European Radiology (2022) 32:8039–8051
https://doi.org/10.1007/s00330-022-08828-x

Radiomics can differentiate high-grade glioma from brain metastasis: a systematic review and meta-analysis

Yuanzhen Li[1] · Yujie Liu[1] · Yingying Liang[2] · Ruili Wei[1,2] · Wanli Zhang[1,2] · Wang Yao[1,2] · Shiwei Luo[1,2] · Xinrui Pang[1,2] · Ye Wang[1,2] · Xinqing Jiang[2] · Shengsheng Lai[3] · Ruimeng Yang[2]

Received: 10 December 2021 / Revised: 5 April 2022 / Accepted: 18 April 2022 / Published online: 19 May 2022
© The Author(s), under exclusive licence to European Society of Radiology 2022

Abstract

Objective (1) To evaluate the diagnostic performance of radiomics in differentiating high-grade glioma from brain metastasis and how to improve the model. (2) To assess the methodological quality of radiomics studies and explore ways of embracing the clinical application of radiomics.

Methods Studies using radiomics to differentiate high-grade glioma from brain metastasis published by 26 July 2021 were systematically reviewed. Methodological quality and risk of bias were assessed using the Radiomics Quality Score (RQS) system and Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool, respectively. Pooled sensitivity and specificity of the radiomics model were also calculated.

Results Seventeen studies combining 1,717 patients were included in the systematic review, of which 10 studies without data leakage suspicion were employed for the quantitative statistical analysis. The average RQS was 5.13 (14.25% of total), with substantial or almost perfect inter-rater agreements. The inclusion of clinical features in the radiomics model was only reported in one study, as was the case for publicly available algorithm code. The pooled sensitivity and specificity were 84% (95% CI, 80–88%) and 84% (95% CI, 81–87%), respectively. The performances of feature extraction from the volume of interest (VOI) or (semi) automatic segmentation in the radiomics models were superior to those of protocols employing region of interest (ROI) or manual segmentation.

Conclusion Radiomics can accurately differentiate high-grade glioma from brain metastasis. The adoption of standardized workflow to avoid potential data leakage as well as the integration of clinical features and radiomics are advised to consider in future studies.

Key Points

- *The pooled sensitivity and specificity of radiomics for differentiating high-grade gliomas from brain metastasis were 84% and 84%, respectively.*
- *Avoiding potential data leakage by adopting an intensive and standardized workflow is essential to improve the quality and generalizability of the radiomics model.*
- *The application of radiomics in combination with clinical features in differentiating high-grade gliomas from brain metastasis needs further validation.*

Keywords Glioma · Brain neoplasms · Artificial intelligence · Magnetic resonance imaging · Quality improvement

✉ Shengsheng Lai
laiss@gdyzy.edu.cn

✉ Ruimeng Yang
eyruimengyang@scut.edu.cn

[1] School of Medicine, South China University of Technology, Guangzhou 510006, People's Republic of China

[2] Department of Radiology, the Second Affiliated Hospital, School of Medicine, South China University of Technology, Guangzhou 510180, Guangdong, People's Republic of China

[3] School of Medical Equipment, Guangdong Food and Drug Vocational College, Guangzhou 510520, Guangdong, People's Republic of China

Abbreviations

AI	Artificial intelligence
AUC	Area under the curve
cML	Conventional machine learning
CT	Computed tomography
DL	Deep learning
DTI	Diffusion tensor imaging
DWI	Diffusion-weighted imaging
ICC	Interclass correlation coefficient
MRI	Magnetic resonance imaging
PRISMA-DTA	Preferred Reporting Items for Systematic Reviews and Meta-analysis for Diagnostic Test Accuracy
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies-2
ROI	Region of interest
RQS	Radiomics Quality Score
SWI	Susceptibility-weighted imaging
VOI	Volume of interest

Introduction

High-grade glioma (World Health Organization (WHO) grade III and IV) and brain metastasis are common malignant brain tumors in adults [1]. Accurate noninvasive differentiation of the two malignancies is crucial for the clinical management and treatment options for patients [2, 3]. Combining clinical history, CT/MRI can improve the diagnostic accuracy of primary and metastatic tumors. However, given the similarity in radiological appearances for necrotic centers, irregular enhancing margins, and peritumoral edema, high-grade glioma, and brain metastasis cannot be accurately differentiated [4].

Radiomics is an emerging method that utilizes the intensity, shape, and texture of medical images invisible to the naked eyes to recognize tumors [5]. It follows some standard steps: (1) image acquisition and segmentation, (2) feature extraction and selection, and (3) modeling and validation [6]. However, the quality, applicability, and diagnostic performance of the method are influenced by methodological modifications at each step such as the type of feature [7] and the placement of the region of interest (ROI) [8] influences the performance of radiomics. At present, there is no consensus on the most accurate or appropriate radiomics model for differentiating high-grade glioma from brain metastasis.

In the present study, we systematically analyzed existing data based on trial radiomics models used in differentiating high-grade glioma from brain metastasis. We aimed to optimize the diagnostic performance and clinical applicability of the radiomics methods. The pooled diagnostic performance of radiomics was analyzed using a quantitative statistical meta-analysis.

