scientific reports

OPEN

# Machine learning for grading prediction and survival analysis in high grade glioma

Xiangzhi Li[1,2,3,6], Xueqi Huang[2,6], Yi Shen[3,6], Sihui Yu[2], Lin Zheng[1], Yunxiang Cai[1], Yang Yang[4], Renyuan Zhang[5], Lingying Zhu[1]✉ & Enyu Wang[1]✉

**We developed and validated a magnetic resonance imaging (MRI)-based radiomics model for the classification of high-grade glioma (HGG) and determined the optimal machine learning (ML) approach. This retrospective analysis included 184 patients (59 grade III lesions and 125 grade IV lesions). Radiomics features were extracted from MRI with T1-weighted imaging (T1WI). The least absolute shrinkage and selection operator (LASSO) feature selection method and seven classification methods including logistic regression, XGBoost, Decision Tree, Random Forest (RF), Adaboost, Gradient Boosting Decision Tree, and Stacking fusion model were used to differentiate HGG. Performance was compared on AUC, sensitivity, accuracy, precision and specificity. In the non-fusion models, the best performance was achieved by using the XGBoost classifier, and using SMOTE to deal with the data imbalance to improve the performance of all the classifiers. The Stacking fusion model performed the best, with an AUC = 0.95 (sensitivity of 0.84; accuracy of 0.85; F1 score of 0.85). MRI-based quantitative radiomics features have good performance in identifying the classification of HGG. The XGBoost method outperforms the classifiers in the non-fusion model and the Stacking fusion model outperforms the non-fusion model.**

**Keywords** Magnetic resonance imaging, Cerebral glioblastoma, Radiomics analysis, Machine learning

Glioblastoma is the most common primary malignant brain tumor, accounting for approximately 57% of all gliomas and 48% of all primary malignant central nervous system (CNS) tumors[1]. Despite recent advances in the multimodal treatment of glioblastoma, the overall prognosis remains poor, with minimal long-term survival rates[2]. Approximately 5 out of every 100 glioblastoma patients survive five years after diagnosis, and more than 300,000 people die from glioma disease each year worldwide[3,4]. WHO Classification of Tumors of the Central Nervous System classified gliomas into grades I-IV, with grades I and II being low-grade gliomas and grades III and IV being high-grade gliomas (HGG)[5]. The prognosis for HGG is even worse, with the majority of HGG patients surviving less than two years, and the median survival for grade IV is only about 14.6 months[6]. ASCO guidelines recommend that all patients with advanced tumors receive palliative care within 8 weeks of diagnosis, but the distinction between patients with intermediate and advanced gliomas remains a challenge[7,8].

The exact grading of glioma still needs to be confirmed by histopathology, but preoperative MRI features can significantly improve the grading accuracy[9]. Although biopsy or surgical excision is the primary method for obtaining pathological specimens, sampling error due to tumor heterogeneity and patient physical limitations remains a clinical challenge[10]. MRI, with its high resolution, non-invasive and quantifiable characteristics, has become a core imaging tool to assist grading and prognostic evaluation[11–13].

Machine learning (ML) is used in various medical fields due to its ability to develop robust risk models and improve predictive power[14,15]. Recent studies have shown that ML algorithms have a strong ability to predict prognostic outcomes of patients using glioblastoma imaging and pathological features[16]. Currently, the most accurate prediction method is to use radiomics and deep learning to build models from manually or automatically extracted image features[17–19]. However, the classification performance of different machine

[1]Taizhou Key Laboratory of Minimally Invasive Interventional Therapy & Artificial Intelligence, Taizhou Branch of Zhejiang Cancer Hospital (Taizhou Cancer Hospital), No.50, Zhenxin Road, Taizhou 317502, China. [2]School of Science & School of Medicine, Guangxi University of Science and Technology, Liuzhou 545006, China. [3]School of Public Health, Youjiang Medical University For Nationalities, Baise 533000, China. [4]Department of Radiation Oncology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China. [5]Department of International Education College, Hainan Medical University, Haikou 571199, China. [6]These authors contributed equally: Xiangzhi Li, Xueqi Huang and Yi Shen. ✉email: whitemouse811@hotmail.com; fishman811@sohu.com

learning algorithms for HGG grading is unclear, and the value of dealing with data imbalances has not been fully determined.This study aims to develop and validate an HGG classification model based on MRI radiomics, focusing on comparing the performance of various machine learning classifiers, including stack-fusion models. In addition, assess the improvements SMOTE has brought in addressing data imbalances. We also conducted survival analysis of HGG based on ML model to explore the survival differences of HGG and the factors affecting survival.

## Materials and methods

### Patients

To conduct this prediction study, we used a dataset from Taizhou Cancer Hospital and the Second Affiliated Hospital of Zhejiang University School of Medicine in Zhejiang, China. A total of 238 patients diagnosed with glioblastoma underwent pathological examination in two hospitals from March 2013 to June 2018 were selected. The exclusion criteria were as follows: (a) patients with missing original MRI (incomplete T1WI dataset) in the format of Digital Imaging and Communications in Medicine (DICOM) incomplete ($n=52$), and (b) unsatisfactory imaging quality ($n=2$). All enrolled cases were classified using the 5th edition of the WHO Classification Criteria for Tumors of the Central Nervous System (CNS5). Figure 1 shows the workflow of this study. Finally, a total of 184 patients (106 males and 78 females; age range, 11–80 years; mean age, 51.1 years) were included in the development of the ML prediction model. Pathological assessment of grade III lesions was performed in 59 cases and grade IV lesions in 125 cases. In the survival analysis of GBM patients, a total of 144 patients were finally included due to missing data in 40 follow-up cases. All procedures involving human participants were performed in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants and/or their legal guardians included in the study. This study complied with the MI-CLAIM reporting guidelines.

### Image data acquisition

Patients were examined by Intera Achieva 1.5T or GE Signa Excite HD 1.5T magnetic resonance equipment. The MRI sequence consisted of axial gradient-echo T1WI imaging, and DWI with free-breathing. Parallel imaging was used, and fat suppression was applied using spectral pre-saturation inversion recovery (SPIR).The T1WI acquisition parameters were: acquisition matrix 256 × 256, slice thickness 6 mm, slice spacing 1 mm, number of excitations (NEX) 1.0, repetition time (TR) 450 ms, echo time (TE) 21 ms, pixel width 0.86 × 0.86 mm, and field of view (FOV) 22.0 × 22.0 cm[20].

### Lesion segmentation

Before segmentation, the low-frequency intensity inhomogeneity in MRI image data was eliminated, and the N4 bias field of MRI image lines was corrected using SimpleITK (version 2.1.1). For T1-weighted MRI, the region of interest (ROI) was obtained by manually circling the tumor along the border, slice by slice. The delineation process included the cystic and hemorrhagic portion of the ROI while avoiding the vascular component and adjacent normal tissues. The ROIs were manually segmented on T1-weighted images by a radiologist with 10 years of experience in brain MRI. These ROIs were then reviewed by a radiologist with more than 20 years of experience. If one more experienced radiologist questioned the ROIs, they would be re-segmented after agreement between the two. The ROIs were manually drawn on axial slices using ITK-SNAP (version 3.6.0), covering the entire lesion. In this way, we ensured that segmentation for subsequent analyses was as accurate as possible. Figure 2 presents an example of a case with grade III and a case with grade IV glioma.

