195331	0.8647078	0.1193631	1	0.860038643	0.2709265	0.289400549	0.815240214
αc_skewness	0.3627599	0.3290015	0.2745984	0.9586886	0.3554096	0.268078053	0.3554096	0.559983311	0.489186612	0.002683	0.864013804	0.6531477	0.8222873	0.100625	0.8600386	1	0.2656724	0.327043339	0.82568054
βc_var	0.435051	0.6096523	0.5820147	0.2806428	0.571527	0.197306291	0.571527	0.101711981	0.380616355	0.6827822	0.180495858	0.5056715	0.5385137	0.0931557	0.2709265	0.265672407	1	0.045724646	0.518963128
α_kurtosis	0.1822402	0.203333	0.0322924	0.3546339	0.1526654	0.180574349	0.1526654	0.855274669	0.061558816	0.0618503	0.237974246	0.3133959	0.3145114	0.0213385	0.2894005	0.327043339	0.0457246	1	0.31758627
ADC_median	0.5362574	0.5490412	0.4696809	0.8089412	0.5353785	0.159125801	0.5353785	0.584011623	0.622770263	0.10158	0.777779164	0.7413501	0.9638618	0.0168136	0.8152402	0.82568054	0.5189631	0.31758627	1

Pink highlight data are highly colinear features (r>0.7) and blue highlight features are finally selected.

Table S2 Combinations of the subgroups of selected 10 features in multiparametric DWI model

Features selected by ML	DDC_min	β_f^*_min	α_c_kurtosis	μ_skewness	Ds_mean	α_skewness	σ_skewness	α_c_var	β_c_var	Ds_min
	S(SEM)	**C(CTRW)**	**F(FROC)**	**S(SM)**	**I(IVIM)**					
Subgroups (DWImodels)	DDC_min α_skewness	α_c_kurtosis α_c_var β_c_var	β_f^*_min μ_skewness	σ_skewness	Ds_mean Ds_min					
Combinations of subgroups	s Ss SFs SFsI SCFsI	S Fs SsI SCFs	F Is FsI SCsI	I SF SCs CFsI	C SI CFs CFsI	FI CsI SCFI	Cs SFI	SC SCF	CF SCI	CI CFI

https://dx.doi.org/10.21037/qims-22-145

Table S3 Integrated training and internal test results for each feature combination and they were sorted according to their AUCs on internal test set

