Hindawi
Scientific Programming
Volume 2021, Article ID 9913466, 11 pages
https://doi.org/10.1155/2021/9913466

Research Article

Tumor Grade and Overall Survival Prediction of Gliomas Using Radiomics

Jianming Ye,[1] **He Huang**,[2] **Weiwei Jiang**,[2] **Xiaomei Xu**,[2] **Chun Xie**,[1] **Bo Lu**,[3] **Xiangcai Wang**,[1] **and Xiaobo Lai**[2]

[1]*The First Affiliated Hospital, Gannan Medical University, Ganzhou 341000, China*
[2]*School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou 310053, China*
[3]*Faculty of Engineering, Shanghai Normal University Tianhua College, Shanghai 201815, China*

Correspondence should be addressed to Bo Lu; lb2364@sthu.edu.cn, Xiangcai Wang; wangxiangcai@csco.ac.cn, and Xiaobo Lai; dmia_lab@zcmu.edu.cn

Received 19 March 2021; Revised 3 April 2021; Accepted 10 April 2021; Published 23 April 2021

Academic Editor: Chenxi Huang

Glioma is one of the most common and deadly malignant brain tumors originating from glial cells. For personalized treatment, an accurate preoperative prognosis for glioma patients is highly desired. Recently, various machine learning-based approaches have been developed to predict the prognosis based on preoperative magnetic resonance imaging (MRI) radiomics, which extract quantitative features from radiographic images. However, major challenges remain for methodologic developments to optimize feature extraction and provide rapid information flow in clinical settings. This study investigates two machine learning-based prognosis prediction tasks using radiomic features extracted from preoperative multimodal MRI brain data: (i) prediction of tumor grade (higher-grade vs. lower-grade gliomas) from preoperative MRI scans and (ii) prediction of patient overall survival (OS) in higher-grade gliomas (<12 months vs. > 12 months) from preoperative MRI scans. Specifically, these two tasks utilize the conventional machine learning-based models built with various classifiers. Moreover, feature selection methods are applied to increase model performance and decrease computational costs. In the experiments, models are evaluated in terms of their predictive performance and stability using a bootstrap approach. Experimental results show that classifier choice and feature selection technique plays a significant role in model performance and stability for both tasks; a variability analysis indicates that classification method choice is the most dominant source of performance variation for both tasks.

1. Introduction

Glioma is one of the most common and deadly malignant brain tumors originating from glial cells. About 50 percent of nervous system tumors and 80 percent of all malignant brain tumors are gliomas. Glioblastoma multiforme (GBM) (also called glioblastoma) is a fast-growing glioma that develops from star-shaped glial cells (astrocytes and oligodendrocytes) that support the health of the nerve cells within the brain. In adults, GBM occurs most often in the cerebral hemispheres, especially in the brain's frontal and temporal lobes of the brain. GBM is a devastating brain cancer that typically results in death in the first 15 months after diagnosis. Traditional treatment of GBM is surgical resection followed by radiation therapy and/or chemotherapy. However, the median survival time of GBM is still less than 15 months despite surgical resection, radiotherapy, and chemotherapy. Therefore, the accurate preoperative prognosis of GBM patients is desired, which can provide essential information for planning the optimized and personalized treatment.

Recently, various machine learning-based approaches have been developed to predict the prognosis based on preoperative magnetic resonance imaging (MRI) radiomics, which is a new cross-field of medical informatics, aiming to extract quantitative features defined by mathematics from medical images, such as shape, intensity, and texture [1, 2]. Particularly, they applied the regression model to predict OS time in days or categorized it into short or long term based on binary classification using radiomic

features extracted from various types of preoperative image data [3]. According to the strategy through which features are extracted, these studies can be roughly divided into two categories: (1) methods based on manual features and (2) methods based on automatically extracted features using machine learning techniques. The basic idea of the method based on manual features is to extract the artificial designed features by semiautomatic or full-automatic method and use the traditional machine learning method to regress or classify the calculated features [4, 5]. For example, in [6], brain tumors' image phenotypic features are calculated from a public preoperative multimodal MRI brain dataset and input into the random forest classifier to learn a regression model for OS prediction. In [7], some manually labeled features are extracted from the BraTS 2017 dataset, such as the volume and surface irregularity of brain tumors, are used to train the artificial neural networks for OS prediction. Although manual feature-based methods have shown promising results, there is no systematic way to determine OS-related manual features but mostly depended on experience. Therefore, the machine learning-based methods have been proposed, which can automatically learn OS-related, deeply embedded MRI image features to better predict OS without prior knowledge [4]. For example, in [8], Nie et al. present a two-stage deep learning-based OS time prediction method of high-grade gliomas patient, where a 3D multichannel convolutional neural network (CNN) is proposed to extract implicit and high-level features automatically for OS prediction from multimodal preoperative MRI brain tumor data, including the contrast-enhanced T1 (T1c), diffusion tensor imaging (DTI), and resting-state functional MRI (rs-fMRI). In [9], a novel three-dimensional detailed delineation algorithm is introduced for GBM tumors in MRI, which efficiently delineates the whole tumor, enhancing core, edema, and necrosis volumes using fuzzy connectivity and multi-thresholding, followed by survival prediction of patients using the concept of habitats.

Although all the studies mentioned above have indicated an essential value of brain imaging phenotype for OS prediction, tumors are often heterogeneous in space and time. There are differences in the cell, gene, and microenvironment for different tumor regions at the same time point or at other time points in the same tumor region, which usually requires multiple biopsies to capture the tumor's molecular heterogeneity, bringing inconvenience and risk to patients. Radiomics can provide a noninvasive way to explore the heterogeneity of tumors [10]. Gliomas are the most common primary malignant brain tumors with high intrinsic heterogeneity. This heterogeneity is evident in radiomic features and morphology, making classification and prognosis more difficult [11]. Radiomics analysis of gliomas can provide additional information about the patient's classification, prognosis, and possible survival outcomes [12, 13].

However, although researchers at home and abroad have done a lot of research on the application of machine learning algorithm in radiomic feature classification and prognosis prediction [14–19], due to the lack of a unified standard, there are still many unknowns about which is the best model in the field. Many studies also use proprietary or in-house software in their radiomic feature extraction/analysis pipeline, severely limiting the community from making advances. Coupled with the fact that patients' medical images are protected by the confidentiality laws, it is incredibly challenging, if not impossible, to reproduce the results. Therefore, it is crucial to utilize publically available datasets and open-source tools to expand the radiomics field.

In our study, two machine learning classification tasks using radiomic features are investigated, which predict tumor grade and patient OS from preoperative MRI scans, respectively. These two tasks utilize the conventional machine learning techniques constructed with various classifier methods. Feature reduction methods also are applied to increase model performance and decrease computational costs. Models are assessed in terms of their predictive performance and stability using a bootstrap approach based on the 2017 BraTS Challenge's MRI data. Experimental results show that the classifier choice and dimensionality reduction technique plays a significant role in model performance and stability for both tasks. Figure 1 shows an outline of the radiomic workflow for the grade classification task, and we utilized a similar scheme for the overall survival classification task.