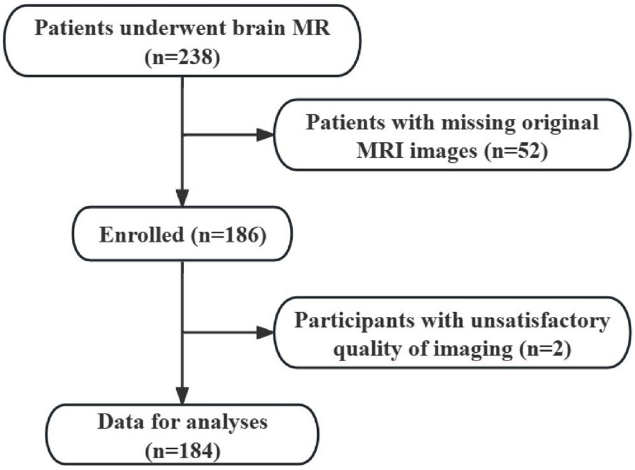
Patients underwent brain MR (n=238)
Patients with missing original MRI images (n=52)
Enrolled (n=186)
Participants with unsatisfactory quality of imaging (n=2)
Data for analyses (n=184)

**Fig. 1**. Patient exclusion criteria flowchart.

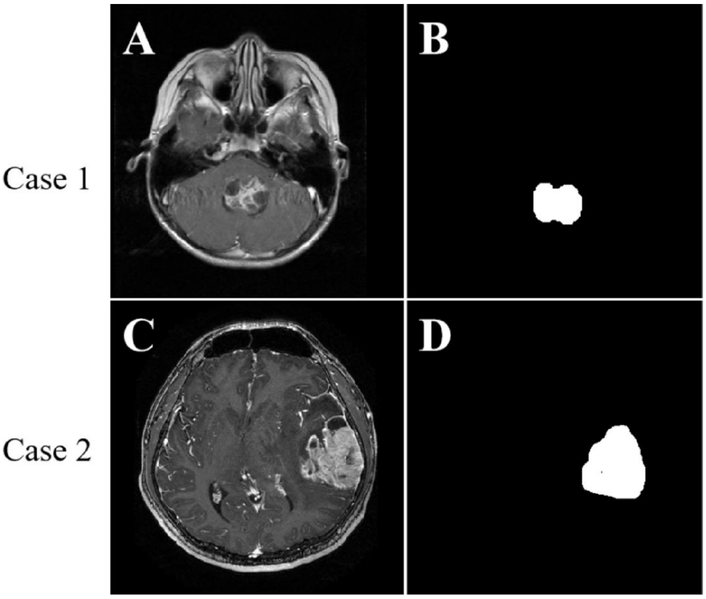

**Fig. 2**. This study uses two subtypes of high-grade glioma as examples. Case 1 (first column): An 11-year-old male with grade III glioma. Case 2 (second column): A 19-year-old female with grade IV glioma.

## Feature extraction

Radiomics features were defined according to the Thermal Radiomics Python package[21]. The T1WI sequence was used to calculate all the features on the optical radiomics list, and the T1WI sequence yielded a total of 107 image features. The features included in the thermal radiomics list include 18 First order features, 14 shape features (3D) and 75 texture features. The texture eigenomics list includes 24 Gy Level Co-occurrence Matrix (GLCM) Features, 16 Gy Level Size Zone Matrix (GLSZM) Features, 16 Gy Level Run Length Matrix (GLRLM) Features, 5 Neighbouring Gray Tone Difference Matrix (NGTDM) Features, and 14 Gy Level Dependence Matrix (GLDM) Features. We extracted the grouping for each of the 2 diagnostic imaging physicians using the within-group correlation coefficient (ICC). The features using intraclass correlation coefficient (ICC) to assess intra- and inter-observer agreement. Typically, an ICC of less than 0.75 was considered to be below reliability[22]. We used the features with good repeatability (ICC > 0.90) for subsequent steps such as feature dimensionality reduction and feature selection.

## Feature selection

Too many features may lead to problems such as model overfitting. Feature selection methods can reduce the dimension of the feature space, i.e., to obtain a "low number but high quality" of features with a low probability of classification error. The Least absolute shrinkage and selection operator (LASSO) is a commonly used feature selection method[23]. We used lasso algorithms for feature selection and screened for features with good repeatability (ICC > 0.90) using consistency tests. These feature selection algorithms were implemented using the Python scikit learning environment (version 0.19.1). LASSO, which minimises the cost function was used and features with non-zero coefficients were identified with a specific penalty coefficient alpha[24].

## Data pre-processing

In the dataset used in this study, patients in class III accounted for 32.1% of all HGG patients, leading to a degradation of classifier performance. To address this category imbalance, we used SMOTE (sampling technique). Synthetic samples were generated by linear combination[25]. We divided the processed data into 70% training set and 30% test set. In the dataset used in this study, count data were presented as values and proportions and analysed using chi-square tests. The training set was used for model development and the test set was used for estimating the generalisability of the model.

## Classifier and model validation

In order to achieve efficient and stable performance for classification, we implemented six ML classifiers using the Python scikit-learning environment (version 0.19.1), namely Logistic Regression, Extreme Gradient Boosting (XGBoost), Decision Tree, Random Forest (RF), Adaboost, and the Gradient Boosting Decision Tree (GBDT). The reason why these six classifiers were chosen and compared is that these classifiers are commonly used to classify related studies glioblastoma, bladder, skin lesions, breast, kidney, colon in previous studies[26–31]. Our study used five cross-validated model validations with baseline features for each cohort as shown in Table 1. The average performance of the model was used to evaluate the classification performance of the model. The shapley additive explanations (SHAP) explainer was constructed to rank the features by using Python package shap v0.39, which was used to calculate feature Contribution[32].

| Variants | Class III<br>*N* = 59 | Class IV<br>*N* = 125 | *P* value |
|---|---|---|---|
| Minimum | 194.44 ± 359.05 | 103.54 ± 186.59 | 0.02 |
| First order Skewness | 0.17 ± 0.70 | 0.48 ± 0.54 | 0.02 |
| Idn | 0.93 ± 0.03 | 0.92 ± 0.02 | < 0.01 |
| Imc1 | -0.16 ± 0.05 | -0.13 ± 0.05 | < 0.01 |
| Inverse Variance | 0.30 ± 0.13 | 0.25 ± 0.10 | < 0.01 |
| Joint Energy | 0.02 ± 0.03 | 0.01 ± 0.01 | < 0.01 |
| MCC | 0.59 ± 0.10 | 0.56 ± 0.12 | 0.06 |
| Large Area Low Gray Level Emphasis | 556.03 ± 1159.00 | 194.46 ± 559.96 | < 0.01 |
| Size Zone Nonuniformity | 2159.71 ± 3811.42 | 3020.99 ± 4891.65 | 0.24 |
| Age | 44.85 ± 13.76 | 54.17 ± 14.57 | < 0.01 |
| Sex | | | 0.88 |
| Male | 33(55.93%) | 73(58.40%) | |
| Female | 26(44.07%) | 52(41.60%) | |

**Table 1**. Characteristics of participants by glioblastoma grading. *Idn* Inverse difference normalization, *Imc1* Relevance Information Metric 1, *MCC* maximum correlation coefficient.