Materials and methods

This systematic review was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-analysis for Diagnostic Test Accuracy (PRISMA-DTA) guidelines [9]. Our protocol is available at PROSPERO (CRD42021269218). Literature search, data extraction, and quality assessment were performed independently by two reviewers (Y. Li and Y. Liu).

Literature search

Relevant articles were systematically searched through PubMed, EMBASE, and Cochran Library. Only original English articles published by up to 26 July 2021 were considered. To avoid erroneous omissions, all articles on differentiation of high-grade glioma from brain metastasis using MR imaging were retrieved. Selection of those on the use of radiomics was then performed manually. The search terms for the relevant articles included ((glioblastoma OR glioma) AND (brain metastases OR brain metastasis OR metastatic brain tumor)) AND (magnetic resonance OR magnetic resonance imaging).

Literature selection

Inclusion criteria

To be included, the articles must have fulfilled the following: (1) population: patients with high-grade gliomas or brain metastasis; (2) index test: quantitative analysis of MR images using radiomics method; (3) outcomes: differentiation of glioma from brain metastasis based on adequate data sufficient to reconstruct the 2 × 2 sensitivity and specificity tables; (4) original articles.

Exclusion criteria

Reviews, guidelines, conference proceedings or case reports/series, articles including low-grade gliomas, papers reporting fewer than 10 patients, and those with overlapping patient populations were excluded from the qualitative systematic review. For articles with overlapping patient cohorts, both were included only if the methodological difference was significant.

For the further quantitative statistical analyses, studies with suspected data leakage (i.e., without explicitly mentioning the use of independent validation/testing, or using the entire dataset in feature selection) were excluded to avoid the potential overestimation.

Data extraction and quality assessment

Following data were extracted from the included studies. (1) Patient characteristics: number of overall cases and patients with glioma or brain metastasis, WHO grade of glioma, mean age, age range, sex ratio of the study participants, and nature of the tumor (primary or metastasis); (2) Study characteristics: authors, institution, duration of patient recruitment, study design (prospective or retrospective), method of patients' recruitment (consecutive or non-consecutive), reference standard, interval between MRI and the reference standard; (3) MRI characteristics: magnetic field strength, vendor, scanner model, and sequence; (4) Radiomics protocols: MRI sequence, ROI/VOI placement, segmentation, feature selection, and artificial intelligence (AI) classifier (if available); and (5) Outcomes: true positive, false positive, false negative, and true negative. If these parameters were not provided, the sensitivity and specificity data from the original article were used for calculations. For articles with multiple sets of experiments (i.e., multiple comparisons), all of the presented groups were included.

The risk of bias and concerns about the applicability of the included studies were assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool [10]. The tool has four domains: "patient selection," "index test," "reference standard," and "flow and timing" each with subsections. Disagreements were arbitrated by a third reviewer (R. Wei).

The methodological quality of the radiomics of the included studies was assessed using the Radiomics Quality Score (RQS) system [6] under the following parameters: image acquisition and segmentation, features extraction and selection, modeling, model validation, and data sharing. The total scores ranging from − 8 to 36 were a product of 16 dimensions (Table 3). In the domains of feature selection and validation, we further assessed when the feature selection was applied in the modeling process, as well as how the data sets were allocated for modeling to identify possible data leakage. Since RQS is a quantitative tool for several ordinal questions, the inter-rater agreement for single items of the RQS was calculated using the Fleiss kappa statistic for ordinal variables [11]. The Kappa result is suggested by Cohen to be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. Inter-rater agreement for the pooled RQS was assessed using the interclass correlation coefficient (ICC).

Statistical analysis

Inter-rater agreement of RQS was calculated using the "IRR" package in R version 4.1.1. Meta-analysis was performed using the Meta-DiSc 1.4 and "MIDAS" modules in Stata 16 software. Most of the included articles used machine learning classifiers that reported results for multiple cohorts, such as training, testing, and validation set; however, we only employed the generalizable data from the validation/testing set. Pooled sensitivity and specificity for homogeneous data were determined using a fixed-effect model. Furthermore, the diagnostic performance of different radiomics protocols was compared according to the following rules: (1) T1WI, T2WI, and T1CE only vs diffusion MRI; (2) inclusion *vs*. exclusion of the peritumoral; (3) feature extraction from ROI *vs*. VOI; (4) segmentation method (manual *vs*. (semi) automatic); (5) presence *vs*. absence of AI classifier. Due to the possible risk of data leakage, pooled sensitivity/specificity of excluded studies with diffusion MRI [7, 8, 12–14] or without AI classifiers [7, 8, 12, 15, 16] is presented here only for reference purposes.

The heterogeneity of the included studies was based on the Cochrane Q-test (p value ≤ 0.05) and I^2 statistic (> 50%) [17]. Spearman coefficient > 0.6 was considered to indicate a considerable threshold effect [18]. Publication bias was analyzed using a funnel plot.