2. Material and Method

2.1. Dataset and Preprocessing. We utilized the 2017 BraTS Challenge's Training Dataset [20], which comprises 210 higher-grade gliomas (HGG) and 75 lower-grade gliomas (LGG) preoperative multimodal MRI scans collected from multiple centers. Each patient's multimodal scans include T1, postcontrast T1-weighted (T1c), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). All the MRI scans have been segmented manually by one to four raters, following the same annotation protocol, and experienced neuroradiologists approved their annotations. Annotations comprise the GD-enhancing tumor (ET-label 4), the peritumoral edema (ED-label 2), and the necrotic and non-enhancing tumor (NCR/NET-label 1). Each sequence was skull-stripped and was resampled to $1\,\text{mm} * 1\,\text{mm} * 1\,\text{mm}$ (isotropic resolution). For the overall survival challenge, age and prognosis of the patient posttreatment were supplied by the organizers. The overall survival data also were available for a subset of the GBM scans. In this study, all samples used for the grade prediction task are referred to as the Tumor Grade Dataset; all samples used for the overall survival prediction classification task are referred to as the Overall Survival Dataset. For the Tumor Grade Dataset, glioblastoma multiforme (GBM) was considered the negative type ($n = 210$) while LGG was considered the positive type ($n = 75$). The Overall Survival Dataset was stratified into binary classes based on median survival rates for GBM; patients who died before 12 months from diagnosis were considered negative ($n = 81$), while patients who died after 12 months were considered positive ($n = 82$). Examples of the four modalities and the corresponding tumor masks from two GBM patients are shown in Figure 2.

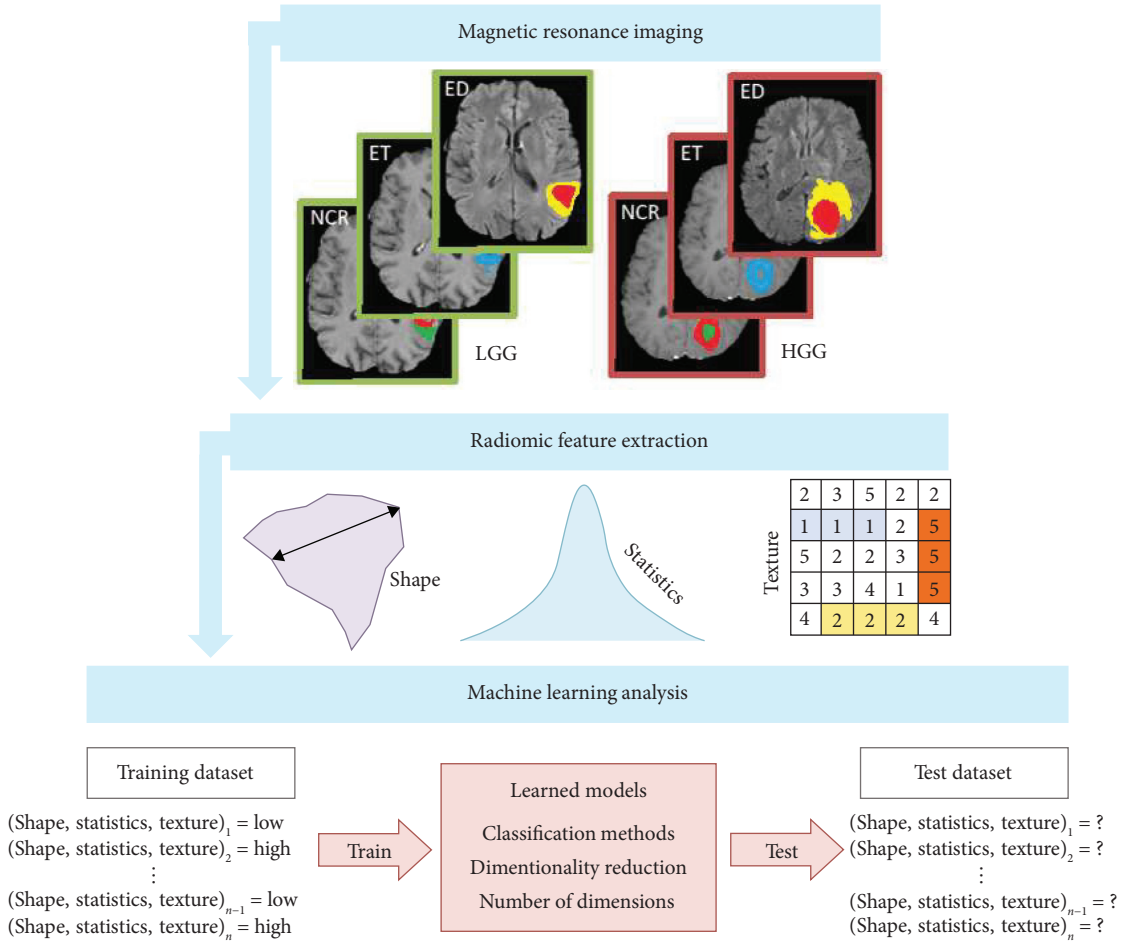

Figure 1: Proposed workflow for grade/survival classification task.

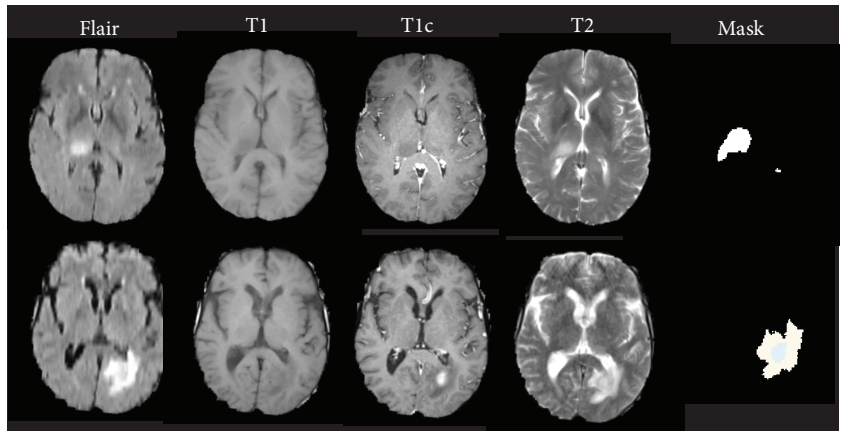

Figure 2: Examples of the four MRI modalities and the corresponding tumor masks from two randomly selected GBM patients.