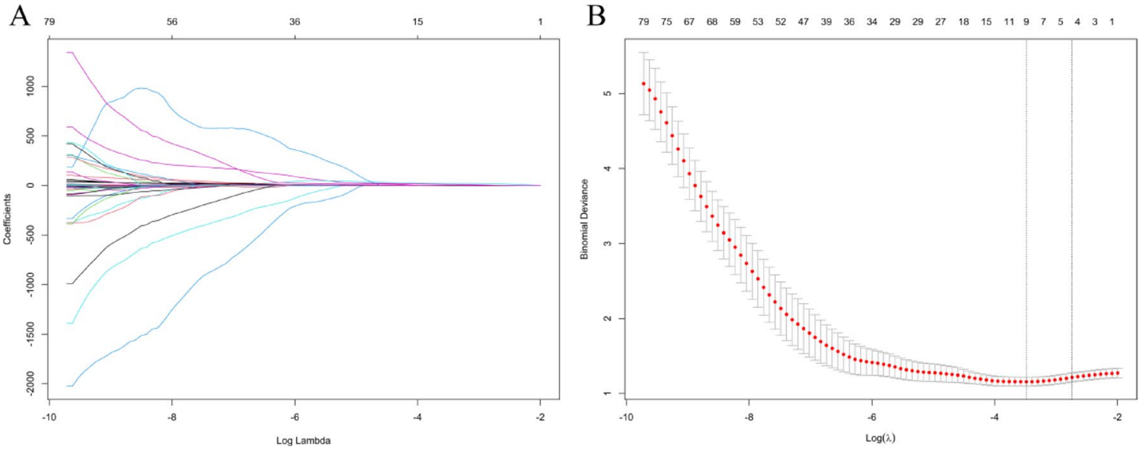

**Fig. 3**. Variable selection by the LASSO regression model. (**A**) Choice of the optimal parameter (λ) in the LASSO regression model with logλ as the horizontal coordinate and regression coefficients as the vertical coordinate; (**B**) Plot of λ vs. number of variables with logλ as the bottom horizontal coordinate, binomial deviance as the vertical coordinate, and number of variables as the top horizontal coordinate.

## Statistical analysis

The area under the ROC curve (AUC), sensitivity, accuracy, precision and specificity were calculated for the classification performance of HGG using the Python scikit-learn environment (version 0.19.1).The ROC curve was based on the mean value calculated from all cross-validation sets for generalisation ability. Confidence intervals (CIs) for the AUCs were obtained using the Python scikit-learning environment (version 0.19.1) for 1000 bootstrap replications. The trained models were interpreted by SHAP. Survival analyses were performed using cox proportional risk, Kaplan-Meier (KM) test. $P < 0.05$ was considered statistically significant difference.

## Results

### Clinical characteristics of the patients

In this study, a prediction model was developed by using MRI of 184 HGG patients with 59 grade III lesions and 125 grade IV lesions with mean ages of 44.8 and 54.2 years, respectively. The results showed that there were statistically significant differences between grade III lesions and grade IV lesions in minimum, fisrt order skewness, inverse difference normalisation, correlation information metric (IMC1), inverse variance, joint energy, large area low gray level emphasis and age (Table 1). Because of the imbalance in the number of cases of grade III lesions and grade IV lesions, we used an oversampling process with synthetic minorities.

### Performance of feature selection methods and classifiers

In our ML model, we tested six classifiers with the LASSO feature selection method, and the LASSO coefficient path diagram is shown in Fig. 3A. As the strength of regularization gradually increases, many feature coefficients

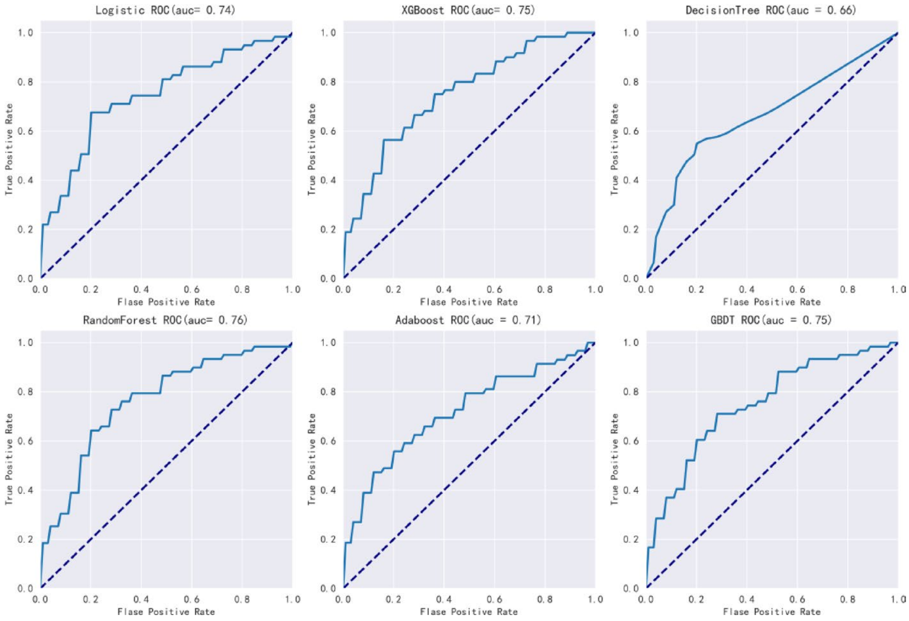

**Fig. 4**. Six model ROC curves for data imbalance not handled with smote.

| Feature class | AUC (95% CI) | ACC | Recall rate | F1 score | Specificity |
|---|---|---|---|---|---|
| Logistic | 0.74(0.65–0.83) | 0.73 | 0.44 | 0.51 | 0.86 |
| XGBoost | 0.75(0.67–0.83) | 0.74 | 0.49 | 0.54 | 0.86 |
| Decision tree | 0.66(0.59–0.75) | 0.72 | 0.51 | 0.53 | 0.82 |
| Random forest | 0.76(0.69–0.83) | 0.72 | 0.32 | 0.42 | 0.90 |
| Adaboost | 0.71(0.64–0.79) | 0.72 | 0.44 | 0.50 | 0.85 |
| GBDT | 0.75(0.68–0.81) | 0.73 | 0.49 | 0.53 | 0.84 |

**Table 2**. Performance metrics for six algorithms that do not use Smote to handle data imbalance. *AUC* Area under the ROC curve, *ACC* Accuracy.

are compressed to zero, allowing most features to be eliminated and the model to be gradually simplified. The cross-validation error plot in the regularization path is shown in Fig. 3B.

