Oncology

Review

Oncology 2021;99:433–443
DOI: 10.1159/000515597

Received: January 16, 2021
Accepted: February 28, 2021
Published online: April 13, 2021

Current Status and Quality of Machine Learning-Based Radiomics Studies for Glioma Grading: A Systematic Review

Mohsen Tabatabaei[a] Ali Razaei[b, d] Amir Hossein Sarrami[c] Zahra Saadatpour[b, d] Aparna Singhal[b, d] Houman Sotoudeh[b, d]

[a]Health Information Management, Office of Vice Chancellor for Research, Arak University of Medical Sciences, Arak, Iran; [b]Department of Radiology, University of Alabama at Birmingham (UAB), Birmingham, AL, USA; [c]University of Semnan, Semnan, Iran; [d]Division of Neuroradiology, Department of Radiology, University of Alabama at Birmingham (UAB), Birmingham, AL, USA

Keywords
Glioma · Neoplasm grading · Artificial intelligence · Systematic review

Abstract
Introduction: Radiomics now has significant momentum in the era of precision medicine. Glioma is one of the pathologies that has been extensively evaluated by radiomics. However, this technique has not been incorporated into clinical practice. In this systematic review, we selected and reviewed the published studies about glioma grading by radiomics to evaluate this technique's feasibility and its challenges. ***Material and Methods:*** Using seven different search strings, we considered all published English manuscripts from 2015 to September 2020 in PubMed, Embase, and Scopus databases. After implementing the exclusion and inclusion criteria, the final papers were selected for the methodological quality assessment based on our in-house Modified Radiomics Standard Scoring (RQS) containing 43 items (minimum score of 0, maximum score of 44). Finally, we offered our opinion about the challenges and weaknesses of the selected papers. ***Results:*** By our search, 1,177 manuscripts were found (485 in PubMed, 343 in Embase, and 349 in Scopus). After the implementation of inclusion and exclusion criteria, 18 papers remained for the final analysis by RQS. The total RQS score ranged from 26 (59% of maximum possible score) to 43 (97% of maximum possible score) with a mean of 33.5 (76% of maximum possible score). ***Conclusion:*** The current studies are promising but very heterogeneous in design with high variation in the radiomics software, the number of extracted features, the number of selected features, and machine learning models. All of the studies were retrospective in design; many are based on small datasets and/or suffer from class imbalance and lack of external validation datasets.
© 2021 S. Karger AG, Basel

Introduction

Gliomas are the most common primary malignancies of the central nervous system. Histologically, they show the glial cells' characteristics and are generally classified due to these similarities [1]. Traditionally, gliomas have been classified into slow-growing lesions (WHO grades 1 and 2) and rapidly progressive lesions (WHO grades 3 and 4). WHO grade 1 and 2 gliomas are considered low-

karger@karger.com
www.karger.com/ocl
© 2021 S. Karger AG, Basel

Correspondence to:
Houman Sotoudeh, hsotoudeh@uabmc.edu

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

grade gliomas (LGG), while grades 3 and 4 are classified as high-grade gliomas (HGG). Glioblastoma multiforme (GBM, WHO grade 4) is the most aggressive type of glioma with poor prognosis and median survival of about 15 months, even after multimodal therapy [2]. However, this classification has been modified in the most recent WHO classification. Based on the 2016 WHO glioma classification, a glioma is classified based on histopathology characteristics and molecular fingerprints [3]. The WHO glioma classification will be modified soon, and a new version will be released by mid-2021. It is estimated that such new classification relies more on the molecular fingerprints [4–8]. Despite the momentum toward the molecular fingerprint, the glioma classification into LGG versus HGG or grade 1, grade 2, grade 3, and grade 4 is still commonly used in clinical practice. Also, glioma classification into low-grade and high-grade is still standard of care in the radiology reports [9–11].

Radiomics includes recently emerging techniques that convert digital medical images into mineable data by extracting quantitative descriptors and can potentially quantify tumor characteristics. Using radiomics, through the mathematical models built based on selected radiomics features, predicting the tumor phenotype, molecular markers, and the prognosis is feasible [12]. Recent studies have shown encouraging results of applying radiomics in oncological practice [13–15]. This method can enhance the traditional imaging analysis and provide personalized medicine for patients [16]. The power of radiomics in quantifying distinct tumor types and, consequently, tumor grading and predicting different cancers' survival has been demonstrated by many experimental studies [16–21]. The preliminary investigations about the role of radiomics for glioma are promising. It seems that radiomics can provide an acceptable method to characterize the histologic and molecular features of different glioma subtypes. By extracting numerous image features based on tumor geometry, histogram, and texture analysis, radiomics can effectively characterize tumor phenotypes [22]. Glioma is one of the widely evaluated tumors by radiomics technology [13, 23, 24]. Despite the facts mentioned above, radiomics is not a part of the standard of care clinical practice [24]. Lack of clinical application can be attributed to several weaknesses in radiomics study designs, including lack of standard guidelines about using a special radiomics software, the number of extracted features, feature selection technique, and the machine learning models [24, 25].

This systematic review was conducted to analyze the most recent studies in glioma grading by machine learning-based radiomics, evaluate the possibility of clinical usage of this technique, and reveal the weaknesses that must be resolved in the future.

Materials and Methods

Article Search Strategy and Study Selection

A database search was conducted in PubMed, Scopus, and EMBASE to identify all relevant published researches. Papers published in these databases from 2015 until 2020 were included. Our research contains papers published before September 2020. The search terms used to find radiomics studies were "glioma" OR "astrocytoma" OR "glioblastoma multiforme" OR "glioblastoma" AND "radiomics" OR "radiogenomics" OR "artificial intelligence." Overall, seven different strings were searched in the datasets mentioned above (online supplement 1; for all online suppl. material, see www.karger.com/doi/10.1159/000515597).