2.2. Multitask VNet for Glioma MRI Data Segmentation. This study uses a multitask VNet framework to segment glioma and its different subregions from the multimodal MR image, shown in Figure 3. The network has two decoder modules with similar structures, and different decoders are assigned different tasks. The mask decoder module performs training-mask segmentation according to pixel classification tasks, and the distance transform decoder module performs regression tasks to realize distance map estimation. The structure of the encoder module and decoder module of the network is similar to the VNet. Its encoder module alternately stacks convolutional layers and downsampling layers to achieve feature extraction of the input signal under different receptive fields.

In contrast, the decoder module alternately stacks deconvolutional layers and convolutional layers in the joint

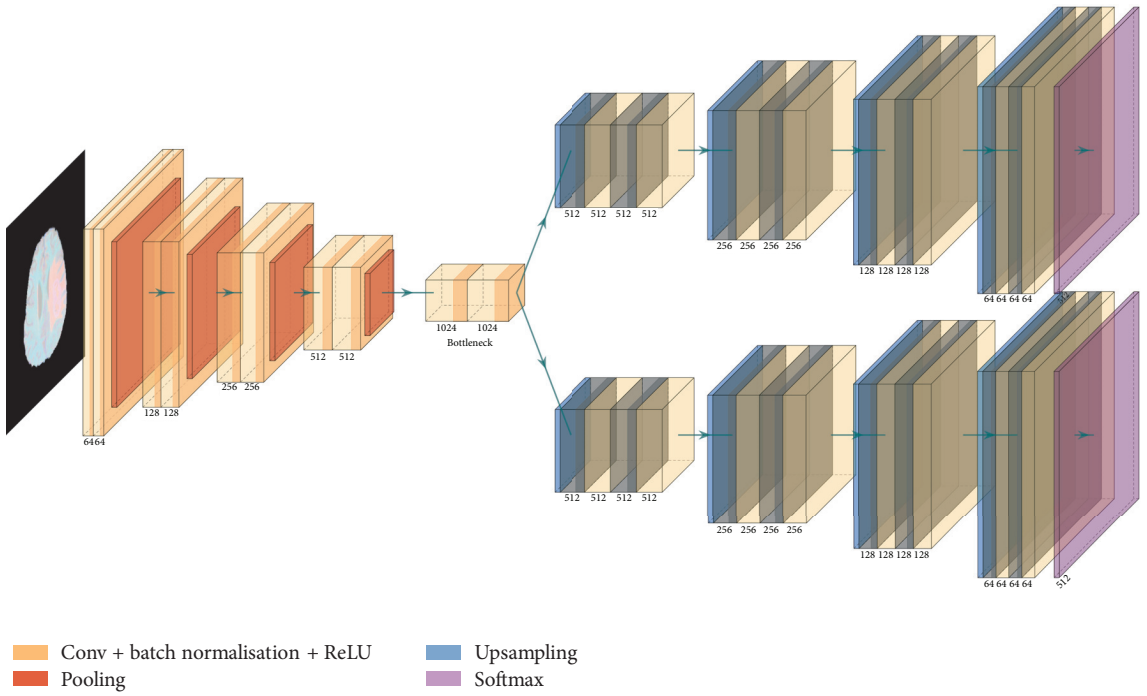

Figure 3: The overall architecture of the proposed multitask VNet.

encoder restore image resolution stage by stage based on the features extracted by the module. The model's loss function is the weighted sum of the categorical focal loss of the mask decoder block and the MSE loss of the distance transform decoder block. Its essence is that the distance map prediction regularizes the template prediction.

2.3. Quantitative Feature Extraction. Medical images contain a lot of information that can reflect the relationship between human macro performance and microenvironment. Up to now, the analysis and diagnosis of medical images are mainly based on human judgment. The disadvantage of this method is that it can only be qualitative but not quantitative. Compared with the qualitative description of human experience, quantitative features can reflect more potential information in the image. Medical imaging has developed from traditional morphological diagnosis to quantitative tumor analysis. The main difference is that the latter needs to extract and analyze more high-order quantitative image features.

Quantitative feature extraction refers to the process of extracting information from images by computer. The performance of a classification model largely depends on the features used. We extracted 16 shapes, 19 first-order statistics, 27 gray-level cooccurrence matrix (GLCM), 16 gray-level size zone matrix (GLSZM), and 16 gray-level run length matrix (GLRLM) features from each phenotype region of interest (ROI). The coiflet wavelet transform filter was also applied to each image to extract eight decompositions; for each phenotype, each decomposition's intensity-based features were calculated. The combination of shape features, first-order features, texture features, and wavelet features extracts 718 features for each image phenotype and 2154 features for each sample. Before extracting these features, voxel intensity values were normalized using the Z-score normalization in the whole brain, discretized with a bin width of 0.1, and constrained to an intensity value range of 3 standard deviations from the mean. For the Tumor Grade Dataset, some LGG samples do not contain ET segmentations. Therefore, regardless of tumor grade, these samples were removed from the analysis to keep the features equal. In addition, several mask combinations suffered from geometry mismatches and were likewise discarded. The removal of these samples from the Tumor Grade Dataset led to 44 LGG samples and 191 GBM samples remaining for the analysis. Similarly, after the removal of inappropriate samples, the Overall Survival Dataset was left with a total of 73 GBM samples with survival <12 months and 77 GBM samples with survival >12 months.

2.4. Feature Selection Methods. Radiomics leads to the creation of several informative features for use in predictive modeling. However, when the number of samples is far less than the number of features, direct classification prediction has a high computational cost and a poor effect. It may even lead to the classification prediction algorithm's failure. Hence, feature selection is needed to obtain the feature set with good performance after image feature extraction.

For machine learning models, there are many methods to reduce the feature space. Common categories of feature selection methods include filter, wrapper, and embedded methods. In addition, compared with the wrapper and embedded methods, the filter methods have the advantages of classifier independence and high computational efficiency [21]. Surprisingly, previous studies have found univariate filter methods that ignore interactions between variables can

be just as effective as multivariate methods that consider these interactions [22]. A possible alternative way for feature selection is dimensionality reduction. The complex interaction between variables is often considered by linear or nonlinear mapping, and the high-dimensional space is transformed into space with lower dimension [23]. It has been recently proposed that unsupervised dimensionality reduction techniques could have better prediction performance than filter methods in radiomic studies [24].

We utilized four unsupervised dimensionality reduction methods to build machine learning models, that is, principal component analysis (PCA), kernel PCA (KPCA), independent component analysis (ICA), and factor analysis (FA). We chose these methods due to their simplicity, computational efficiency, and easily available implementation. Moreover, these methods were compared with a univariate filter technique, ANOVA *F*-score with the top 30 features selected (FILT), and maximum 2D diameter features from each phenotype (DIAM). DIAM was chosen to investigate how our radiomic methods would compare against a commonly utilized prognostic radiological metric [25].

2.5. Models' Building. The prediction of tumor grade or overall survival in this paper is a small sample binary classification problem. To solve this problem, supervised learning in machine learning is more targeted. Supervised learning uses the training data to find rules through training to predict new samples. Training data consists of examples represented by a set of input features (radiomic features) and an output value (tumor grade or overall survival class). Once an intelligent prediction model is built from labeled data using a classifier and feature selection method, it can predict an unlabeled sample class.