## Model evaluation and comparison

When SMOTE was not used to handle data imbalance, XGBoost performed better than logistic regression, decision tree, random forest, Adaboost, and GBDT classifiers with ACC of 0.74, 0.73, 0.72, 0.72, 0.72, 0.73 respectively (Fig. 4). Table 2 shows the performance metrics for the six algorithms at each step, including AUC (95% CI), ACC, F1 Score, Sensitivity, and Specificity. When using smote to make the data balanced, XGBoost performs better than logistic regression, decision tree, random forest, Adaboost, and GBDT classifiers, with ACC of 0.78, 0.73, 0.74, 0.76, 0.76, 0.76, respectively (Fig. 5). After SMOTE processing, the balanced data resulted in the improved ACC values for all six models. The Stacking fusion model was constructed based on SMOTE to deal with the data imbalance problem and using the six ML models, GBDT, logistic regression, XGBoost, decision tree, random forest and Adaboost, as the primary learners (Fig. 6). The model uses the primary learners to make predictions on the data and then uses these predictions as inputs to the secondary learners (logistic regression) for final prediction or classification. Ultimately, the performance metrics of the seven models including the Stacking fusion model are shown in Table 3, with the Stacking fusion model having the best AUC value of 0.95. DCA decision curve is a tool to evaluate the value of the model forest farm (Figure S1A and Figure S1C), which is the decision curve without and after oversampling respectively, in which XGBoost algorithm has the highest net benefit. After Stacking is integrated, the net income of models is significantly increased (Figure S1E). As can be seen from the Calibration Curve (Figure S1B and Figure S1D) and the Precision-Recall curve (Figure S6 and Figure S7), SMOTE has improved the sample equilibrium of the model after SMOTE treatment, thus achieving better performance in predictive calibration. In addition, the calibration curve (Figure S1F) and the Precision-Recall curve (Figure S8) of the Stacking fusion model also performs well, further verifying its advantages in improving model robustness and prediction accuracy.

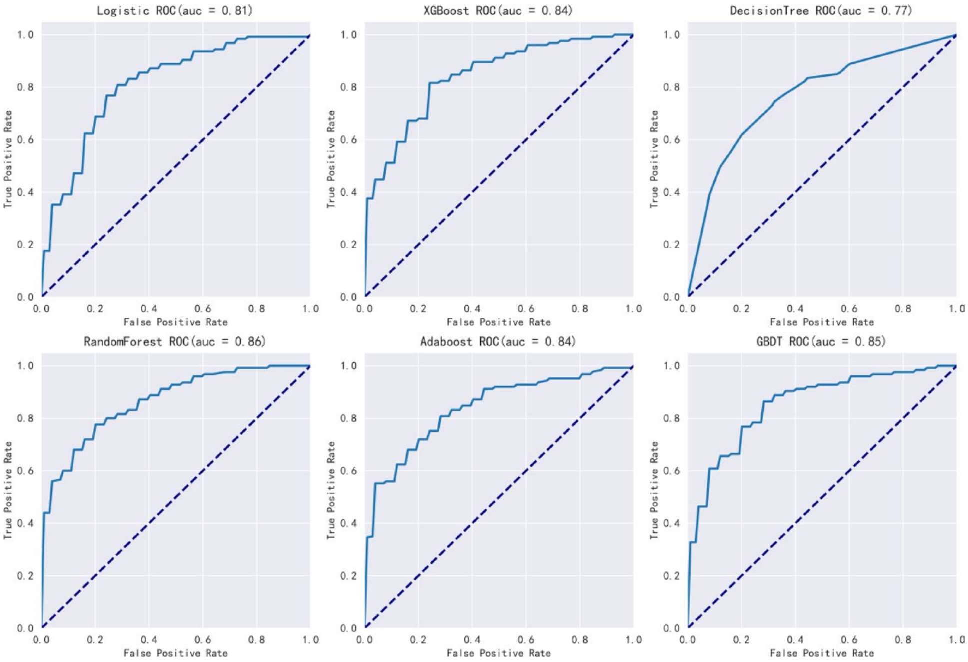

**Fig. 5**. Six model ROC curves after using smote to deal with data imbalances.

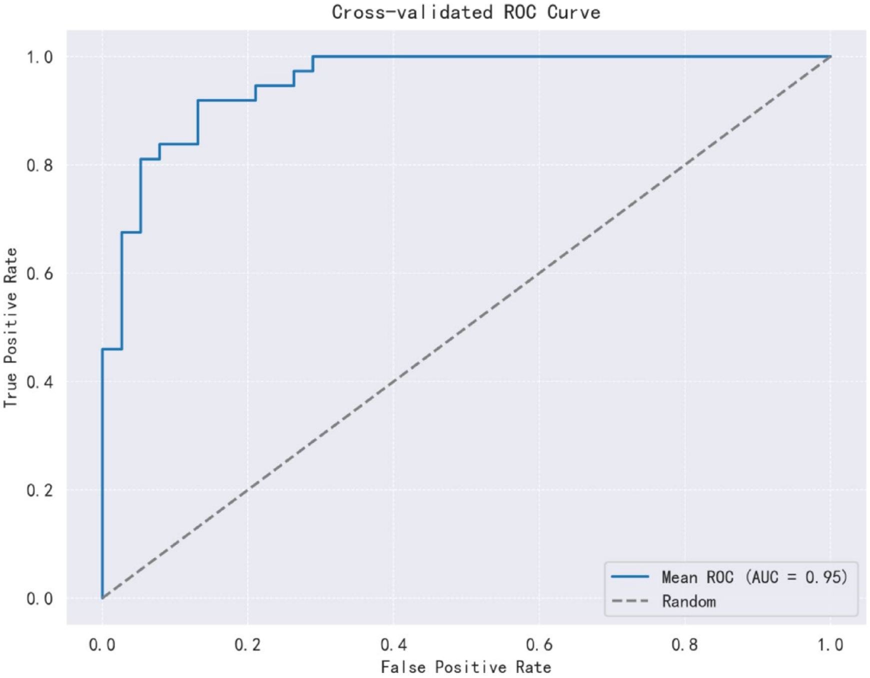

**Fig. 6**. ROC curve for the stacking fusion model with an AUC of 0.95.

## Model interpretation

To better understand the relationship between the model and the data, we use SHAP to provide a more intuitive explanation of how these variables affect model predictions. However, SHAP is not interpretable for the fusion model, and is therefore interpreted for the best performing XGBOOST model. Figure S2A shows the important features in the model, where the features are ranked on the y-axis, indicating their importance to the prediction model. The results show a high correlation between "SizeZoneNonUniformity", "Idn", "Skewness", "Minimum", "InverseVariance", "JointEnergy", "Imc1" and "LargeAreaLowGrayLevelEmphasis" and the prediction of grade IV glioma. Figure S2B is used to explain the decision-making process of the machine learning model and help analyze the influence of various features on the model's prediction results. On the analysis of the results,