Data Extraction and Analysis

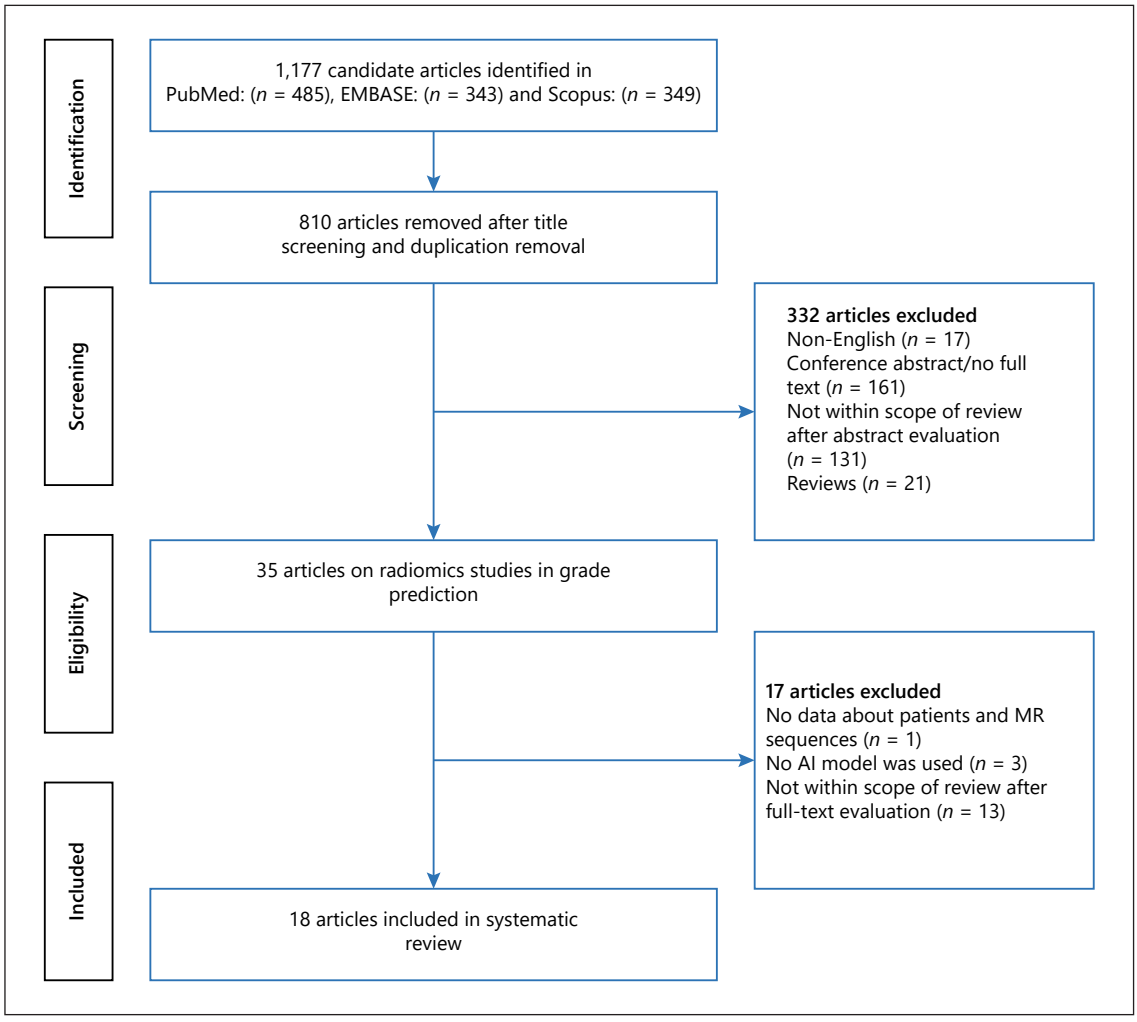
The search results were then reviewed by two experienced reviewers (H.S., with 16 years of experience in neuro-oncologic imaging and artificial intelligence, and M.T., with 9 years of experience in health information management and artificial intelligence) by paper title. The papers that were related to the grading of glioma and radiomics were then selected by these two reviewers by consensus. Subsequently, five reviewers (Z.S., A.R., A.H.S., who had 14 years of experience in medical imaging, H.S., and M.T) evaluated the eligible radiomics performance studies. The PRISMA flowchart of study is shown in Figure 1.

Before performing their analysis, an online expert panel was convened to review and discuss the items listed in the RQS and ensure they all had explicit knowledge of RQS. The detailed RQS score was adopted and modified from another group's prior published study [26]. The reviewers extracted the data using a predetermined RQS evaluation according to 43 components for each article: Title (1 item), Abstract (5 items), Introduction (6 items), Materials and Methods (23 items), Results (1 item) and Discussion (7 items). For each item, score 1 or 0 was assigned to the papers. The only exception was "using the external dataset," for which a score of 2 was assigned with total score range for any paper between 0 and 44 (online supplement 2). Also, for each article, a questionnaire consisting of 10 items regarding the radiomics pipeline was filled, which included: 1. type of dataset, 2. MRI sequences used, 3. number of patients, 4. number of extraction features before and after feature selection, 5. software for radiomics, 6. AI model, 7. number of patients in each grade, 8. type of performance metrics, 9. most important findings, and 10. limitations and weaknesses. Each article was evaluated by two of the five independent reviewers. Disagreements between any two reviewers were discussed at a research meeting attended by the reviewers and an additional reviewer until the consensus was reached. The entire processes of RQS evaluation and questionnaire filling was double-checked separately by H.S. and M.T. and disagreement was resolved by consensus.