We selected nine conventional machine learning techniques constructed with various classifier methods and two deep learning-based models for comparison, that is, decision trees (DT), random forest (RF), bagging (BAG), boosting (BST), Gaussian naïve Bayes (NB), multilayer perceptron (MLP), support vector machines (SVM), logistic regression (LR), *k*-nearest neighbors (KNN), convolutional neural networks (CNN), and deep neural networks (DNN). These models were chosen for their widespread use in radiomic studies and simple implementation. Models built with conventional machine learning and deep learning-based techniques are displayed in Table 1. Although the hyperparameter of classifiers can be tuned by cross-validation to improve the model performance, in our study, the classifiers were used together with the default hyperparameter settings to maintain simplicity and reduce the computational cost. Intelligent prediction models were built from combinations of feature reduction methods and classifier methods.

3. Experiment

3.1. Experimental Details. To analyze our results, a split was made by the patient. For each dataset (Tumor Grade Dataset $n = 245$ and Overall Survival Dataset $n = 150$), data were randomly split into training and testing sets with a test size = 0.2, yielding training sets containing 196/120 samples and testing sets containing 49/30 samples, respectively. To prevent the class imbalances from affecting the models' performance, we applied the synthetic minority oversampling (SMOTE) [26] technique to the tumor grade training dataset due to the existing radiomics studies that have shown SMOTE can effectively improve the classification predictive performance when the classes are imbalanced. However, SMOTE was not applied to the Overall Survival Dataset since classes were already balanced. Moreover, multicenter data and magnetic field inhomogeneities often contribute to the intensity inhomogeneities in the MR images. Therefore, we use the *Z*-score normalization as a necessary preprocessing step in dimensionality reduction to standardize features concerning the training set.

To investigate and compare the performance of different dimensionality reduction and classification approaches, a three-dimensional parameter grid for analysis was constructed in this study. For any of the four dimensionality reduction approaches, we took two as the step size ($n = 1, 3, 5, 15$) incrementally selected the number of dimensions from 1 to 15 (e.g., principal component). The training data and 11 machine learning models evaluate these dimension subsets to build the machine learning prediction model. The area under the receiver operating curve (AUC) score was calculated to evaluate the model quantitatively on the test set, which was repeated 100 times for each combination with different random splits through a bootstrap approach. The mean of the AUC values (μ_{AUC}) over all iterations was calculated to determine the given model's final AUC value. By calculating the mean over 100 iterations, we can ensure a more representative value for each model. Similarly, an empirical metric for stability, relative standard deviation (RSD) was previously defined as follows [22]:

$$\mathrm{RSD} = \frac{\sigma_{\mathrm{AUC}}}{\mu_{\mathrm{AUC}}}, \tag{1}$$

where σ_{AUC} and μ_{AUC} were the standard deviation and mean of the 100 AUC values, respectively. It should be noted that higher stability corresponds to lower RSD values.

We apply the popular open-source machine learning python library scikit-learn for model building and analysis in Python 3.6. The training and testing experiments are performed on an NVIDIA GeForce Titan RTX 24G GPU with Intel Xeon Silver 4210 2.2G GPU. The presented figures are generated using the plotting library Matplotlib. An open-source radiomics toolbox, Pyradiomics, was used for radiomic feature extraction.

3.2. Performance Measurements. There are three main experimental factors in our study which can affect the radiomics-based prediction, that is, prediction model (RF, NB, DT, BAG, BST, SVM, LR, MLP, KNN, CNN, and DNN), feature selection method (PCA, KPCA, ICA, and FA), and the number of dimensions selected (1, 3, 5, . . ., 15). Multivariate analysis of variance (ANOVA) was used to quantify these factors' impacts on AUC scores and their interactions in each classification task. To compare the variability contributed by

Table 1: Models built with various machine learning techniques.

Classifier methods	Dimensionality reduction methods	Feature selection methods
Decision trees (DT)	Principal component analysis (PCA)	ANOVA *F*-score (FILT)
Random forest (RF)	Kernel PCA (KPCA)	Max 2D diameter (DIAM)
Bagging (BAG)	Independent component analysis (ICA)	—
Boosting (BST)	Factor analysis (FA)	—
Naïve bayes (NB)	—	—
Multilayer perceptron (MLP)	—	—
Support vector machine (SVM)	—	—
Logistic regression (LR)	—	—
k-Nearest neighbor (KNN)	—	—

each factor, the variance (sum of squares) calculated for each factor was divided by total variance and multiplied by 100 to yield the percent variance for each factor.

In our study, a total of 2154 features were extracted from the segmented tumor regions of the preoperative MRI scans from the BraTS 2017 glioma dataset. For the Tumor Grade Dataset, the output classes were LGG or HGG, while for the Overall Survival Dataset, the output classes were <12-month or >12-month survival. For both classification tasks, feature selection and classification training were made using the training set, whereas the testing set was used to assess performance and stability.

3.2.1. Predictive Performance. Figure 4 depicts the performance of dimensionality reduction and classification methods using 11 dimensions for both tasks. Performance from models constructed using ANOVA *F*-score univariate filter method (FILT) and diameter features (DIAM) are also displayed. For the grade classification task, the best results among the four dimensionality reduction techniques are achieved by FA, while ICA usually performs the worst effects. Moreover, FA has comparable results to FILT, which generally has the highest predictive performance. Additionally, using diameter features alone scores much lower than any dimensionality reduction techniques. In terms of classifiers, most classifier methods show similar results except DT, which is noticeably lower. For the survival classification task, the best results among the four dimensionality reduction techniques are also often achieved by FA. Otherwise, performance results are more similar than in the grade classification task. Worthy of note is that using diameter features alone often scores comparable or higher than any dimensionality reduction techniques. Again, most classifier methods show similar results except decision trees and support vector machines (for PCA, KPCA, and ICA), which are noticeably lower. Additionally, AUC scores for the survival classification task are much lower (<0.65) than for the grade classification task (>0.80).

In addition, we repeated the above experiment by varying the number of dimensions. Figures 5 and 6 show the predictive performance corresponding to 1, 3, 5, 7, 9, 11, 13, and 15 dimensions for each feature selection method for both tasks, respectively.

3.2.2. Stability and Predictive Performance. Four AUC/RSD values corresponding to different dimensionality reduction techniques (PCA, KPCA, ICA, and FA) are generated for each prediction method. We took the median of all four AUC/RSD values for each prediction task as the representative AUC/RSD of a model. Figure 7 shows the evaluation of models' representative stability and predictive performance in each classification task. In addition, as deep learning models show significantly performance in these tasks. Hence, the performance of both the conventional machine learning techniques and two deep learning-based models (DNN and CNN) is also evaluated. Figure 7(a) shows MLP, LR, KNN, and BST should be preferred as their stability and predictive performance were higher than the corresponding median values across all classifiers. Similarly, in Figure 7(b), MLP, LR, and KNN should be preferred, with BST on the borderline of top performance and stability.