Results

Literature search

The detailed literature search and selection process is shown in Fig. 1. The initial search identified 1,456 articles, of which 390 were duplicates. After screening through titles and abstracts of the remaining 1,066, an additional 1,045 articles were removed. After reading through the remaining 21 full texts, five more articles were removed for the following reasons: (1) two studies contained data insufficient to generate the 2 × 2 tables [19, 20]; (2) in one study, there was no comparison between glioma and metastasis [21]; (3) one study reported overlapping cohort [22]; (4) one study included low-grade glioma [23]. In one article [24], a deep learning model (DL) was independently validated. Considering its difference from conventional machine learning (cML), this article was regarded as two distinct studies (named Bae (cML) and Bae (DL)). For qualitative systematic reviews, 17 studies with a combined total of 1,717 patients were included [4, 7, 8, 12–16, 24–31]. Seven of them were suspected of possible data leakage and were therefore excluded from quantitative statistical analysis (Supplementary Table S1) [7, 8, 12–16]. Ultimately, the results of quantitative statistics were determined by the remaining 10 studies only [4, 24–31].

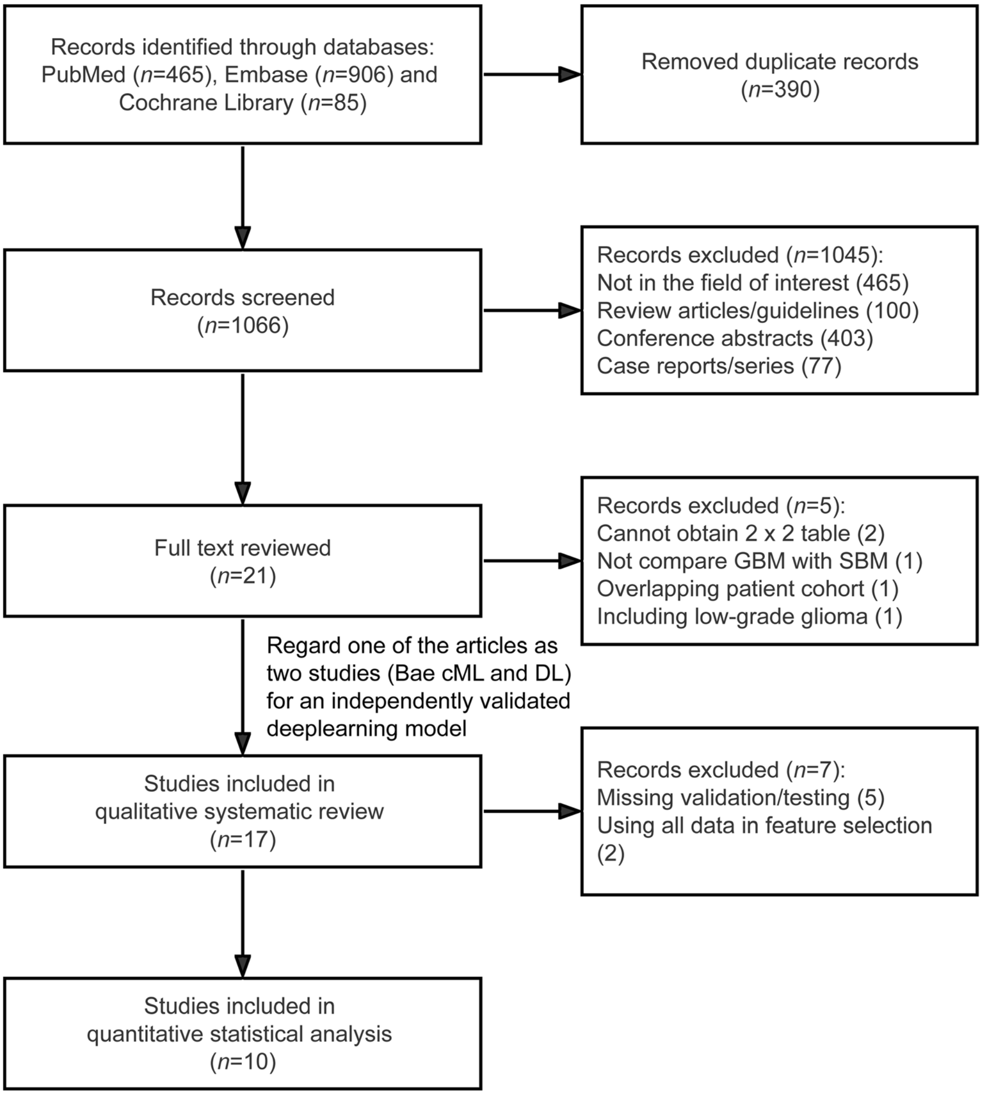

Records identified through databases: PubMed (*n*=465), Embase (*n*=906) and Cochrane Library (*n*=85)
Removed duplicate records (*n*=390)
Records screened (*n*=1066)
Records excluded (*n*=1045):
Not in the field of interest (465)
Review articles/guidelines (100)
Conference abstracts (403)
Case reports/series (77)
Full text reviewed (*n*=21)
Records excluded (*n*=5):
Cannot obtain 2 x 2 table (2)
Not compare GBM with SBM (1)
Overlapping patient cohort (1)
Including low-grade glioma (1)
Regard one of the articles as two studies (Bae cML and DL) for an independently validated deeplearning model
Studies included in qualitative systematic review (*n*=17)
Records excluded (*n*=7):
Missing validation/testing (5)
Using all data in feature selection (2)
Studies included in quantitative statistical analysis (*n*=10)