Feature-combination	Feature numbers	Prediction methods	trainCV acc	trainCV auc	test acc	test auc	tpr	tnr	Cut-off
ADC	1	LRfold25	0.7939	0.783	0.6	0.6429	0.4286	1	0.872
DDC_min'	1	'model_KNNs'	0.7	0.7994	0.7333	0.7411	0.5714	0.875	0.5455
Ds min'	1	model KNNs'	0.8333	0.9195	0.6667	0.6875	0.8571	0.5	0.6667
βf* min'	1	model_SVMs'	0.7667	0.7638	0.6	0.5982	0.5714	0.625	—
βc var'	1	'model_Stacks'	0.6	0.727	0.6	0.625	0.5714	0.75	0.5601
α skewness'	1	'model RFs'	0.6333	0.8741	0.6	0.5	0.4286	0.75	0.5847
αc var'	1	'model KNNs'	0.6333	0.823	0.5333	0.5982	0.4286	0.875	0.7143
s	1	'model_RFs'	0.6667	0.9782	0.4667	0.6161	0.7143	0.5	0.4431
μ skewness'	1	'model RFs'	0.7	1	0.4667	0.5357	0.5714	0.75	0.74
αc kurtosis'	1	'model LRCVs'	0.6667	0.7425	0.4667	0.5179	0.8571	0.375	0.4511
Ds_aver'	1	'model_RFs'	0.8	0.9253	0.4667	0.5893	0.5714	0.75	0.721
F	2	LRfold5'	0.8382	0.8184	0.6667	0.7679	1	0.5	0.442
SM	2	'LRfold12'	0.8106	0.7921	0.5333	0.6071	0.4286	1	0.8665
S	2	'LRfold24'	0.807	0.8748	0.6667	0.6786	0.8571	0.625	0.4924
SEM	2	'RFfold20'	0.7909	0.7857	0.7333	0.7143	0.7143	0.75	0.5583
DKI	2	'Stackfold12'	0.8376	0.8341	0.6667	0.6071	0.7143	0.75	0.3458
I	2	'KNNfold8'	0.6845	0.731	0.5333	0.5536	0.2857	1	1
Ss	3	'KNNfold24'	0.8376	0.8743	0.6	0.6696	0.4286	0.875	0.7143
FROC	3	'LRfold2'	0.8112	0.8079	0.6667	0.6429	0.4286	0.875	0.5855
IVIM	3	'Stackfold18'	0.7764	0.735	0.4667	0.5357	0.4286	0.875	0.5655
CTRW	3	'Stackfold15'	0.8042	0.7838	0.5333	0.5357	0.8571	0.375	0.5162
sI	3	'RFfold3'	0.7273	0.8385	0.4	0.5268	0.8571	0.375	0.506
Fs	3	'KNNfold3'	0.787	0.8519	0.6667	0.5179	0.5714	0.75	0.5556
C	3	NIfold5'	0.8182	0.7864	0.5333	0.5179	0.5714	0.625	0.4581
SI	4	'RFfold24'	0.8182	0.8635	0.8667	0.8214	0.7143	0.7143	0.7983
SF	4	'KNNfold24'	0.8312	0.9078	0.7333	0.6518	0.7143	0.7143	0.5294
FI	4	'Stackfold5'	0.7273	0.835	0.5333	0.5714	0.4286	0.4286	0.7126
Cs	4	'KNNfold9'	0.7761	0.8419	0.5333	0.5268	1	1	0.1667
SC	5	'Stackfold1'	0.8921	0.8863	0.7333	0.7857	0.7143	0.875	0.54
CF	5	'KNNfold24'	0.757	0.8509	0.5333	0.6696	1	0.375	0.25
SFs	5	'KNNfold3'	0.8585	0.9007	0.6667	0.625	0.5714	0.75	0.6429
FsI	5	'RFfold24'	0.9091	0.8587	0.5333	0.625	0.8571	0.5	0.4442
CI	5	'RFfold19'	0.6364	0.7798	0.6	0.6161	0.7143	0.625	0.3205
SsI	5	'RFfold12'	0.8182	0.8668	0.6	0.6161	0.7143	0.625	0.4288
SCs	6	'Stackfold4'	0.8955	0.8939	0.6667	0.7679	0.7143	0.875	0.5243
SFI	6	'KNNfold24'	0.8273	0.9013	0.6667	0.6518	0.7143	0.625	0.6
CFs	6	'RFfold3'	0.8182	0.858	0.7333	0.5893	0.5714	0.875	0.5739
CsI	6	'Stackfold4'	0.7273	0.8391	0.5333	0.5714	0.7143	0.625	0.4629
SCF	7	'Stackfold4'	0.9091	0.8898	0.8	0.8393	0.8571	0.75	0.5051
CFI	7	'RFfold13'	0.8182	0.8638	0.6667	0.6964	0.5714	0.875	0.5547
SCI	7	'LRfold24'	0.7976	0.8578	0.6667	0.6786	0.7143	0.75	0.491
SFsI	7	'KNNfold3'	0.8718	0.9037	0.6667	0.5982	0.7143	0.625	0.4667
SCFs	8	'Stackfold7'	0.9091	0.9021	0.7333	0.7679	0.5714	0.875	0.5308
CFsI	8	'RFfold3'	0.9091	0.8731	0.6	0.6429	0.7143	0.625	0.3624
SCsI	8	'Stackfold1'	0.9091	0.8815	0.6	0.6429	0.5714	0.75	0.5474
SCFI	9	'LRfold24'	0.8139	0.8727	0.6667	0.6607	0.8571	0.625	0.1751
All features	10	'RFfold24'	0.8755	0.8873	0.6667	0.7321	0.7143	0.75	0.5698

"tpr" refers to the sensitivity and "tnr" refers to the specificity. The highlighted row represents the best prediction model which achieves the highest AUC in internal test set.