3.2.3. Experimental Factor Effect. To quantify the effect of classification methods, dimensionality reduction methods, and the number of selected dimensions, multivariate ANOVA was performed on AUC scores in this study. In Figure 8, we observed that all three experimental factors and their interactions affect both classification tasks' prediction performance. The classification method was the most dominant source of variability as it explained 36% and 37% of the total variance in AUC scores for tumor grade and survival classification tasks, respectively. The number of dimensions used was the second most dominant source of variability for both tasks as it explained 28% and 20% of the total variance in AUC scores for tumor grade and survival classification tasks, respectively. The dimensionality reduction method was the least dominant source of variability for both tasks as it explained 3% and 2% of the total variance in AUC scores for tumor grade and survival classification tasks, respectively. Interaction terms between the experimental factors followed similar trends.

3.3. Discussion. Several studies have built radiomics-based predictive models for various clinical factors such as tumor grade, prognostic outcome, treatment response, and more. However, to expand the radiomics community, studies utilizing open-source data, tools, and machine learning models, such as those used in our current investigation, are necessary. In a series of papers by Parmar et al., they evaluated the predictive performance and stability of computed tomography (CT) radiomic machine learning models constructed with various feature selection filter

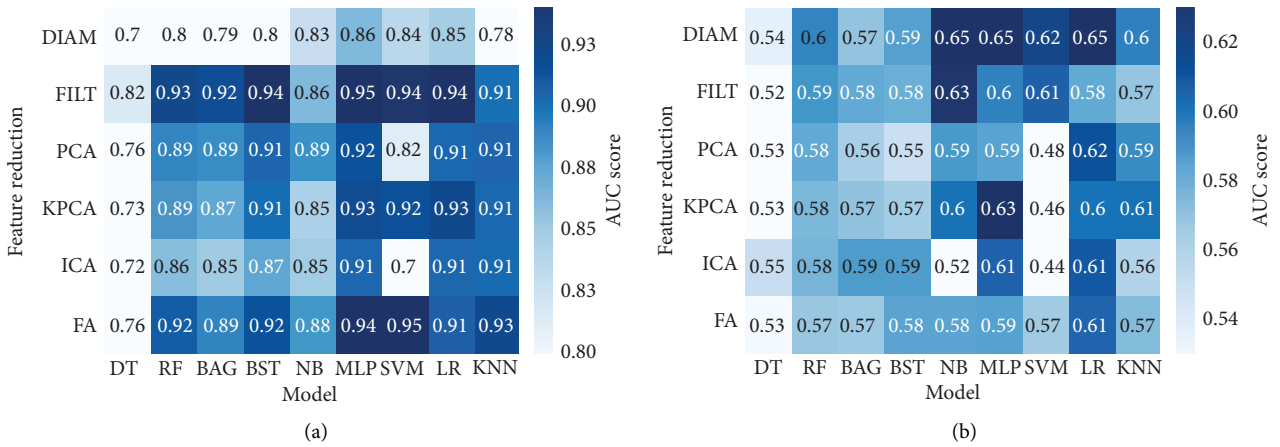

Figure 4: Predictive performance of feature reduction and classification methods. (a) Grade classification task and (b) survival classification task.

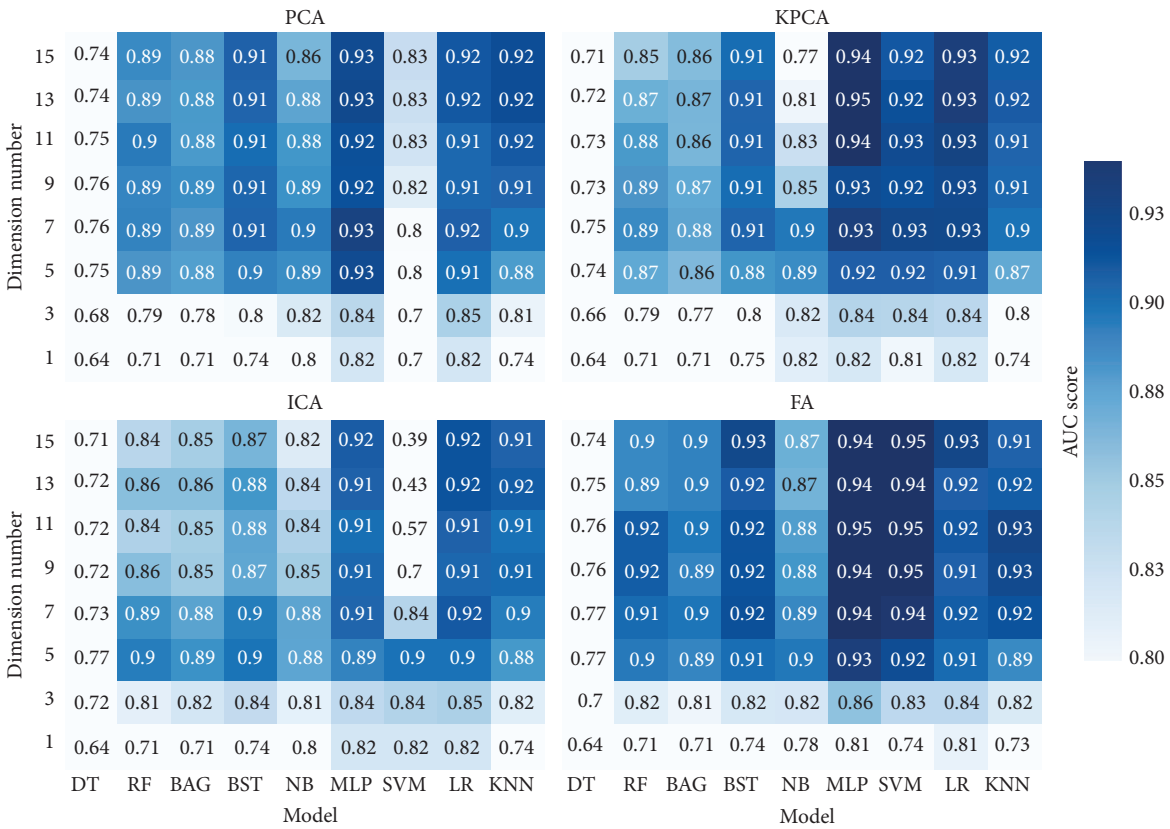

Figure 5: Predictive performance corresponding to classification methods and the number of dimensions for each dimensionality reduction method for grade classification task.

methods and classifier methods [22]. Results show that specific machine learning models perform differently depending on the cancer type, e.g., head and neck vs. lung. Therefore, it is vital to test these methods in different cancer types and various imaging modalities.