Fig. 1 A flowchart showing the selection process for included studies

Characteristics of the included studies

The characteristics of the included studies are shown in Table 1. The radiomics protocols are shown in Table 2 while their proportions are presented in Fig. 2C–G. Eight studies only investigated T1WI, T2WI, or T1CE images [4, 15, 24–26, 28, 29], five examined diffusion MRI [7, 8, 12–14], and four combined multiple MRI techniques [16, 27, 30, 31]. In addition, 10 studies only investigated the tumor area [4, 8, 13, 14, 16, 25, 26, 28, 30, 31], two only investigated the peritumor area [12, 15], whereas five studies investigated both [7, 24, 27, 29]. Semi and automatic segmentation was applied in seven studies [4, 7, 13–15, 24, 27], whereas manual segmentation of images was reported in 10 studies [8, 12, 16, 25, 26, 28–31]. Feature selection was performed in nine studies [4, 8, 13, 15, 24–28] to reduce the risk of over-fitting [6]. AI classification was reported in 12 studies [4, 13, 14, 24–31]. The details of the radiomics protocols are shown in Supplementary Table S2.

Quality assessment

QUADAS-2

For the patient selection domain, all studies used a case-control design and the exclusion criteria in two studies [4, 7] were unclear. In the domain of index test, the thresholds of the test indexes were not pre-specified in 11 studies [7, 8, 12–16, 26, 27, 30, 31]. Regarding the reference standards, two studies were unclear [7, 27], whereas the time intervals between MRI and the standard reference were not reported in 14 studies. However, there was no concern regarding applicability considered in the QUADAS-2 tool except for two studies that did not report the reference standards. The most common risk factors for bias factors identified in the included studies are summarized in Fig. 3. The initial flow of assessments of the two reviewers is presented in supplementary Table S3.

Table 1 Characteristics of the included studies. Those for Bae et al cML and DL share the same row due to the same patient cohort. Studies in boldface are included in the quantitative statistical analysis

Author	Year	Institution	Study design	Number of cases (n)	Glioma (n, %)	Histology of glioma (WHO 2016)	Age (Mean ± SD)	Male: female	Reference standard	Magnet strength (T)	Vendor	Scanner
Artzi et al	**2019**	Tel Aviv Sourasky Medical Center, Israel	Retrospective	439	212 (48.3%)	Glioblastoma (grade IV)	60 ± 13	230:209	Histopathology	1.5/3.0	Siemens GE	NA
Bae et al (cML & DL)	**2020**	National Health Insurance Service Ilsan Hospital, Korea	Retrospective	248	159 (64.1%)	Glioblastoma (grade IV)	NA	157:91	Histopathology	3.0	GE Philips	Achieva, Ingenia, Discovery MR750
Liu et al	**2021**	Xiangya Hospital, Central South University, China	Retrospective	268	140 (52.2%)	Glioblastoma (grade IV)	55 ± 9	170:98	Histopathology	3.0	NA	NA
Ortiz-Ramón et al	**2020**	Universitat Politècnica de València, Spain	Retrospective	100	50 (50.0%)	Glioblastoma (grade IV)	61	63:37	Histopathology clinical follow-up	1.5	Philips	Achieva
Priya et al	**2021**	University of Iowa Hospital and Clinics USA	Retrospective	120	60 (50.0%)	Glioblastoma (grade IV)	GBM: 62 ± 11 BM: 62 ± 10	60:60	NA	1.5	Siemens	NA
Qian et al	**2019**	Beijing Tiantan Hospital, Beijing, China	Retrospective	412	242 (58.7%)	Glioblastoma (grade IV)	GBM: 56 ± 15 BM: 56 ± 10	241:170 (1 NA)	Histopathology	3.0	Siemens	Trio
Shin et al	**2021**	Yonsei University College of Medicine, Korea	Retrospective	598	382 (63.9%)	Glioblastoma (grade IV)	57 ± 15	357:241	Histopathology	3.0	Siemens GE Philips	Genia, Achieva, Discovery, MR750, Tim Trio
Swinburne et al	**2019**	Icahn School of Medicine at Mount Sinai, USA	Retrospective	18	9 (50.0%)	Glioblastoma (grade IV)	NA	NA	Histopathology	3.0	NA	NA
Tateishi et al	**2020**	Kumamoto University, Japan	Retrospective	126	73 (57.9%)	Glioblastoma (grade IV)	NA	68:58	Histopathology	3.0	Siemens	Magnetom Trio
Csutak et al	2020	County Emergency Hospital, Cluj-Napoca, Romania	Retrospective	42	16 (38.1%)	High-grade glioma (III = 5, IV = 11)	63 ± 12	25:17	Histopathology	1.5	GE	Signa explorer
Petrujkić et al	2019	Centre for Radiology and Magnetic Resonance, Serbia	Retrospective	55	30 (54.5%)	Glioblastoma (grade IV)	NA	32:23	Histopathology	3.0	Siemens	Magnetom Skyra
Skogen et al	2019	Oslo University Hospitals, Ullevål, Norway	Retrospective	43	22 (51.2%)	Glioblastoma (grade IV)	GBM: 58 BM: 63	NA	Histopathology	3.0	GE	Signal HDx
Vamvakas et al	2018	University of Patras, Greece	Retrospective	20	10 (50.0%)	Glioblastoma (grade IV)	NA	NA	NA	3.0	GE	Signa HDx
Yang et al	2014	University of London, UK	Retrospective	48	30 (62.5%)	Glioblastoma (grade IV)	NA	NA	Histopathology	1.5	GE	Signa LX, Signa HD
Yang et al	2016	University of London, UK	Retrospective	48	30 (62.5%)	Glioblastoma (grade IV)	NA	NA	Histopathology	1.5	GE	Signa LX, Signa HD
Zhang et al	2019	Guangzhou First People's Hospital, Guangzhou, China	Retrospective	62	36 (58.1%)	Glioblastoma (grade IV)	GBM: 45 ± 17 BM: 56 ± 9	33:29	Histopathology	3.0	Siemens	Verio