| Feature class | AUC | ACC | Recall rate | F1 score | Specificity | *P* |
|---|---|---|---|---|---|---|
| Logistic | 0.81(0.77–0.84) | 0.73 | 0.72 | 0.72 | 0.74 | < 0.001 |
| XGBoost | 0.84(0.78–0.88) | 0.78 | 0.79 | 0.77 | 0.77 | < 0.001 |
| Decision tree | 0.77(0.66,0.84) | 0.74 | 0.74 | 0.74 | 0.74 | < 0.001 |
| Random forest | 0.86(0.82,0.92) | 0.76 | 0.73 | 0.75 | 0.80 | < 0.001 |
| Adaboost | 0.84(0.80,0.87) | 0.76 | 0.74 | 0.75 | 0.77 | < 0.001 |
| GBDT | 0.85(0.80,0.90) | 0.76 | 0.76 | 0.76 | 0.76 | < 0.001 |
| Stacking | 0.95(0.92–0.99) | 0.85 | 0.84 | 0.85 | 0.87 | |

**Table 3**. Performance metrics of six algorithms after using Smote to handle data imbalance. AUC Area under the ROC curve, ACC Accuracy, P values are used for stacking delong tests with other models.

"SizeZoneNonUniformity", "Skewness" ,"Minimum", "LargeAreaLowGrayLevelEmphasis" and "InverseVariance" this a few indicators of red points are mainly concentrated on the left side, When these eigenvalues are small, the prediction results of the model are more in favor of grade 4 glioma. The red spots of the remaining indicators are mainly concentrated on the right side, indicating that when these characteristic values are large, it is more likely to predict grade IV brain glioma. This phenomenon suggests that these features play an important role in the model to distinguish between different grades of brain gliomas. Figure S2C is a Decision plot showing the contribution degree of each feature to the model output. The distribution of SHAP values for most indicators is concentrated in the range of 0.2–0.8, indicating that these features have a high importance in the model prediction process and play a key role in the final decision. Figure S2D is a Partial Dependence Plot (PDP). It is used to show the average influence of a certain feature on the model prediction results under different values. PDP helps to understand the relationship between the feature and the target variable by simulating the predicted trend when a single feature changes, thus revealing the decision logic of the model.

## Survival analysis

We analyzed the survival of 144 patients with high-grade glioma HGG, of whom 43 (29.9%) had censored data due to lost follow-up or non-event endpoints, and 101 (70.1%) reached the end point of death. We did survival analysis using the KM test and Cox proportional risk model, where HGG staging was included in the Cox model in addition to the nine MRI features screened. KM test revealed that gender and MGMT protein had no significant effect on survival. However, HGG stage, IDH1 classification, KPS score, Ki-67, and surgical mode (total vs. subtotal) had significant effects on survival (Figure S3). Specifically, the median survival for Grade III patients was 667 days, compared to 531 days for Grade IV patients.

The Cox proportional risk model was used to investigate the significant contribution of MRI covariates in predicting OS. Both univariate and multivariate analyses were performed. Table 4 summarises all results, including effect sizes expressed as hazard ratios (HR). In univariate Cox regression, each covariate was assessed independently and found to be statistically significant for IMC1, MCC, IDH1, Ki-67, surgical approach (total resection and subtotal resection), tumor grading, KPS score.

In the multivariate Cox regression, we did find significant variables for IDH1, Ki-67, tumor grading, KPS score. In the multivariate Cox analysis of stepwise regression, we finally filtered out six statistically significant covariates. Increased IMC1 was associated with shortened OS (HR = 1.37). To verify the robustness of the model, we performed a subgroup analysis (Figure S4). In certain clinical subgroups of patients with Ki-67 > 20, KPS score > 70, grade IV gliomas, and wild-type IDH1, Imc1 is a imaging feature of great value in predicting the prognosis of patients.

To verify the robustness of the model under different sample distributions and clinical contexts, we constructed a balanced dataset using undersampling under the results and conducted a survival analysis. (Figures S5-S8, Table 5). The results showed that the radiological feature Imc1 always had significant and stable prognostic value, providing strong support for personalized survival prediction and treatment.

## Discussion

In this study, we developed and validated an MRI-based radiomics approach for classifying HGG and interpreted the results of the model using SHAP values. We identified influential features, such as "SizeZoneNonUniformity", "Skewness", "Idn", "Minimum", "Imc1", "JointEnergy", and "InverseVariance". In addition, we demonstrated how each feature impacts the model's prediction of glioma grade. The best performance in the non-fusion model was achieved by using the XGBoost classifier, and using the SMOTE to deal with data imbalance improves the performance of all classifiers. This is consistent with previous studies that LASSO and XGBoost classifiers outperform other classifiers for classification of benign and malignant ovarian cysts and classification of skin lesions[28,33]. The Stacking fusion model performed best with AUC = 0.95 (sensitivity of 0.84; accuracy of 0.85; F1 score of 0.95) by interpreting the SHAP values, we found that size region non-uniformity, skewness, inverse difference normalisation, minimum eigenvalue, and correlation information metrics were highly correlated with the model predictions. In the survival analysis, we found that gender had no significant effect on survival, but there were significant differences in survival across GBM staging. In addition, by Cox proportional risk model, we determined the contribution of MRI covariates to predict OS and found that IMC1 were associated with survival.