Additionally, Zhang et al. performed a similar study on lung CT with unsupervised dimensionality reduction methods and proposed dimensionality reduction methods have the potential to be superior to filter methods [27]. This study further demonstrates the variability of machine learning models constructed from different classifiers and dimensionality reduction techniques in a different cancer type (glioma) and imaging modality (MRI). We demonstrate that dimensionality reduction techniques are often lower than or comparable to filtering methods for both tasks. Specifically, we show that FA can be an improvement over PCA, which was suggested by Zhang et al. to be the best method for dimensionality reduction in radiomic studies.

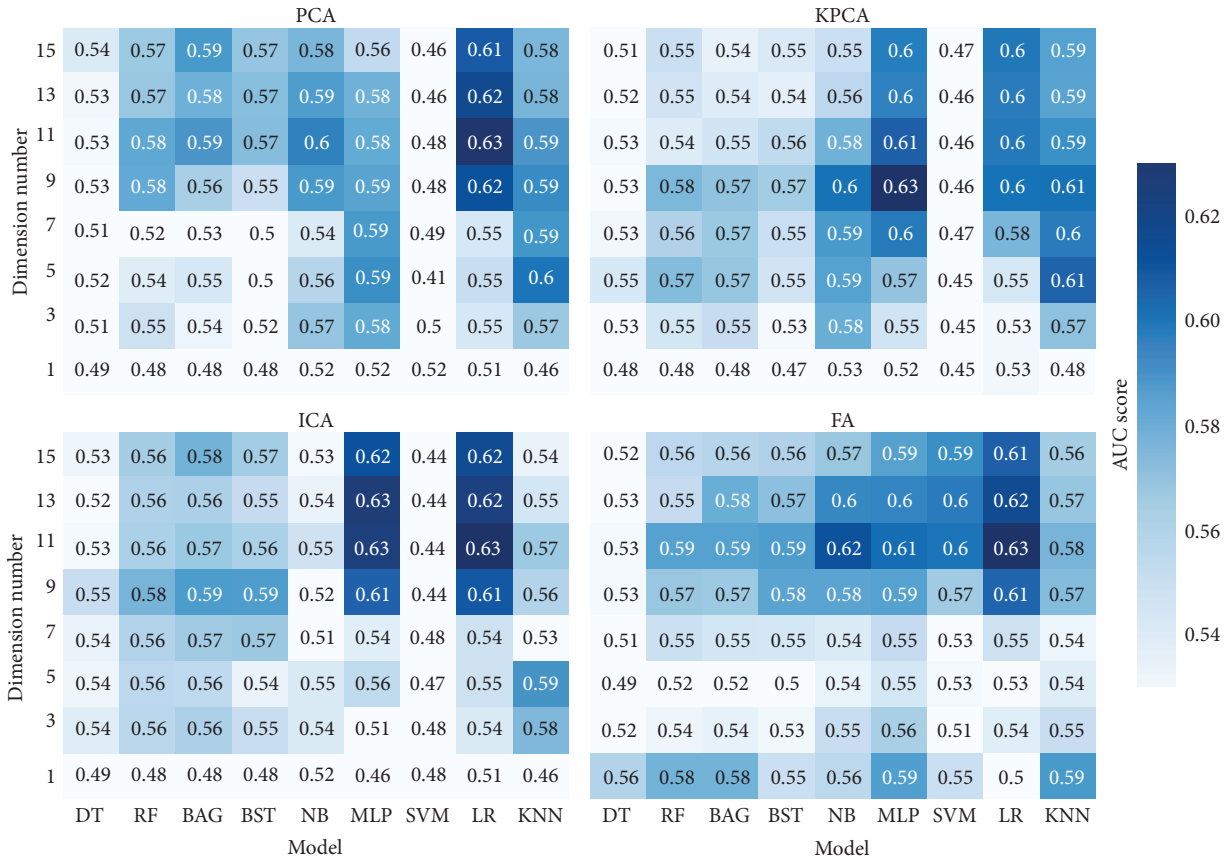

Figure 6: Predictive performance corresponding to classification methods and the number of dimensions for each dimensionality reduction method for survival classification task.

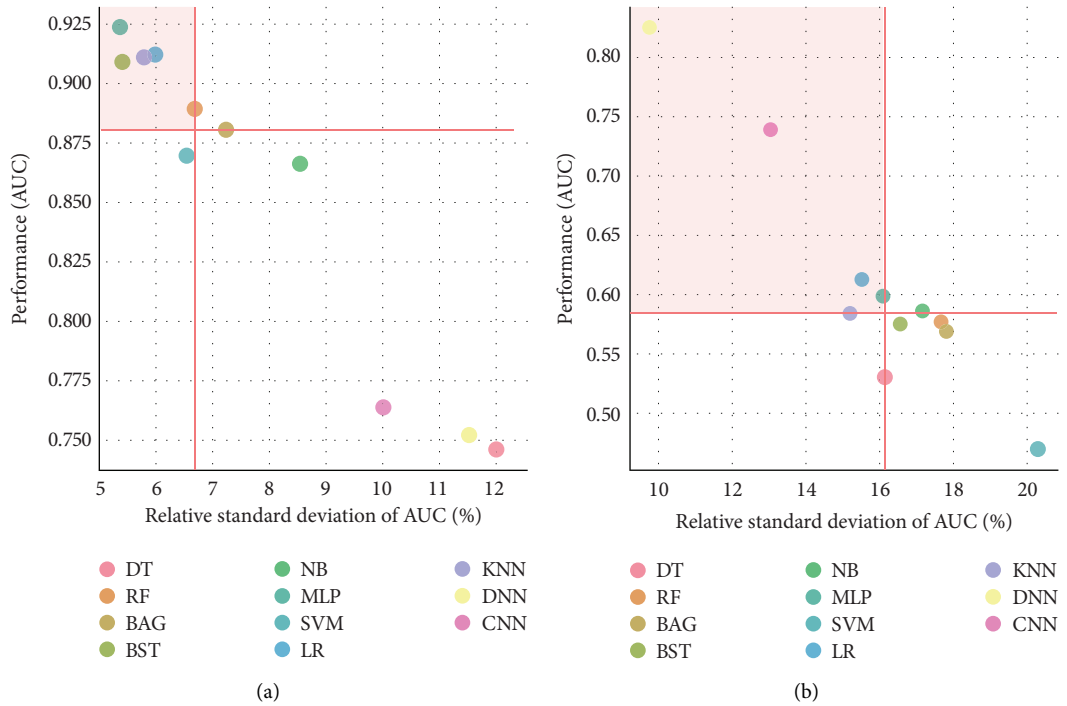

Figure 7: Scatterplots between representative stability and predictive performance of classification methods. (a) Grade classification task and (b) survival classification task.