BM brain metastasis, *DL* deep learning, *GBM* glioblastoma multiforme, *GE* General Electric, c*ML* conventional machine learning, *NA* not available, *SD* standard deviation, *WHO* World Health Organization

Table 2 Radiomics protocols which yielded the highest AUC in each study. For studies that did not provide results of feature selection (marked with *), all extracted features are presented in the table instead of the portion actually used in the model. Studies marked in boldface are included in the quantitative statistical analysis

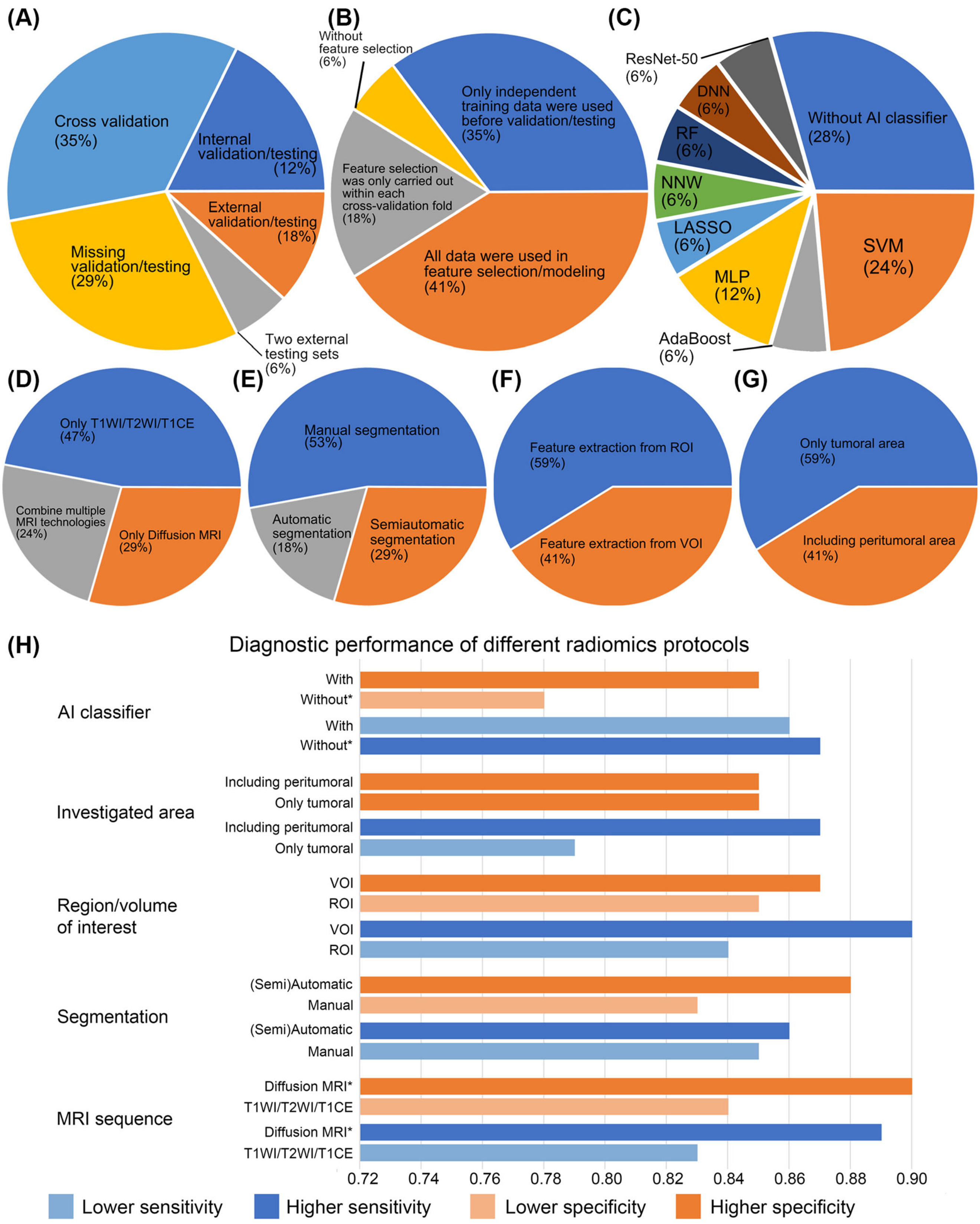
Study ID	Image acquisition and segmentation			Feature extraction and selection		Modeling
	Investigated sequence(s)	ROI/VOI	Segmentation	Amount of features (selected /extracted)	Selected or extracted features	Classifier
Artzi 2019*	T1CE	ROI: tumoral	Automatic	15/760	Clinical, location, first-order, GLCM, GLRLM, wavelet, morphological	SVM
Bae 2020 (cML)*	T2WI, T1CE	ROI: tumoral and peritumoral	Semiautomatic	97/265	Shape, first-order, GLCM, GLRLM, GLSZM, NGTDM	AdaBoost
Bae 2020 (DL)*	T2WI, T1CE	ROI: tumoral and peritumoral	Semiautomatic	NA	NA	Deep Learning (DNN)
Liu 2021*	T1WI, T2WI, T1CE	VOI: tumoral	Manual	259/2106	Shape, first-order, GLCM, GLDM, GLRLM, GLSZM, NGTDM, wavelet, LoG, deep learning-based	RF
Ortiz-Ramón 2020*	T1CE	ROI: tumoral	Manual	82/88	First-order, GLCM, GLRLM, GLSZM, NGTDM, local binary patterns	MLP
Priya 2021*	T1WI, T2WI, T1CE, FLAIR, DWI	VOI: tumoral and peritumoral	Semiautomatic	NA/1070	Shape, first-order, GLCM, GLDM, GLRLM, GLSZM, NGTDM	LASSO
Qian 2019	T1CE	ROI: tumoral	Manual	24/1303	Shape, texture, square, square root, logarithm, exponential, LoG, wavelet	SVM
Shin 2021*	T2WI, T1CE	VOI: tumoral and peritumoral	Manual	NA	NA	Deep Learning (ResNet-50)
Swinburne 2019*	DSC, DCE, DWI	VOI: tumoral	Manual	NA/NA	NA	MLP
Tateishi 2020	T2WI, TICE, DWI	ROI: tumoral	Manual	12/12	First-order, GLCM	SVM
Csutak 2020	T2WI	ROI: peritumoral	Semiautomatic	10/hundreds	First-order, GLRLM, wavelet	NA
Petrujkić 2019	T2WI, T1CE, SWI	ROI: tumoral	Manual	NA/24	GLCM	NA
Skogen 2019	DTI	ROI: peritumoral	Manual	1/30	First-order	NA
Vamvakas 2018	DTI	VOI: tumoral and peritumoral	Semiautomatic	5/32	3D curvature	NA
Yang 2014	DTI	VOI: tumoral	Automatic	5/61	Shape	NNW
Yang 2016	DTI	VOI: tumoral	Automatic	NA/6	Shape	SVM
Zhang 2019	DWI	ROI: tumoral	Manual	2/11	GLCM	NA