| Feature | HR (univariable) | HR (multivariable) | HR (final) |
|---|---|---|---|
| Minimum | 0.80(0.60–1.08, $P$ =0.142) | | |
| Skewness | 1.21(0.99–1.47, $P=0.066$) | | |
| Idn | 0.83(0.69-1.00, $P=0.054$) | | |
| Imc1 | 1.49(1.16–1.91, $P=0.002$) | 1.33(0.98–1.79, $P=0.066$) | 1.37(1.07–1.74, $P=0.012$) |
| Inverse Variance | 0.94(0.77–1.14, $P=0.519$) | | |
| MCC | 0.77(0.63–0.95, $P=0.001$) | 0.96(0.74–1.25, $P=0.761$) | |
| Large Area Low Gray Level Emphasis | 0.67(0.39–1.15, $P=0.144$) | | |
| Size Zone Nonuniformity | 0.97(0.80–1.17, $P=0.728$) | | |
| Joint Energy | 0.91(0.73–1.13, $P=0.389$) | | |
| IDH1 | | | |
| Wild | | | |
| Mutant | 0.39(0.23–0.67, $P<0.001$) | 0.55(0.30-1.00, $P=0.050$) | 0.54(0.30–0.97, $P=0.039$) |
| MGMT | | | |
| Negative | | | |
| Positive | 1.56(0.96–2.54, $P=0.071$) | | |
| Ki_67 | | | |
| ≤ 20 | | | |
| > 20 | 1.76(1.15–2.68, $P=0.009$) | 1.56(1.00-2.42, $P=0.050$) | 1.54(0.99-2.40, $P=0.053$) |
| Sex | | | |
| Female | | | |
| Male | 1.21(0.82–1.79, $P=0.341$) | | |
| Operate | | | |
| Part | | | |
| All | 0.66(0.45–0.98, $P=0.038$) | 0.70(0.46–1.06, $P=0.088$) | 0.69(0.46–1.03, $P=0.068$) |
| Grade | | | |
| III | | | |
| IV | 2.73(1.59–4.68, $P<0.001$) | 1.93(1.07–3.48, $P=0.029$) | 1.91(1.06–3.45, $P=0.030$) |
| KPS | | | |
| ≤ 70 | | | |
| > 70 | 0.41(0.26–0.64, $P<0.001$) | 0.38(0.24–0.60, $P<0.001$) | 0.38(0.24–0.61, $P<0.001$) |

**Table 4**. Risk values for different factors in univariate and multivariate analyses. *Idn* Inverse difference normalization, *Imc1* Relevance Information Metric 1, *MCC* maximum correlation coefficient, *KPS* Karnofsky Performance Status.

Most ML prediction models use public MRI databases due to the scarcity of glioblastoma data and the very short survival period[34]. However, some studies have found that public MRI storage inventories have significant shortcomings, including a lack of expert tumour segmentation, which can lead to a reduction in predictive reliability[35]. Therefore, one of the strengths of this study is the use of a larger sample size of MRI data from glioblastoma patients in both centers and the application of ML algorithms to construct predictive models with higher model reliability. In the process of constructing the graded prediction model, we encountered several important issues. First, unbalanced data can significantly affect model performance in the biomedical field[36,37]. For example, if 30% of the patients with grade III glioblastoma are in the brain, even if the model predicts all outcomes as normal, the accuracy of predicting grade IV is still 0.7, which is clearly incorrect. This imbalance also led to a tendency for the model to predict grade IV more frequently, with lower accuracy in predicting grade III samples. However, the above study did not take this fully into account during the model construction process. In ML, it is recommended that methods such as oversampling or undersampling be used to address data imbalance[38]. Therefore, we use SMOTE sampling technique to balance the sample sizes of class III and class IV, which improves the prediction accuracy and stability of the model. Finally, model performance evaluation is also a challenge. AUC is the most widely used metric. In this study, the AUC of the six ML algorithms ranged from 0.89 to 0.99, and few studies have been conducted to predict GMB patient levels. Overfitting is another issue to be considered in model evaluation[39]. During the modelling process, even with methods such as cross-validation, some models still showed very high AUC values and high accuracy even on the test data. However, statistical tests showed differences in AUC between the two datasets, suggesting that the generalisation ability of models trained only with internal validation is questionable. Therefore, in the absence of fully independent external validation data, we recommend splitting part of the data and processing it before normalisation and variable selection. In conclusion, despite the challenges we face in constructing predictive models for disability risk, ML algorithms have the potential to address these issues. By addressing data imbalance, efficiently selecting relevant variables, improving model accuracy, and controlling overfitting, a predictive model for classifying HGG patients with high predictive and generalisation abilities was built.

| Feature | HR (univariable) | HR (multivariable) | HR (final) |
|---|---|---|---|
| Minimum | 0.70 (0.43–1.13, $P=0.147$) | | |
| Skewness | 1.31 (0.92–1.87, $P=0.141$) | | |
| Idn | 0.85 (0.63–1.16, $P=0.306$) | | |
| Imc1 | 2.01 (1.30–3.10, $P=0.002$) | 1.60 (0.95–2.70, $P=0.077$) | 1.74 (1.16–2.60, $P=0.007$) |
| Inverse Variance | 0.97 (0.71–1.31, $P=0.826$) | | |
| MCC | 0.62 (0.44–0.89, $P=0.009$) | 0.90 (0.60–1.35, $P=0.620$) | |
| Large Area Low Gray Level Emphasis | 0.81 (0.44–1.48, $P=0.489$) | | |
| Size Zone Nonuniformity | 0.82 (0.49–1.37, $P=0.447$) | | |
| Joint Energy | 0.90 (0.62–1.30, $P=0.568$) | | |
| IDH1 | | | |
| Wild | | | |
| Mutant | 0.37 (0.16–0.82, $P=0.015$) | 1.03 (0.43–2.51, $P=0.942$) | |
| MGMT | | | |
| Negative | | | |
| Positive | 1.42 (0.70–2.89, $P=0.334$) | | |
| Ki_67 | | | |
| <=20 | | | |
| > 20 | 1.86 (1.00-3.46, $P=0.049$) | 1.99 (0.97–4.08, $P=0.060$) | 1.93 (0.95–3.93, $P=0.069$) |
| Sex | | | |
| Female | | | |
| Male | 1.44 (0.77–2.71, $P=0.252$) | | |
| Operate | | | |
| Part | | | |
| All | 0.62 (0.34–1.12, $P=0.115$) | | |
| Grade | | | |
| III | | | |
| IV | 4.49 (1.76–11.45, $P=0.002$) | 3.37 (1.23–9.22, $P=0.018$) | 3.38 (1.29–8.83, $P=0.013$) |
| KPS | | | |
| <=70 | | | |
| > 70 | 0.22 (0.10–0.47, $P<0.001$) | 0.16 (0.07–0.39, $P<0.001$) | 0.16 (0.07–0.39, $P<0.001$) |

**Table 5**. The risk values of different factors in the univariate and multivariate analyses after downsampling. *Idn* Inverse difference normalization, *Imc1* Relevance Information Metric 1, *MCC* maximum correlation coefficient, *KPS* Karnofsky Performance Status.

Our imaging model contributes to the development of networks that can be used to aid decision-making in hospitals. From a model interpretation perspective, traditional ML algorithms are often criticised for their lack of transparency and interpretability[40,41]. In order to better understand the internal logic and decision rules behind the model predictions, another strength of this study is the use of SHAP values to interpret these ML models. XGBoost is the best performing non-fusion model in this study, and it is also the explanatory model we focus on. In these test data, we calculated the SHAP values for each feature variable to assess their contribution to the prediction results. The overall SHAP summary plot helps us to understand which features positively and negatively influence the predicted results, while the importance feature plot provides an average assessment of the importance of the features across the entire dataset. SHAP helps us to understand the contribution of each feature to the predicted results of this overall picture, and provides useful information for further analysis and interpretation of the model. In addition, the SHAP dependency graph helps to observe how these features affect the output of the prediction model at different levels. In this study, the influence of various imaging features on the grade of brain glioma can be clearly seen. Overall, the SHAP values used in this study provide a way to reduce the difficulty in interpreting ML models and increase their interpretability and transparency. This allows us to better understand the prediction results of the XGBoost model for grading HGG patients. By analysing the SHAP values, we can quantitatively assess the extent to which these features influence the prediction results, identify potential risk groups of HGG patients based on MRI features, and provide a basis for interventions. This helps physicians to make early clinical decisions, alleviate the suffering of HGG patients, and reduce the burden on the healthcare system.