Statistical Analysis

All statistical analyses were performed using SPSS (SPSS version 22; SPSS, Chicago, IL) and R (R version 3.3.3; R Foundation for Statistical Computing, Vienna, Austria).

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM



Color version available online

Fig. 1. The PRISMA flowchart of this systematic review.

Results

A total of 1,177 records were identified until September 13, 2020. Retrieved article titles were screened for eligibility. After title screening and removal of duplicates, 810 papers were excluded. Screening of the abstracts of the remaining 367 articles was performed. Abstract review further excluded 332 articles for the following reasons: non-English (n = 17), conference abstract/no full text (n = 161), not within scope of review (n = 131), reviews (n = 21), and pediatric (n = 2). Full-text reviews of the 35 potential articles was performed. After full-text evaluation, one paper was excluded because the number of patients and MR sequences were not reported. Thirteen papers were out of the scope of this review. Three papers did not use machine learning models, which revealed 18 papers for the final review (four papers from 2020, seven papers from 2019, five from 2018, and two from 2017). The total RQS score ranged from 26 (59% of maximum possible score) to 43 (97% of maximum possible score) and with a mean of 33.5 (76% of maximum possible score).

The design of the final papers has been summarized in Table 1. All papers used MRI for the imaging modality, and all of them were retrospective in design. The patient number ranged from 40 to 285 (mean: 157, SD:94). In seven studies, public datasets were used. In nine studies, in-house datasets were used. In two papers, the radiomics pipeline was developed on an in-house dataset and was tested on an external dataset. The used softwares for feature extraction were IBEX: n = 1, Pyradiomics: n = 7, Matlab: n = 11, MaZda: n = 1. The number of extracted fea-

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

Oncology 2021;99:433–443
DOI: 10.1159/000515597

Table 1. The study design of the eighteen final papers

Authors, date of publication	Type of dataset	MRI modality/sequences used	Patients, *n*	Number of extraction features (per patient)		Software for feature extraction	Feature selection	AI model	Type of grade (number of patients)	
				Before feature selection	After feature selection				LGG	HGG
Cui et al., 2018d [27]	Public	T1, T2, T1+C, FLAIR	80	6,920	52	IBEX	LASSO, *t* test	DT, LASSO	40	40
Chen et al., 2020 [28]	Public	T1, T2, T1+C	285	505	50	Pyradiomics, Matlab	PLS	SVM, RF, LR	75	210
Xiao et al., 2019 [29]	Public	T1+C	285	4,956	25	Pyradiomics	RFE	LR, SVM, LDA	75	210
Wu et al., 2018 [30]	In-house	T1, T2, T1+C, FLAIR	161	346	68/19	Matlab	LASSO and Glmnet	Linear regression	52	109
Vamvakas et al., 2019 [31]	In-house	T1, T1+C, T2, FLAIR, MR perfusion DCE, DWI [ADC], DTI, MRS	40	581	21	Matlab, MaZda	SVM-RFE algorithm, SVM classifier	SVM	20	20
Lu et al., 2018 [32]	Public, in-house	T1+C, FLAIR, T2	214	39,212	20–1,960	Matlab, MR Radiomic platform	*t* test	SVM	108	106
Hashido et al., 2020 [33]	In-house	MR perfusion ASL, MR perfusion DSC	46	91	75	Pyradiomics	LASSO + *t* test	LR	15	31
Cho et al., 2018 [34]	Public	T1, T2, FLAIR, T1+C	285	468	5	Matlab, Pyradiomics	MRMR algorithm	SVM, LR, RF	75	210
Lin et al., 2017 [35]	In-house	T1, T2, FLAIR, T1+C	161	346	19	Matlab	Elastic net	LR	N/S	N/S
Cho et al., 2017 [36]	Public	T1, T2, FLAIR, T1+C	108	180	16–34	N/S	LASSO	LR	54	54
Takahashi et al., 2019 [37]	In-house	T2, DWI [ADC], DTI, DKI	55	2,856	6	Matlab	RFE	LR, SVM	14	41
Park et al., 2019 [38]	In-house and public	T1+C, T2, FLAIR	204	250	N/S	Pyradiomics	*t* test, RFE, ROSE	Elastic net, LDA, RF, GBD	121	83
Wang et al., 2019 [39]	In-house	T1+C, DWI [ADC], T2	85	652	15	Matlab	LASSO	LR	50	35
Tian et al., 2018 [40]	In-house	DWI [ADC], T2, T1, T1+C, MR perfusion ASL	153	420	28/30	Matlab	RFE	SVM	42	111
Wu et al., 2018 [41]	In-house	T1+C	156	360	31	Matlab	Pearson's R, K-mean	LR, RF, KNN, SVM	46	110
Gao et al., 2020 [42]	In-house	T1+C	369	1,372	15	Pyradiomics	χ^2, heatmap, RF	LR, SVM, and RF	147	222
Çinarer et al., 2020 [43]	Public	T2 and FLAIR	121	744	126	Pyradiomics	Mann-Whitney U test	DNN	77	44

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

Table 1 (continued)

Authors, date of publication	Type of dataset	MRI modality/sequences used	Patients, *n*	Number of extraction features (per patient)		Software for feature extraction	Feature selection	AI model	Type of grade (number of patients)	
				Before feature selection	After feature selection				LGG	HGG
Bi et al., 2019 [44]	Public	T1, T2, FLAIR	50	483	N/S	Matlab	LASSO, decision tree SVM		17	33