AUC area under curve, *cML* conventional machine learning, *DCE* dynamic contrast enhanced, *DL* deep Learning, *DNN* deep neural networks, *DSC* dynamic susceptibility contrast, *DTI* diffusion tensor imaging, *DWI* diffusion-weighted imaging, *FLAIR* fluid-attenuated inversion recovery, *GLCM* gray level co-occurrence matrix, *GLDM* gray level dependence matrix, *GLRLM* gray level run length matrix, *GLSZM* gray level size zone matrix, *LASSO* Least Absolute Shrinkage and Selection Operator, *LoG* Laplacian of Gaussian, *MLP* multi-layer perceptron, *NA* not available, *NGTDM* neighboring gray tone difference matrix, *NNW* neural networks, *RF* random forest, *ROI* region of interest, *SVM* support vector machine, *SWI* susceptibility weighted imaging, *T1CE* T1 contrast enhancement, *T1WI* T1 weighted image, *T2WI* T2 weighted image, *VOI* volume of interest

RQS

The RQS results of the included studies were as follows: mean ± standard deviation −5.13 ± 5.82, median −4.25, and the total scores ranged from −3 to 15 (0–42%). The average rating and inter-rater agreement test results for each item are shown in Table 3, whereas Table S4 presents the individual and summarized scores for each study. In most studies, multiple segmentation, feature selection, discrimination statistics, and biological correlation were performed. The potential clinical utility was also explored. In contrast, no studies

Fig. 2 Visual summary of the studies. **A**, Proportion of validation/testing. ▶ **B**, Proportion of potential sources of data leakage. **C**, Proportion of machine-learning methods used. **D**, Proportion of MRI modalities used. **E**, Proportion of segmentation methods. **F**, Proportion of ROI and VOI. **G**, Proportion of investigated areas. **H**, Diagnostic performance of different radiomics strategies, the sensitivities/specificities marked with * were regarded as studies with suspected data leakage. AI: artificial intelligence, DNN: deep neural networks, LASSO: Least Absolute Shrinkage and Selection Operator, MLP: multi-layer perceptron, NNW: neural networks, RF: random forest, ROI: region of interest, SVM: support vector machine, T1CE: T1 contrast enhancement, T1WI: T1 weighted image, T2WI: T2 weighted image, VOI: volume of interest



explored the stability of radiomics features under different scanners or at different time points. Also, calibration statistics and evaluation of the cost-effectiveness of the protocols were not performed in any of the studies.