A limitation of our study is the lack of external validation and protocol analysis of other centers. In addition, we used only one commonly used feature selection method and the performance of six classification methods for MRI-based identification of HGG patient classification. Finally, our results are based on basic sequences that are easy to be applied clinically and only a single T1 sequence was used; other MRI sequences were not taken into account. Another study also found that the prediction and classification of brain glioma lesions found that multiple sequences were superior to single sequences in terms of features[42]. This was not done and this was due to the absence of other sequences in some of the patients. Nonetheless, future studies could try a multi-feature selection approach and data from multiple sequences to build predictive models.

In this two-center study, we developed a model to predict survival up to 8 months after radiotherapy. The model was designed to be used in cohorts of patients receiving optimal treatment as well as in cohorts of patients receiving modified treatment. The neural network with T1 showed generalisable classification in both retrospective and external, prospective test cohorts. If validated in a large prospective study, these methods could be used to differentiate between patients with an initial response to radiotherapy and those who require closer image monitoring and second-line treatment (or termination of ineffective treatment).

In conclusion, this paper proposes a multi-parameter MRI-based grading prediction model for HGG patients. The experimental results show that the radiomics analysis based on multiparameter MRI is effective for grading HGG patients. XGBoost is the optimal non-fusion ML method, and Stacking fusion model has the best performance. After using SMOTE to make the data balanced can improve the performance of all models. If validated in a large prospective study, this method could be used to differentiate disease stages in HGG patients.

## Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 20 February 2025; Accepted: 6 May 2025
Published online: 15 May 2025

## References

1. Ostrom, Q. T. et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the united States in 2011–2015. *Neuro-Oncology* **20** (suppl_4), iv1–86. https://doi.org/10.1093/neuonc/noy131 (2018).
2. Taphoorn, M. J., Sizoo, E. M. & Bottomley, A. Review on quality of life issues in patients with primary brain tumors. *Oncologist* **15** (6), 618–626. https://doi.org/10.1634/theoncologist.2009-0291 (2010).
3. Huang, P., Li, L., Qiao, J., Li, X. & Zhang, P. Radiotherapy for glioblastoma in the elderly: A protocol for systematic review and meta-analysis. *Medicine* **99** (52), e23890. https://doi.org/10.1097/MD.0000000000023890 (2020).
4. Zhao, X. & Zhao, X. M. Deep learning of brain magnetic resonance images: A brief review. *Methods* **192**, 131–140. https://doi.org/10.1016/j.ymeth.2020.09.007 (2021).
5. Louis, D. N. et al. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* **131** (6), 803–820. https://doi.org/10.1007/s00401-016-1545-1 (2016).
6. Bao, Z. et al. Intratumor heterogeneity, microenvironment, and mechanisms of drug resistance in glioma recurrence and evolution. *Front. Med-Prc.* **15** (4), 551–561. https://doi.org/10.1007/s11684-020-0760-2 (2021).
7. Ferrell, B. R. et al. Integration of palliative care into standard oncology care: American society of clinical oncology clinical practice guideline update. *J. Clin. Oncol.* **35** (1), 96–112. https://doi.org/10.1200/JCO.2016.70.1474 (2017).
8. Fink, L. et al. Palliative care for in-patient malignant glioma patients in Germany. *J. Neuro-Oncol.* **167** (2), 323–338. https://doi.org/10.1007/s11060-024-04611-8 (2024).
9. Cao, H. et al. A quantitative model based on clinically relevant MRI features differentiates lower grade gliomas and glioblastoma. *Eur. Radiol.* **30** (6), 3073–3082. https://doi.org/10.1007/s00330-019-06632-8 (2020).
10. Buda, M., AlBadawy, E. A., Saha, A. & Mazurowski, M. A. Deep radiogenomics of Lower-Grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images. *Radiol. Artif. Intell.* **2** (1), e180050. https://doi.org/10.1148/ryai.2019180050 (2020).
11. Gutman, D. A. et al. MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* **267** (2), 560–569. https://doi.org/10.1148/radiol.13120118 (2013).
12. Matsui, Y. et al. Prediction of lower-grade glioma molecular subtypes using deep learning. *J. Neuro-Oncol.* **146** (2), 321–327. https://doi.org/10.1007/s11060-019-03376-9 (2020).
13. Liu, H., Shen, L., Huang, X. & Zhang, G. Maximal tumor diameter in the preoperative tumor magnetic resonance imaging (MRI) T2 image is associated with prognosis of grade II glioma. *Medicine* **100** (10), e24850. https://doi.org/10.1097/MD.0000000000024850 (2021).
14. Zhou, H. et al. Machine learning for the prediction of all-cause mortality in patients with sepsis-associated acute kidney injury during hospitalization. *Front. Immunol.* **14**, 1140755. https://doi.org/10.3389/fimmu.2023.1140755 (2023).
15. Zhu, Y. et al. Deep learning radiomics of dual-modality ultrasound images for hierarchical diagnosis of unexplained cervical lymphadenopathy. *Bmc Med.* **20** (1), 269. https://doi.org/10.1186/s12916-022-02469-z (2022).
16. Zhang, D. et al. Comparison of MRI radiomics-based machine learning survival models in predicting prognosis of glioblastoma multiforme. *Front. Med-Lausanne*. **10**, 1271687. https://doi.org/10.3389/fmed.2023.1271687 (2023).
17. Choi, Y. S. et al. Machine learning and radiomic phenotyping of lower grade gliomas: improving survival prediction. *Eur. Radiol.* **30** (7), 3834–3842. https://doi.org/10.1007/s00330-020-06737-5 (2020).
18. Sun, L., Zhang, S., Chen, H. & Luo, L. Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning. *Front. Neurosci-Switz*. **13**, 810. https://doi.org/10.3389/fnins.2019.00810 (2019).
19. Ou, J. et al. CT radiomics features to predict lymph node metastasis in advanced esophageal squamous cell carcinoma and to discriminate between regional and non-regional lymph node metastasis: a case control study. *Quant. Imag. Med. Surg.* **11** (2), 628–640. https://doi.org/10.21037/qims-20-241 (2021).
20. Yu, X. et al. Pediatric diffuse intrinsic Pontine glioma radiotherapy response prediction: MRI morphology and T2 intensity-based quantitative analyses. *Eur. Radiol.* https://doi.org/10.1007/s00330-024-10855-9 (2024).
21. van Griethuysen, J. et al. Computational radiomics system to Decode the radiographic phenotype. *Cancer Res.* **77** (21), e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339 (2017).
22. He, Z. et al. Machine learning-based radiomics for histological classification of Parotid tumors using morphological MRI: a comparative study. *Eur. Radiol.* **32** (12), 8099–8110. https://doi.org/10.1007/s00330-022-08943-9 (2022).
23. Wang, H. et al. Radiomics analysis of multiparametric MRI for the preoperative evaluation of pathological grade in bladder cancer tumors. *Eur. Radiol.* **29** (11), 6182–6190. https://doi.org/10.1007/s00330-019-06222-8 (2019).
24. Sauerbrei, W., Royston, P. & Binder, H. Selection of important variables and determination of functional form for continuous predictors in multivariable model Building. *Stat. Med.* **26** (30), 5512–5528. https://doi.org/10.1002/sim.3148 (2007).
25. Zheng, Y., Zhang, C. & Liu, Y. Risk prediction models of depression in older adults with chronic diseases. *J. Affect. Disorders*. **359**, 182–188. https://doi.org/10.1016/j.jad.2024.05.078 (2024).