The table includes the type of datasets, the MR sequences used for radiomics analysis, number of patients in each study, number of extracted features (before and after feature selection), the software used for feature selection, the model used for feature selection, the artificial intelligence model that was trained on the selected features, and number of patients in low-grade glioma (LGG) and high-grade glioma (HGG) cohorts. MRI, magnetic resonance imaging; IBEX, Imaging Biomarker Explorer; AI, artificial intelligence; LGG, low-grade glioma; HGG, high-grade glioma; AK, Artificial intelligence Kit; GBDT, gradient descent algorithm; LASSO, least absolute shrinkage and selection operator; PLS, partial least squares; LDA, linear discriminant analysis; RF, random forest; LR, logistic regression; SVM, support vector machine; Glmnet, elastic net regularized generalized linear model; SRC, Spearman rank correlation; N/S, not specified; RFE, Recursive feature elimination; MRMR, minimum redundancy maximum relevance; ROSE, random oversampling examples; GBM, gradient boosting machine; DNN, deep neural network; KNN, k-nearest neighbors; ANOVA, analysis of variance.

tures for each patient ranged from 91 to 39,212 (mean: 3,552 features). The used dimensionality reduction techniques were: least absolute shrinkage and selection operator ($n = 6$), *t* test ($n = 4$), partial least squares ($n = 1$), recursive feature elimination ($n = 5$), elastic net regularized generalized linear model ($n = 1$), support vector machine ($n = 1$), minimum redundancy maximum relevance algorithm ($n = 1$), elastic net ($n = 1$), random oversampling examples ($n = 1$), Pearson's *R* ($n = 1$), K-mean ($n = 1$), χ^2 ($n = 1$), heatmap ($n = 1$), random forest ($n = 1$), Mann-Whitney U test ($n = 1$), and decision tree ($n = 1$) (many studies used several dimensionality reduction techniques). The used AI models were: decision tree ($n = 1$), least absolute shrinkage and selection operator ($n = 1$), support vector machine ($n = 9$), random forest ($n = 5$), logistic regression ($n = 10$), linear discriminant analysis ($n = 1$), linear regression ($n = 1$), elastic net ($n = 1$), gradient descent algorithm ($n = 1$), k-nearest neighbors ($n = 1$), deep neural network ($n = 1$) (several studies used more than one AI model). The used MR sequence used for feature extraction were: T1 ($n = 9$), T2 ($n = 14$), FLAIR ($n = 10$), T1+C ($n = 14$), DWI/ADC ($n = 4$), MR perfusion ($n = 3$), diffusion tensor imaging/diffusion kurtosis imaging (DTI/DKI) ($n = 2$), MR spectroscopy ($n = 1$). The performance of each study and its most important findings and limitations have been summarized in Table 2.

Discussion

Our systematic review identified and evaluated the research papers dealing with radiomics analysis of gliomas for grading purposes. The role of radiomics feature extraction has been explored in association with different types of machine learning. The preliminary results are promising with high sensitivity, specificity, accuracy, and AUC, as described in Table 2. However, the performance metrics provided by many studies are based on the "test and train" on a single. In two studies, the developed radiomics platforms were used for external datasets with reported AUCs of 94% [32] and 72% [38], which would be more realistic performances.

Based on our analysis, we now know several facts about radiomics and glioma grade prediction: A) It appears that using multiple MR sequences for feature extraction is more efficient than a single sequence. Nevertheless, T1+C was reported as the most important sequence. B) Adding advanced MR techniques (MR perfusion and MRS) can improve radiomics performance. C) Feature extraction of multiple areas (intratumoral, enhancing tumor, and as-

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

Oncology 2021;99:433–443
DOI: 10.1159/000515597

Table 2. Performance, important findings, and limitations of the reviewed papers

Article author [Ref]	AI model	Sen, %	Spec, %	ACC, %	AUC	Important findings	Limitations and weaknesses
Cui et al. [27]	RF	N/S	N/S	91.3	0.956	– Combining four sequences is more accurate than using a single sequence – The most important sequence for feature extraction is T1+C – The most predictive features are shape features – For T1 and T1+C sequences, the most important features are the "shape" and then the "intensity" features – For FLAIR and T2, after shape features, the most predictive features are GLCM	– Retrospective – Public dataset – The number of patients in each grade is unknown
Chen et al. [28]	SVM RF LR	94 94 93	95 93 95	94 94 94	0.99 0.98 0.98	– Wavelet features are more predictive than traditional features – Feature extraction from intratumoral + peritumoral regions is more predictive than each alone – Accuracy: SVM > LR > RF (SVM on wavelet features from intratumoral + peritumoral has the best performance) – The wavelet features are less susceptible to image intensity variation and image deformity	– Retrospective – Public dataset – The number of patients in each grade is unknown – Class imbalance
Xiao et al. [29]	LR SVM LDA Accumulate	88.1 92.4 92.3 90.9	86.7 77.3 68 77.3	87.7 88.4 86 87.4	0.927 0.933 0.912 0.924	– SVM is more accurate than LR or LDA – The best features: GLDM > NGTDM = GLRLM > GLCM > First orders – Feature extraction based on deep learning is less effective than traditional radiomics feature extraction – Combination of "deep learning" + "traditional" feature extraction is more predictive than each of them alone	– Retrospective – Public dataset – The number of patients in each grade is unknown – Class imbalance
Wu et al. [30]	Linear regression	93.5	86.5	91.3	0.96	– Dimensionality reduction improves the prediction	– Retrospective – Number of patients in each grade is unknown – Class imbalance
Vamvakas et al. [31]	SVM	95	96	95.5	0.955	– MR spectroscopy is feasible for radiomics – The lipids/Cr level (from MR spectroscopy) was the most predictive sequence in this study – GLCM features are more predictive than GLRLM features	– Retrospective – Small size of the study – Only the intra-lesion area was used for feature extraction
Lu et al. [32]	SVM (validation)	82.6	90.5	87.7	0.94	– T1+C was the most important sequence to differentiate LGG from HGG. – Adding T2 and FLAIR can improve the classification – Texture features are more important than shape and size features – Using more MR sequences enhances the classifier performance – Used external dataset. No class imbalance. The reported classifying performance would be more realistic	– Retrospective
Hashido et al. [33]	LR	85.7	85.7	N/S	0.88	– DCS better than ASL for differentiation LGG versus HGG	– Retrospective – Small size of study – Class imbalance
Cho et al. [34]	SVM LR RF Accumulate	94 96 94 95	69 68 73 70	88 88 88 88	0.88 0.90 0.92 0.90	– RF was the most predictive model – The ensemble techniques were not able to improve the RF performance – This study used different areas for feature extraction, which include enhancing tumors, non-enhancing tumors, necrosis, and edema	– Retrospective – Class imbalance
Lin et al. [35]	LR	96.3	73	88.8	0.949	– The semi-automatic segmentation of the glioma on MRI is feasible and, in association with radiomics techniques, has an acceptable performance to detect glioma grade	– Retrospective