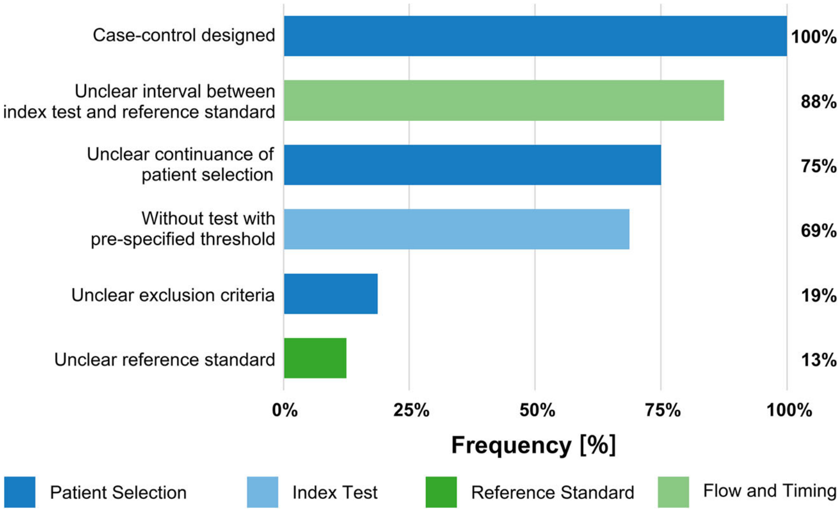


Fig. 3 Causes of risk of bias classified and colored according to the dimensions of the QUADAS-2 tool. The length of the bars represents the frequency with which one cause of the risk of bias was identified in the included studies

Modeling of clinical features of patients was only performed in a single study [4]. MRI protocol was provided in most of the studies. The segmentation method was provided in four studies [4, 13, 14, 27]. The GitHub code was only available in one study [25].

The proportions of validation/testing and potential sources of data leakage are shown in Fig. 2A and B. Six studies used independent validation/testing sets: two used internal sets [4, 28] and four used external sets [24, 25, 29] and reported that only independent training set data were used before validation/testing to avoid potential data leakage. Six studies employed cross-validation [13, 14, 26, 27, 30, 31]: three declared that feature selection was only performed in each fold of cross-validation [26, 27, 30], two used the entire dataset for feature selection [13, 14], and one did not conduct selection. Five studies reported no validation/testing [7, 8, 12, 15, 16]. Studies that did not perform validation/testing [7, 8, 12, 15, 16], as well as those that used the entire dataset in feature selection [13, 14] were excluded from quantitative statistics because of suspected data leakage. The detailed criteria for determining whether studies are with data leakage suspicion are shown in Supplementary Table S1.

Substantial or almost perfect inter-rater agreement was achieved for all individual items of the RQS. The ICC for the total RQS was 0.98 (95% CI 0.93–0.99).

Heterogeneity and diagnostic accuracy

Based on the Cochran's Q tests ($Q = 13.81$ for sensitivity and $Q = 3.10$ for specificity, both $p > 0.05$) and I^2 statistics (I^2 =

Table 3 Ratings of individual RQS items of the included studies and Fleiss's kappa values for the two raters

RQS scoring item	Average (range)	Fleiss's Kappa
Image protocol	0.59 (0–2)	0.73
Multiple segmentations	0.59 (0–1)	0.73
Phantom study on all scanners	0.00 (0–1)	1.00
Imaging at multiple time points	0.00 (0–1)	1.00
Feature reduction or adjustment for multiple testing	0.88 (− 3–3)	0.86
Multivariable analysis with non-radiomics features	0.06 (0–1)	1.00
Detect and discuss biological correlates	0.66 (0–1)	0.86
Cut-off analyses	0.00 (0–1)	1.00
Discrimination statistics	1.72 (0–2)	0.85
Calibration statistics	0.00 (0–2)	1.00
Prospective study registered in a trial database	0.00 (0–7)	1.00
Validation	− 2.53 (− 5–5)	0.87
Comparison to 'gold standard'	0.31 (0–2)	0.76
Potential clinical utility	1.31 (0–2)	0.86
Cost-effectiveness analysis	0.00 (0–1)	1.00
Open science and data	1.41 (0–4)	0.87

RQS radiomics quality score

34.84% for sensitivity and I^2 = 0.00% for specificity), the studies were homogenous. Spearman correlation coefficient was 0.097 (p = 0.789), indicating lack of threshold effect. The pooled sensitivity and specificity of the radiomics based on the 10 studies was 84% (95% CI, 80–88%) and 84% (95% CI, 81–87%) (Fig. 4), respectively. Deeks' funnel plot revealed a low likelihood of publication bias (p = 0.08; Fig. 5).

As shown in Table 4 and Fig. 2H, the higher sensitivity and specificity can be attributed to (1) use of DWI or DTI [7, 8, 12–14], (2) application of (semi) automatic segmentation [4, 24, 27], and (3) features extracted from VOI [25, 27, 29, 30]. However, including the peritumoral area [24, 27, 29] increased the sensitivity of the model but not the specificity. Applying the AI classifier [4, 24–31] improved the specificity but slightly decreased the sensitivity.