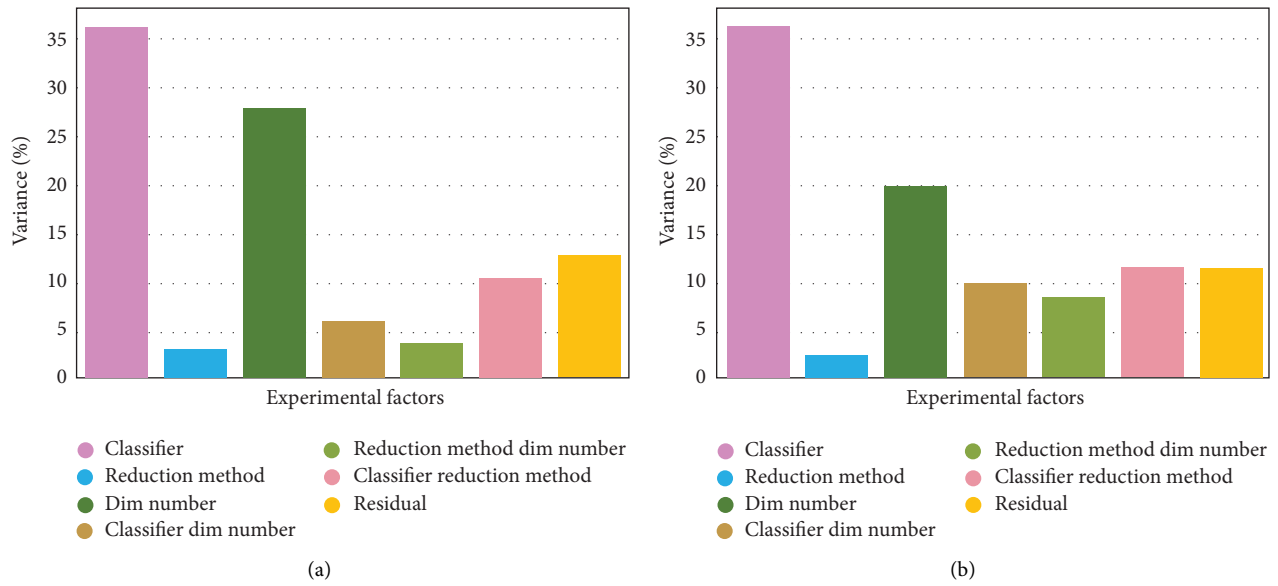

Figure 8: Variation of AUC explained by experimental factors and their interactions. (a) Grade classification task and (b) survival classification task.

ET in T1c MRI scans often used as a distinctive marker when attempting to distinguish LGG from HGG. However, since we have only used LGG samples that contain ET components, we suggest radiomics provides novel information about underlying phenotype, usually not possible in the radiological setting. Glioma grade is histopathologically diagnosed; i.e., a biopsy must be taken for classification [28]. With our radiomics approach, we suggest that imaging data may be a useful supplement to histological data. In this study, we have only classified LGG from HGG, but more grade subclasses can be assessed using these radiomics methods, e.g., grade 1 vs. grade 2 vs. grade 3 vs. grade 4. Previous studies have attempted to build machine learning models for glioma grade classification with dimensionality reduction techniques [29] or other feature selection methods [30]. Still, our results show higher predictive performance, possibly due to a more extensive training set and class balancing with SMOTE.

Predictive performance for grade classification is much higher when compared to survival classification, which is not surprising as each classification task has its own set of optimal radiomic biomarkers linked to underlying biological significance. For example, the combination of shape, first-order statistics, texture, and wavelet features utilized through dimensionality reduction leads to higher predictive performance than diameter features alone for the grade classification task. However, this is not the case for the survival classification task. Moreover, using diameter features alone in survival prediction leads to higher predictive performance than dimensionality reduction or filter techniques with all radiomic features. Previous studies have shown that texture features are challenging to gain predictive power from in GBM, with AUC values routinely falling <0.6 [17, 31]. It may be the case that current intensity-based features are not strongly linked to survival outcome in GBM, but further studies are necessary before coming to these conclusions. This study has taken a coarse approach to build machine learning models, so it may very well be the case that more refined models for survival prediction can create useful texture-based radiomics signatures for GBM survival prediction with high AUC values.

For both classification tasks, the classifier method was the most significant contribution to variability in predictive performance. A trend has commonly been observed in radiomic studies investigating machine learning models using different classifiers and feature selection methods [22]. Oppositely, Wang et al. observed that the dimensionality reduction method plays a larger role in predictive performance variability [24]. Our study has also investigated the role the number of dimensions has in variability, and it was found that it has a larger role than the dimensionality reduction method used. To our knowledge, no other studies have investigated this factor's effect on predictive performance.

Some limitations of our study are as follows. Regarding image preprocessing, we have only utilized a simple method of intensity normalization (Z-score) due to its availability in Pyradiomics. Unlike CT imaging, MRI intensity is expressed in arbitrary units, necessitating intensity standardization before radiomic analysis. More sophisticated intensity normalization methods, such as histogram-based method [32], should be explored in future studies. In addition, we have not taken advantage of classifier hyperparameter tuning and instead relied on default hyperparameter settings to save on computational costs. Future studies should employ hyperparameter tuning to increase predictive performance and stability. While our research explores our classifiers' stability, it should be noted that RSD is only an empirical method that should not be directly compared with other studies but only as a relative reference between classifiers in a given study. Additionally, our definition of a top classifier is relative to other classifiers studied, so it should not be taken as all-encompassing.

4. Conclusion

In this study, we investigate two machine learning classification tasks using radiomic features: (i) prediction of tumor grade (higher-grade vs. lower-grade gliomas) and (ii) prediction of overall survival in higher-grade gliomas (<12 months vs. > 12 months). These tasks are attempted using machine learning models constructed with various classifier methods and dimensionality reduction techniques. Models are assessed in terms of their predictive performance and stability using a bootstrap approach. Our results demonstrate that for both classification tasks, among dimensionality reduction methods, FA yielded the highest predictive performance. Similarly, MLP, LR, and KNN produced the highest predictive performance and stability among classifier methods. In addition, DT tended to perform poorly for both classification tasks. This possibly points to an underlying radiomic structure in the BraTS dataset that is preferentially fit by specific machine learning models. Where results start to diverge significantly is in the implementation of the SVM classifier. For the grade classification task, SVMs tend to perform relatively well with all feature selection methods except ICA. For the survival classification task, SVMs tend to perform poorly with all feature selection methods except FA. Interestingly, previous studies in different cancer types have suggested RF to be the best classifier method for radiomics studies. Still, it does not score among the best classifier methods for either task in our research.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was funded in part by the National Natural Science Foundation of China (Grant nos. 62072413 and 61602419), in part by the Natural Science Foundation of Zhejiang Province of China (Grant no. LY16F010008), in part by Medical and Health Science and Technology Plan of Zhejiang Province of China (Grant no. 2019RC224), and also in part by the Teacher Professional Development Project of Domestic Visiting Scholar in Colleges and Universities of Zhejiang Province of China (Grant nos. 2020-19 and 2020-20).