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

Table 2 (continued)

Article author [Ref]	AI model	Sen, %	Spec, %	ACC, %	AUC	Important findings	Limitations and weaknesses
Cho et al. [36]	LR	88	90	89	0.88	– T1-contrast image is the best sequence to difference between HGG and LGG	– Retrospective
Takahashi et al. [37]	LR SVM	N/S N/S	N/S N/S	91 91	90 93	– 6 features from ADC and MK had the best performance	– Retrospective – T1+C was not used – The patient assignment to train and test groups were not random
Park et al. [38]	Elastic net RFE + GBM ROSE + RF	92.9 72.6 83.3	70 60.4 77.4	79.4 66.7 78.4	0.85 0.72 0.82	– The trained models have been used on the external validation cohort. The performance metric reported in this study would be more realistic – Radiomics is helpful to differentiate grade 2 versus 3 – T1+C is likely the most important sequence – Performance of radiomics is weak for areas of non-enhancing tumor – The performance of the best classifier was good in the internal validation set (AUC, 0.85) and fair in the external validation set (AUC, 0.72) to predict LGG – For the non-enhancing LGG subgroup, the performance of the best classifier was good in the internal validation set (AUC, 0.82) but low in the external validation set (AUC, 0.68)	– Retrospective – Cases from TGCA are heterogeneous
Wang et al. [39]	LR	N/S	N/S	N/S	N/S	– Radiomics signature is associated with grade. The combination of multiple sequences is better than a single sequence – T1+C is the most important sequence	– Sen, Spec, ACC, or AUC were not reported – Sex and age were used conjoined with radiomics, so evaluating the role of pure radiomics is not possible
Tian et al. [40]	SVM	96.4	97.3	96.8	0.98	– Texture features are better than histogram parameters	– Class imbalance but was solved by SMOTE
Wu et al. [41]	LR	94	88	88	0.85	– LR is better than RF, KNN, and SVM for differentiation of LGG from HGG – Feature selection increases the ACC, Sen, and AUC for 5% and Spec for 13%	– Class imbalance
Gao et al. [42]	RF	63	89	81	0.79	– RF is more predictive than LR and SVM	– Class imbalance but was solved by SMOTE
Çinarer et al. [43]	DNN	100	N/S	96.1	0.75	– 3D feature extraction and wavelet feature using DNN are feasible for grading	– Only T2 and FLAIR sequences were used – Class imbalance
Bi et al. [44]	SVM	80	N/S	72	0.65	– Both LASSO and information gain are useful for feature selection – T2 features are more predictive than T1 and FLAIR features – GLCM is the most predictive texture feature	– Class imbalance

Sen, sensitivity; Spec, specificity; ACC, accuracy; AUC, area under the curve.

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

sociated edema) is more efficient than a single lesion. D) There is no agreement about the number of the extracted features and the selected features. E) The Matlab and Pyradiomics libraries are the most commonly used softwares for feature extraction. F) Adding wavelet features to the traditional features will improve radiomics performance. G) Dimensionality reduction and feature selection is a universal approach and increases the prediction accuracy. LASSO is the most commonly used technique for feature extraction. However, many other feature selection techniques were also used. H) Traditional feature extraction (e.g., using Pyradiomics and MatLab) is more effective than deep-learning-based feature selection. I) Gray-level co-occurrence matrix (GLCM) features are probably the most predictive features. J) SVM, LR, and RF are the most promising machine learning models for grading prediction. The deep neural networks are not common for glioma grade prediction.