https://dx.doi.org/10.21037/qims-22-145

Table S4 R-squared of each DWI model to assess the goodness of fitting

DWI model	Mean
ADC	0.9449
IVIM	0.9569
SEM	0.9959
SM	0.9699
DKI	0.9699
FROC	0.9788
CTRW	0.9801

Table S5 Row 'Feature combinations' gives the selected features based on single DWI model and radiomics respectively

DWI models	Traditional methods based on the single DWI model		Traditional DWI radiomics		Feature combinations	Serial numbers of selected features
	Abbreviation	Features	Feature types	Features/serial number		
ADC	Amean	*ADC*_mean	b =3,500	HHL_Maximum/ ② ; HHL_kurtosis/ ⑧	TOP1	①
IVIM	Imean	f_mean, D_S_mean, D_f_mean	b =4,000	HHH_Kurtosis/ ⑥	TOP2	①②
SEM	Smean	*DDC*_mean, α_mean	b =0	HHL_Kurtosis/ ④	TOP3	①②③
SM	smean	ADC_S_mean, σ_mean	Sequential	Kurtosis_original_shape_Minor_Axis_Length/ ① ; Skewness_original_shape_Minor_AxisLength/ ③	TOP4	①②③④
DKI	Dmean	D_K_mean, K_mean		Kurtosis_HHH_InterquartileRange/ ⑤ ; Skewness_origianl_glszm_LargeAraHighGrayLevelEmphasis/ ⑦	TOP5	①②③④⑤
					TOP6	①②③④⑤⑥
FROC	Fmean	D_f_mean, β^*_c_mean, μ_mean		Skewness_original_shape_Maximum2DDiameterColumn/ ⑨	TOP7	①②③④⑤⑥⑦
CTRW	Cmean	D_C_mean, α_c_mean, β_c_aver			TOP8	①②③④⑤⑥⑦⑧
					TOP9	①②③④⑤⑥⑦⑧⑨

The 'serial number ① - ⑨ ' was determined by the ranking after step 4.

https://dx.doi.org/10.21037/qims-22-145

Table S6 Internal test results of two traditional DWI methods

Combination name	Feature-num	Prediction estimator	trainCV_acc	trainCV_auc	test_acc	test_auc	best_tpr	best_tnr	Cut-off
Single DWI model									
ADC	1	LRfold25	0.7939	0.783	0.6	0.6429	0.4286	1	0.872
SM	2	LRfold12	0.8106	0.7921	0.5333	0.6071	0.4286	1	0.8665
SEM	2	RFfold20	0.7909	0.7857	0.7333	**0.7143**	0.7143	0.75	0.5583
DKI	2	Stackfold12	0.8376	0.8341	0.6667	0.6071	0.7143	0.75	0.3458
CTRW	3	Stackfold15	0.8042	0.7838	0.5333	0.5357	0.8571	0.375	0.5162
FROC	3	LRfold2	0.8112	0.8079	0.6667	0.6429	0.4286	0.875	0.5855
IVIM	3	Stackfold18	0.7764	0.735	0.4667	0.5357	0.4286	0.875	0.5655
Radiomics									
TOP6	6	Stackfold3	0.9558	0.9821	0.8	**0.8393**	0.8571	0.875	0.6111
TOP4	4	LRfold3	0.9421	0.944	0.6667	0.75	1	0.625	0.3141
TOP7	7	Stackfold3	0.9624	0.9551	0.7333	0.75	0.8571	0.625	0.429
TOP5	5	LRfold4	0.9321	0.9399	0.7333	0.75	0.8571	0.75	0.3827
TOP8	8	Stackfold2	0.9761	0.9821	0.8	0.75	0.5714	1	0.595
TOP9	9	Stackfold1	0.9794	0.9868	0.7333	0.6964	0.5714	0.875	0.6148
TOP3	3	LRfold1	0.9082	0.9044	0.6667	0.6607	1	0.375	0.3405
TOP2	2	LRfold10	0.8821	0.8801	0.5333	0.5179	0.4286	0.875	0.7174

https://dx.doi.org/10.21037/qims-22-145