26. Thenier-Villa, J. L., Martinez-Ricarte, F. R., Figueroa-Vezirian, M. & Arikan-Abello, F. Glioblastoma pseudoprogression discrimination using multiparametric magnetic resonance imaging, principal component analysis, and supervised and unsupervised machine learning. *World Neurosurg.* **183**, e953–e962. https://doi.org/10.1016/j.wneu.2024.01.074 (2024).
27. Xu, S. & Huang, J. Machine learning algorithms predicting bladder cancer associated with diabetes and hypertension: NHANES 2009 to 2018. *Medicine* **103** (4), e36587. https://doi.org/10.1097/MD.0000000000036587 (2024).
28. Moldovanu, S., Miron, M., Rusu, C. G., Biswas, K. C. & Moraru, L. Refining skin lesions classification performance using geometric features of superpixels. *Sci. Rep-Uk.* **13** (1), 11463. https://doi.org/10.1038/s41598-023-38706-5 (2023).
29. Mo, W., Ding, Y., Zhao, S., Zou, D. & Ding, X. Identification of a 6-gene signature for the survival prediction of breast cancer patients based on integrated multi-omics data analysis. *Plos One.* **15** (11), e241924. https://doi.org/10.1371/journal.pone.0241924 (2020).
30. Lu, X. et al. Application of interpretable machine learning algorithms to predict acute kidney injury in patients with cerebral infarction in ICU. *J. Stroke Cerebrovasc.* **33** (7), 107729. https://doi.org/10.1016/j.jstrokecerebrovasdis.2024.107729 (2024).
31. Wang, X. et al. Development and validation of artificial intelligence models for preoperative prediction of inferior mesenteric artery lymph nodes metastasis in left colon and rectal cancer. *Ejso-Eur J. Surg. Onc.* **48** (12), 2475–2486. https://doi.org/10.1016/j.ejso.2022.06.009 (2022).
32. Gao, T. et al. Machine learning-based prediction of in-hospital mortality for critically ill patients with sepsis-associated acute kidney injury. *Ren. Fail.* **46** (1), 2316267. https://doi.org/10.1080/0886022X.2024.2316267 (2024).
33. Seo, M., Choi, M., Lee, Y. J., Jung, S. E. & Rha, S. E. Evaluating the added benefit of CT texture analysis on conventional CT analysis to differentiate benign ovarian cysts. *Diagn. Interv Radiol.* **27** (4), 460–468. https://doi.org/10.5152/dir.2021.20225 (2021).
34. Madan, R. et al. Prospective phase II study of radiotherapy dose escalation in grade 4 glioma using (68)Ga-Pentixafor PET scan. Clin Oncol-Uk. (2024). https://doi.org/10.1016/j.clon.2024.04.011
35. Cepeda, S. et al. The Rio hortega university hospital glioblastoma dataset: A comprehensive collection of preoperative, early postoperative and recurrence MRI scans (RHUH-GBM). *Data Brief.* **50**, 109617. https://doi.org/10.1016/j.dib.2023.109617 (2023).
36. Wang, L. et al. Classifying 2-year recurrence in patients with Dlbcl using clinical variables with imbalanced data and machine learning methods. *Comput. Meth Prog Bio.* **196**, 105567. https://doi.org/10.1016/j.cmpb.2020.105567 (2020).
37. Noguer, J., Contreras, I., Mujahid, O., Beneyto, A. & Vehi, J. Generation of Individualized Synthetic Data for Augmentation of the Type 1 Diabetes Data Sets Using Deep Learning Models. *Sensors-Basel* ;**22**(13). https://doi.org/10.3390/s22134944. (2022).
38. Shatnawi, M., Zaki, N. & Yoo, P. D. Protein inter-domain linker prediction using random forest and amino acid physiochemical properties. *BMC Bioinform.* **15** (Suppl 16), S8. https://doi.org/10.1186/1471-2105-15-S16-S8 (2014).
39. Kernbach, J. M. & Staartjes, V. E. Foundations of machine Learning-Based clinical prediction modeling: part II-Generalization and overfitting. *Acta Neurochir. Suppl.* **134**, 15–21. https://doi.org/10.1007/978-3-030-85292-4_3 (2022).
40. Ching, T. et al. Opportunities and Obstacles for deep learning in biology and medicine. *J. R Soc. Interface.* **15** (141). https://doi.org/10.1098/rsif.2017.0387 (2018).
41. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* **19** (6), 1236–1246. https://doi.org/10.1093/bib/bbx044 (2018).
42. Herr, J., Stoyanova, R. & Mellon, E. A. Convolutional neural networks for glioma segmentation and prognosis: A systematic review. *Crit. Rev. Oncog.* **29** (3), 33–65. https://doi.org/10.1615/CritRevOncog.2023050852 (2024).

## Author contributions

The study was conceived and designed by XL, XH, and YS. Material preparation was performed by SY, LZ, and YC; data collection and analysis were performed by XL, XH, YY, and RZ. The first draft of the manuscript was written by XL, XH, and YS, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

## Funding

This work was supported by the Science and Technology Program of Traditional Chinese Medicine of Zhejiang Province(2024ZL1261), Zhejiang Science and Technology Plan for Disease Prevention and Control (2025JK137), the Science and Technology Plan Projects of Liuzhou (2022CAC0229) and the Zhejiang Provincial Science and Technology Program for Medicine and Health (No. 2025KY1898), the Taizhou Social Development Science and Technology Program(No. 23yvb138 and No. 23yvb140), the Wenling Social Development Science and Technology Program (No.2021S00042, No.2021S00088 and No.2023S00011).

## Declarations

## Competing interests

The authors declare no competing interests.

## Ethics approval consent to participate

This study was performed in line with the principles of the Declaration of Helsinki. Approved was granted by Ethics Committee of Taizhou Cancer Hospital.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-01413-4.

**Correspondence** and requests for materials should be addressed to L.Z. or E.W.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.