The most common limitations in the selected papers are as below:

1. Retrospective studies: The retrospective nature of the studies is a very common limitation for radiomics studies. All of the selected studies in our systematic review are retrospective. Such a study design is susceptible to various biases, including selection bias.
   Potential solution: By conducting prospective studies, it is possible to use accurate inclusion and exclusion criteria and to have a more realistic estimation of radiomics performance. Nevertheless, such prospective studies can be very time-consuming in relatively small medical centers and small patient populations [45, 46].
2. Dataset: Using public datasets is a very common approach in radiomics studies. In our review, seven studies have been performed on the Cancer Genome Atlas (TCGA) dataset [47–49]. That means many of the studies are essentially using a similar patient population again and again. Their results cannot be generalized to the real world. Also, images in these datasets are from different medical centers and different MR scanner vendors with significant heterogeneity in image acquisition and reconstruction. In addition, these public datasets suffer from old technology. Many cases in TCGA datasets belong to the early 2000s. These MR images have been performed by 90s MR scanners, which now are nearly completely out of use. Sequence parameters, reconstruction techniques, noise, and many other image properties of such images are entirely different than today's state-of-the-art MR images. The estimated radiomics results based on old images cannot be generalized to today's patient population.
   Potential solution: Using an in-house dataset of recent patients would be a more realistic estimation of the radiomics approach. Also, for using a public dataset, the researchers must consider the time of the scan, scanner vendor, sequence physics, reconstruction techniques, and noise before including the case in their study. The old images must be excluded [45, 46, 50].
3. Class imbalance: GBM is the most common subgroup of glioma and comprises most of the datasets (public and in-house). If the prevalence of one condition (GBM) is significantly higher than other conditions (grades 1–3), the study is susceptible to imbalance and overfitting. The developed model works perfectly in the training dataset but very poorly in the real world. In our review, 44.4% of studies suffer from imbalance.
   Potential solution: By arranging the dataset before the study and assigning a relatively same number of different grades, the imbalance can be avoided. This approach is challenging if the entire dataset is small. In this situation, there are several techniques to avoid imbalance. In our review, SMOTE was the most common approach to subside the imbalance between the different grades [51].
4. External validation: All of the challenges mentioned above can reduce the generalization of developed models because of overfitting. Developing a radiomics solution on one training dataset and using it for another dataset from another medical center (external validation) is the best strategy to ensure that the model works well. Unfortunately, this approach is not popular in most of the radiomics and medical AI literature. So far, only about 6% of medical AI publications used external validation [46]. In our review about glioma grading, only two studies used external datasets [32, 38]. Park et al. [38] reported a decrease in the performance for the external dataset. This fact raises concern about the reported performances in other studies as well. The performances of the other developed models are likely overestimated by overfitting [38].
   Potential solution: Testing the trained radiomics model on a new dataset from another patient population and medical center is the most robust technique to ensure that the study does not suffer from variable biases and poor design [46].
5. Moving target: The grading guidelines of malignancies, including gliomas, are continually evolving. The WHO has released the glioma grading system in 2007 and 2016, and the next guideline will be released by mid-2021 [52]. These systems are gradually evolving from a pure histopathologic approach to the molecular

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

fingerprint for classification. In this context, one tumor can be assigned to different grades using different WHO criteria. This effect would be more dramatic by using the old public datasets.
Potential solution: A prospective study design would decrease this limitation. Also, in a retrospective study, repeat grading is necessary. Repeat grading using the molecular diagnosis can be very costly and even impossible if the tissue is not available to repeat the molecular diagnosis [52].

6. Vague patient population: In our review, five studies mentioned the exact number of patients in each grade [37, 38, 40, 42, 43]. Other papers have used patients in LGG and HGG classes.
Potential solution: The exact number of patients in each grade must be reported. In this way, the role of radiomics for each grade can be evaluated separately. Radiomics may work well for one grade but may be poor for another grade. At this time, the performance of radiomics for each glioma grade remains unknown [37, 40].
7. Small dataset: In our review, the average number of patients in studies that used the in-house and public dataset was 157. Generalization of the results of these studies to the real world would be challenging.
Potential solution: Combining datasets from different medical centers will increase the patient population and improve the studies [53].
8. Heterogeneity of radiomics techniques: There is no standard approach for feature extraction. In many studies, the features have been extracted by in-house developed software, which are not reproducible by other researchers. There is no consensus about the number of extracted features, traditional versus wavelet features, MR sequence, segmentation areas on images, feature selection technique, and machine learning models [54].
Potential solution: Using a standard radiomics pipeline will facilitate the incorporation of radiomics into daily practice. In this context, we suggest using an open-source feature extraction software such as Pyradiomics, extracting wavelet along with traditional features, using at least three areas (intratumoral, enhancing part of the tumor and associated edema), using different MR sequences (T1, T2, FLAIR, T1+C, DWI/ADC, MR perfusion, and MRS), using LASSO for feature selection, and using SVM, LR, or RF for the pipeline.

The limitations of this systematic review are as follows. First, only three literature databases (PubMed, Scopus, Embase) were included. Second, only published full-text English language articles with available full text were included. The conference abstracts were not included. Third, only a few papers have reported the exact number of patients in each grade, so we included the papers that reported LGG versus HGG as well.

Conclusion

We reviewed the prior studies about radiomics geared toward glioma grading. By implementation of our criteria, 18 studies remained for the final review. Their results appear promising for grade prediction from MR images using the radiomics techniques. However, there is no agreement about the radiomics pipeline, and the prior studies are very heterogeneous regarding the software used, the number of extracted features, MR sequences, and machine learning technique. Only two studies have implemented their model on external datasets, likely providing a more realistic estimation of grading prediction by radiomics. All of the studies were retrospective in design, and many of them are based on small size datasets. Before the clinical implementation of glioma grading by radiomics, more standardized research is needed.

Statement of Ethics

The paper is exempt from ethical committee approval because it does not involve human subjects and it is a review article.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

None.

Author Contributions

M.T.: study design, literature review, data gathering, statistical analysis, and manuscript writing. A.R.: literature review, data gathering, and manuscript writing. A.H.S.: literature review, data gathering, and manuscript writing. Z.S.: literature review, data gathering, and manuscript writing. A.S.: manuscript writing. H.S.: study design, literature review, data gathering, statistical analysis, manuscript writing, and manuscript submitting.