References

[1] L. Xu, P. Yang, W. Liang et al., "A radiomics approach based on support vector machine using MR images for preoperative lymph node status evaluation in intrahepatic cholangiocarcinoma," *Theranostics*, vol. 9, no. 18, pp. 5374–5385, 2019.

[2] Y. Wu, L. Xu, P. Yang et al., "Survival prediction in high-grade osteosarcoma using radiomics of diagnostic computed tomography," *EBioMedicine*, vol. 34, pp. 27–34, 2018.

[3] Z. Tang, Y. Xu, L. Jin et al., "Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 2100–2109, 2020.

[4] L. Liu, H. Zhang, J. Wu et al., "Overall survival time prediction for high-grade glioma patients based on large-scale brain functional networks," *Brain Imaging and Behavior*, vol. 13, no. 5, pp. 1333–1351, 2019.

[5] B. Jie, D. Q. Zhang, W. Gao et al., "Integration of network topological and connectivity properties for neuroimaging classification," *IEEE Transactions on Bio-Medical Engineering*, vol. 6, no. 2, pp. 576–589, 2014.

[6] B. H. Menze, J. Andras, B. Stefan et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transaction on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[7] A. Jungo, R. McKinley, R. Meier et al., "Towards uncertainty-assisted brain tumor segmentation and survival prediction," *International MICCAI Brainlesion Workshop*, vol. 25, pp. 474–485, 2017.

[8] D. Nie, J. F. Lu, H. Zhang et al., "Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages," *Scientific Reports*, vol. 9, pp. 1–14, 2019.

[9] S. Bhadani, S. Mitra, and S. Banerjee, "Fuzzy volumetric delineation of brain tumor and survival prediction," *Soft Computing*, vol. 24, no. 17, pp. 13115–13134, 2020.

[10] H. J. W. L. Aerts, R. V. Emmanuel, T. H. Ralph et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, pp. 1–8, 2014.

[11] D. J. Brat, R. G. W. Verhaak, K. D. Al-Dape et al., "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas," *New England Journal of Medicine*, vol. 372, no. 26, pp. 2481–2498, 2015.

[12] A. Kotrotsou, P. O. Zinn, and R. R. Colen, "Radiomics in brain tumors," *Magnetic Resonance Imaging Clinics of North America*, vol. 24, no. 4, pp. 719–729, 2016.

[13] S. Narang, M. Lehrer, D. Yang, J. Lee, and A. Rao, "Radiomics in glioblastoma: current status, challenges and potential opportunities," *Translational Cancer Research*, vol. 5, no. 4, pp. 383–397, 2016.

[14] J. Wang, C.-J. Wu, M.-L. Bao, J. Zhang, X.-N. Wang, and Y.-D. Zhang, "Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer," *European Radiology*, vol. 27, no. 10, pp. 4082–4090, 2017.

[15] L. Zhang, Z. Ye, L. Ruan, and M. Jiang, "Pretreatment MRI-derived radiomics may evaluate the response of different induction chemotherapy regimens in locally advanced nasopharyngeal carcinoma," *Academic Radiology*, vol. 27, no. 12, pp. 1655–1664, 2020.

[16] M. Ingrisch, M. J. Schneider, D. Nörenberg et al., "Radiomic analysis reveals prognostic information in T1-weighted baseline magnetic resonance imaging in patients with glioblastoma," *Investigative Radiology*, vol. 52, no. 6, pp. 360–366, 2017.

[17] P. Grossmann, V. Narayan, K. Chang et al., "Quantitative imaging biomarkers for risk stratification of patients with recurrent glioblastoma treated with bevacizumab," *Neuro-Oncology*, vol. 19, no. 12, pp. 1688–1697, 2017.

[18] B. Zhang, X. He, F. Ouyang et al., "Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma," *Cancer Letters*, vol. 403, no. 10, pp. 21–27, 2017.

[19] P. P. J. H. Langenhuizen, S. Zinger, S. Leenstra et al., "Radiomics-based prediction of long-term treatment response of vestibular schwannomas following stereotactic radiosurgery," *Otology & Neurotology*, vol. 41, no. 10, pp. E1321–E1327, 2020.

[20] J. Long, G. Ma, H. Liu et al., "Cascaded hybrid residual U-Net for glioma segmentation," *Multimedia Tools and Applications*, vol. 79, no. 33-34, pp. 24929–24947, 2020.

[21] M. M. Rahman, O. L. Usman, R. C. Muniyandi et al., "A Review of machine learning methods of feature selection and classification for autism spectrum disorder," *Brain Sciences*, vol. 10, no. 12, pp. 1–23, 2020.

[22] C. Parmar, P. Grossmann, J. Bussink et al., "Machine learning methods for quantitative radiomic biomarkers," *Scientific Reports*, vol. 5, pp. 1–11, 2015.

[23] L. Senigagliesi, M. Baldi, and E. Gambi, "Comparison of statistical and machine learning techniques for physical layer authentication," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1506–1521, 2021.

[24] T. Wang, J. Deng, Y. She et al., "Radiomics signature predicts the recurrence-free survival in stage I non-small cell lung cancer," *The Annals of Thoracic Surgery*, vol. 109, no. 6, pp. 1741–1749, 2020.

[25] N. Upadhyay and A. D. Waldman, "Conventional MRI evaluation of gliomas," *The British Journal of Radiology*, vol. 84, no. 2, pp. 107–111, 2011.

[26] J. Wei, Z. Lu, K. Qiu, P. Li, and H. Sun, "Predicting drug risk level from adverse drug reactions using smote and machine learning approaches," *IEEE Access*, vol. 8, pp. 185761–185775, 2020.

[27] Y. C. Zhang, A. Oikonomou, A. Wong et al., "Radiomics-based prognosis analysis for non-small cell lung cancer," *Scientific Reports*, vol. 7, pp. 1–8, 2017.

[28] A. Golebiewska, A.-C. Hau, A. Oudin et al., "Patient-derived organoids and orthotopic xenografts of primary and recurrent gliomas represent relevant patient avatars for precision oncology," *Acta Neuropathologica*, vol. 140, no. 6, pp. 919–949, 2020.

[29] E. I. Zacharaki, V. G. Kanas, and C. Davatzikos, "Investigating machine learning techniques for MRI-based classification of brain neoplasms," *International Journal of Computer Assisted Radiology and Surgery*, vol. 6, no. 6, pp. 821–828, 2011.

[30] J.-b. Qin, Z. Liu, H. Zhang et al., "Grading of gliomas by using radiomic features on multiple magnetic resonance imaging (MRI) sequences," *Medical Science Monitor*, vol. 23, no. 9, pp. 2168–2178, 2017.

[31] D. Yang, G. Rao, J. Martinez, A. Veeraraghavan, and A. Rao, "Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma," *Medical Physics*, vol. 42, no. 11, pp. 6725–6735, 2015.

[32] R. T. Shinohara, E. M. Sweeney, J. Goldsmith et al., "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, no. 8, pp. 9–19, 2014.