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

References

1 Chen R, Smith-Cohn M, Cohen AL, Colman H. Glioma Subclassifications and Their Clinical Significance. Neurotherapeutics. 2017; 14(2):284–97.
2 Lin Z, Yang R, Li K, Yi G, Li Z, Guo J, et al . Establishment of age group classification for risk stratification in glioma patients. BMC Neurol. 2020;20(1):310.
3 Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131(6):803–20.
4 Louis DN, Giannini C, Capper D, Paulus W, Figarella-Branger D, Lopes MB, et al. cIMPACT-NOW update 2: diagnostic clarifications for diffuse midline glioma, H3 K27M-mutant and diffuse astrocytoma/anaplastic astrocytoma, IDH-mutant. Acta Neuropathol. 2018;135(4):639–42.
5 Brat DJ, Aldape K, Colman H, Holland EC, Louis DN, Jenkins RB, et al. cIMPACT-NOW update 3: recommended diagnostic criteria for "Diffuse astrocytic glioma, IDH-wildtype, with molecular features of glioblastoma, WHO grade IV". Acta Neuropathol. 2018; 136(5):805–10.
6 Ellison DW, Hawkins C, Jones DTW, Onar-Thomas A, Pfister SM, Reifenberger G, et al. cIMPACT-NOW update 4: diffuse gliomas characterized by MYB, MYBL1, or FGFR1 alterations or BRAF(V600E) mutation. Acta Neuropathol. 2019;137:683–7.
7 Brat DJ, Aldape K, Colman H, Figrarella-Branger D, Fuller GN, Giannini C, et al. cIMPACT-NOW update 5: recommended grading criteria and terminologies for IDH-mutant astrocytomas. Acta Neuropathol. 2020; 139(3):603–8.
8 Ellison DW, Aldape KD, Capper D, Fouladi M, Gilbert MR, Gilbertson RJ, et al. cIMPACT-NOW update 7: advancing the molecular classification of ependymal tumors. Brain Pathol. 2020;30(5):863–6.
9 Thust SC, Heiland S, Falini A, Jager HR, Waldman AD, Sundgren PC, et al. Glioma imaging in Europe: A survey of 220 centres and recommendations for best clinical practice. Eur Radiol. 2018;28:3306–17.
10 Gui C, Lau JC, Kosteniuk SE, Lee DH, Megyesi JF. Radiology reporting of low-grade glioma growth underestimates tumor expansion. Acta Neurochir (Wien). 2019;161(3):569–76.
11 Bink A, Benner J, Reinhardt J, De Vere-Tyndall A, Stieltjes B, Hainc N, et al. Structured Reporting in Neuroradiology: Intracranial Tumors. Front Neurol. 2018;9:32.
12 Kong Z, Lin Y, Jiang C, Li L, Liu Z, Wang Y, et al. (18)F-FDG-PET-based Radiomics signature predicts MGMT promoter methylation status in primary diffuse glioma. Cancer Imaging. 2019;19:58.
13 Lohmann P, Galldiks N, Kocher M, Heinzel A, Filss CP, Stegmayr C, et al. Radiomics in neuro-oncology: Basics, workflow, and applications. Methods. 2020.
14 Tagliafico AS, Piana M, Schenone D, Lai R, Massone AM, Houssami N. Overview of radiomics in breast cancer diagnosis and prognostication. Breast. 2020;49:74–80.
15 Schick U, Lucia F, Bourbonne V, Dissaux G, Pradier O, Jaouen V, et al. Use of radiomics in the radiation oncology setting: Where do we stand and what do we need? Cancer Radiother. 2020;24(6–7):755–61.
16 Liu L, Yi X, Lu C, Qi L, Zhang Y, Li M, et al. Applications of radiomics in genitourinary tumors. Am J Cancer Res. 2020;10(8):2293–308.
17 Cho HH, Lee SH, Kim J, Park H. Classification of the glioma grading using radiomics analysis. PeerJ. 2018;6:e5982.
18 Conti A, Duggento A, Indovina I, Guerrisi M, Toschi N. Radiomics in breast cancer classification and prediction. Semin Cancer Biol. 2020.
19 Gu H, Zhang X, di Russo P, Zhao X, Xu T. The Current State of Radiomics for Meningiomas: Promises and Challenges. Front Oncol. 2020; 10:567736.
20 Machicado JD, Koay EJ, Krishna SG. Radiomics for the Diagnosis and Differentiation of Pancreatic Cystic Lesions. Diagnostics (Basel). 2020;10.
21 Delgadillo R, Ford JC, Abramowitz MC, Dal Pra A, Pollack A, Stoyanova R. The role of radiomics in prostate cancer radiotherapy. Strahlenther Onkol. 2020;196(10):900–12.
22 Lu CF, Hsu FT, Hsieh KL, Kao YJ, Cheng SJ, Hsu JB, et al. Machine Learning-Based Radiomics for Molecular Subtyping of Gliomas. Clin Cancer Res. 2018;24(18):4429–36.
23 Kocher M, Ruge MI, Galldiks N, Lohmann P. Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. Strahlenther Onkol. 2020;196(10):856–67.
24 Chaddad A, Kucharczyk MJ, Daniel P, Sabri S, Jean-Claude BJ, Niazi T, et al. Radiomics in Glioblastoma: Current Status and Challenges Facing Clinical Implementation. Front Oncol. 2019;9:374.
25 Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to Radiomics. J Nucl Med. 2020; 61(4):488–95.
26 Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging. 2019 Dec;46(13):2656–72.
27 Cui G, Jeong J, Lei Y, Wang T, Dong X, Liu T, et al., editors. Machine-Learning-Based Classification of Low-Grade and High-Grade Glioblastoma Using Radiomic Features in Multiparametric MRI. Medical Physics; 2018: Wiley, Hoboken, NJ USA.
28 Chen Q, Wang L, Wang L, Deng Z, Zhang J, Zhu Y. Glioma Grade Prediction Using Wavelet Scattering-Based Radiomics. IEEE Access. 2020;8:106564–75.
29 Xiao T, Hua W, Li C, Wang S, editors. Glioma Grading Prediction by Exploring Radiomics and Deep Learning Features. Proceedings of the Third International Symposium on Image Computing and Digital Medicine; 2019.
30 Wu Y, Liu B, Wu W, Lin Y, Yang C, Wang M. Grading glioma by radiomics with feature selection based on mutual information. J Ambient Intell Human Comput. 2018;9(5):1671–82.
31 Vamvakas A, Williams SC, Theodorou K, Kapsalaki E, Fountas K, Kappas C, et al. Imaging biomarker analysis of advanced multiparametric MRI for glioma grading. Phys Med. 2019;60:188–98.
32 Lu CF, Hsu FT, Hsieh KL, Kao YJ, Cheng SJ, Hsu JB, et al. Machine Learning-Based Radiomics for Molecular Subtyping of Gliomas. Clin Cancer Res. 2018;24(18):4429–36.
33 Hashido T, Saito S, Ishida T. A radiomics-based comparative study on arterial spin labeling and dynamic susceptibility contrast perfusion-weighted imaging in gliomas. Sci Rep. 2020;10(1):6121–10.
34 Cho HH, Lee SH, Kim J, Park H. Classification of the glioma grading using radiomics analysis. PeerJ. 2018;6:e5982.
35 Lin Y, Wu Y, Pang H, Wu W, Liu T, Wang M, editors. A precise grading method for glioma based on radiomics. 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI); 2017: IEEE.
36 Cho H-h, Park H, editors. Classification of low-grade and high-grade glioma using multi-modal image radiomics features. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2017: IEEE.
37 Takahashi S, Takahashi W, Tanaka S, Haga A, Nakamoto T, Suzuki Y, et al. Radiomics analysis for glioma malignancy evaluation using diffusion kurtosis and tensor imaging. Int J Radiat Oncol Biol Phys. 2019 Nov 15;105(4): 784–91.
38 Park YW, Choi YS, Ahn SS, Chang JH, Kim SH, Lee SK. Radiomics MRI phenotyping with machine learning to predict the grade of lower-grade gliomas: a study focused on nonenhancing tumors. Korean J Radiol. 2019; 20(9):1381–9.
39 Wang Q, Li Q, Mi R, Ye H, Zhang H, Chen B, et al. Radiomics nomogram building from multiparametric MRI to predict grade in patients with glioma: a cohort study. J Magn Reson Imaging. 2019;49(3):825–33.
40 Tian Q, Yan LF, Zhang X, Zhang X, Hu YC, Han Y, et al. Radiomics strategy for glioma grading using texture features from multiparametric MRI. J Magn Reson Imaging. 2018; 48(6):1518–28.

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM

41 Wu Y, Liu B, Zhao G, Yang H, Chen Y, Lv Q, et al., editors. Robust Feature Selection Method of Radiomics for Grading Glioma. The International Conference on Healthcare Science and Engineering; 2018: Springer.
42 Gao M, Huang S, Pan X, Liao X, Yang R, Liu J. Machine Learning-Based Radiomics Predicting Tumor Grades and Expression of Multiple Pathologic Biomarkers in Gliomas. Front Oncol. 2020;10.
43 Çinarer G, Emiroğlu BG, Yurttakal AH. Prediction of Glioma Grades Using Deep Learning with Wavelet Radiomic Features. Applied Sciences. 2020;10(18):6296.
44 Bi X, Liu JG, Cao YS, editors. Classification of Low-grade and High-grade Glioma using Multiparametric Radiomics Model. 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC); 2019: IEEE.
45 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500–10.
46 Aneja S, Chang E, Omuro A. Applications of artificial intelligence in neuro-oncology. Curr Opin Neurol. 2019;32(6):850–6.
47 Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26(6):1045–57.
48 Scarpace L, Mikkelsen T, Cha S. Radiology Data from The Cancer Genome Atlas Glioblastoma Multiforme [TCGA-GBM] collection (2016). URL
49 Pedano N, Flanders AE, Scarpace L, Mikkelsen T, Eschbacher J, Hermes B, et al. Radiology data from the cancer genome atlas low grade glioma [TCGA-LGG] collection. The Cancer Imaging Archive. 2016;2.
50 Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin. 2019;69(2):127–57.
51 Lin WJ, Chen JJ. Class-imbalanced classifiers for high-dimensional data. Brief Bioinformatics. 2013;14(1):13–26.
52 Louis DN, Schiff D, Batchelor T, Wen PY. Classification and pathologic diagnosis of gliomas. UpToDate, Waltham, MA: Walters Kluwer Health. 2020.
53 Rudie JD, Rauschecker AM, Bryan RN, Davatzikos C, Mohan S. Emerging Applications of Artificial Intelligence in Neuro-Oncology. Radiology. 2019;290(3):607–18.
54 Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, et al. Radiomics and radiogenomics in lung cancer: A review for the clinician. Lung Cancer. 2018;115:34–41.

Downloaded by:
University of Edinburgh
129.215.16.31 - 8/11/2021 7:57:33 PM