Discussion

In the present study, we assessed the diagnostic performance of radiomics in differentiating high-grade gliomas from brain metastasis by systematically reviewing existing studies. To the best of our knowledge, there are only three systematic reviews on this topic that merely evaluated the diagnostic accuracy of manual evaluation of diffusion or perfusion MRI [32–34]. We found that the pooled sensitivity and specificity of the radiomics models were 84% (95% CI, 80–88%) and 84% (95% CI, 81–87%), respectively. In particular, the diagnostic performance of radiomics in analyzing DWI or DTI [7, 8, 12–14] (a pooled sensitivity and specificity of 89% and 90%) was much superior to that of a manual review of DWI or DTI (a pooled sensitivity and specificity of 80% and 81%, respectively) evaluated in another meta-analysis [32].

Radiomics has been used for almost a decade and despite its significant advantages, its clinical application suffers from numerous limitations [35, 36]. Based on the findings of this study, radiomics has the unique advantage of quantitatively combining clinical information and medical imaging. However, the underutilization of this hinders its clinical application. The USA National Institute of Health (NIH) describes radiomics features as a type of "markers" or "characteristics" that are objectively measured and evaluated to reflect the state of a biological or pathogenic process or pharmacologic responses to a therapeutic intervention [37]. Nevertheless, clinical data such as age, sex, and exposure to ionizing radiation

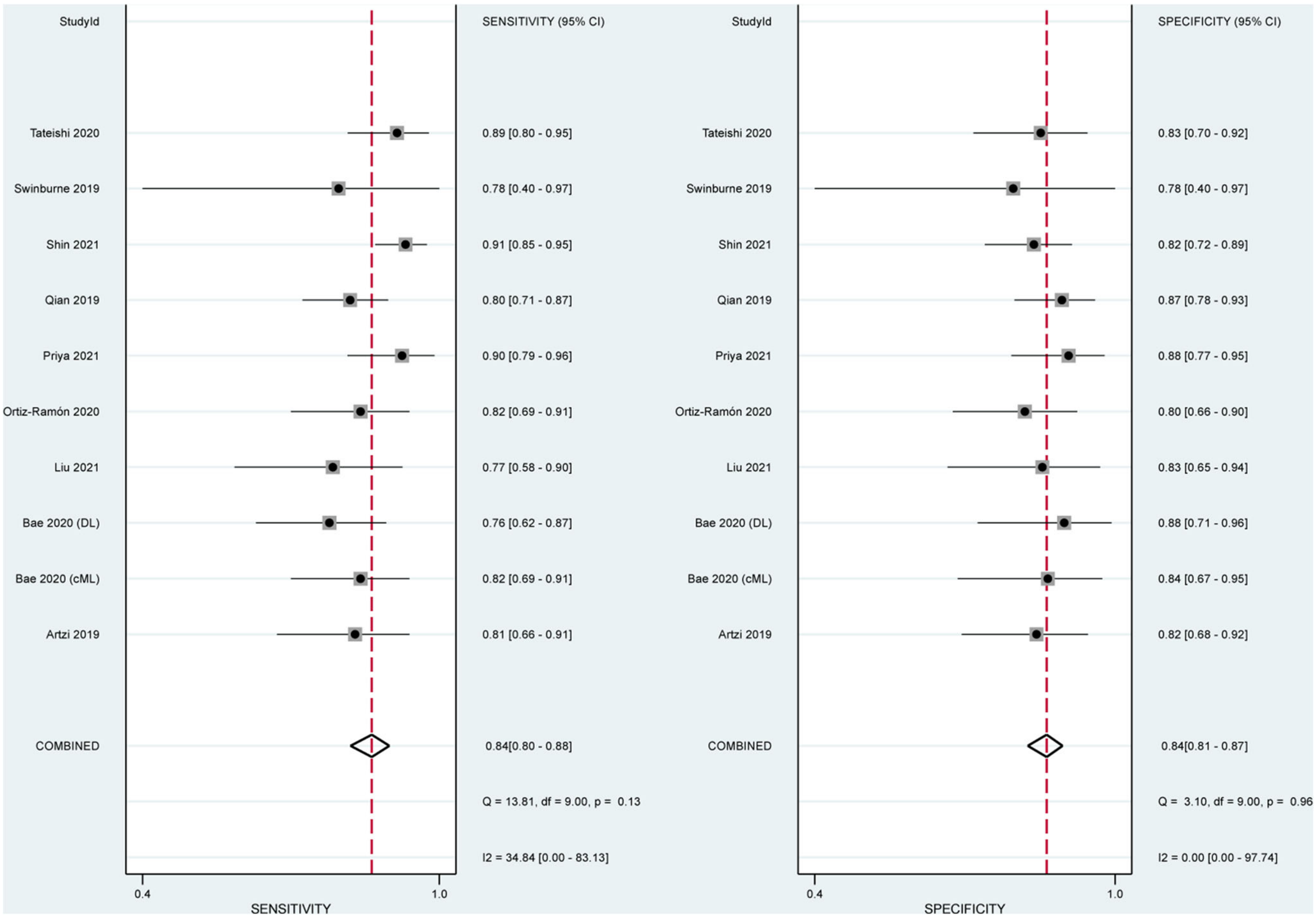


Fig. 4 Coupled Forest plots of pooled sensitivity and specificity

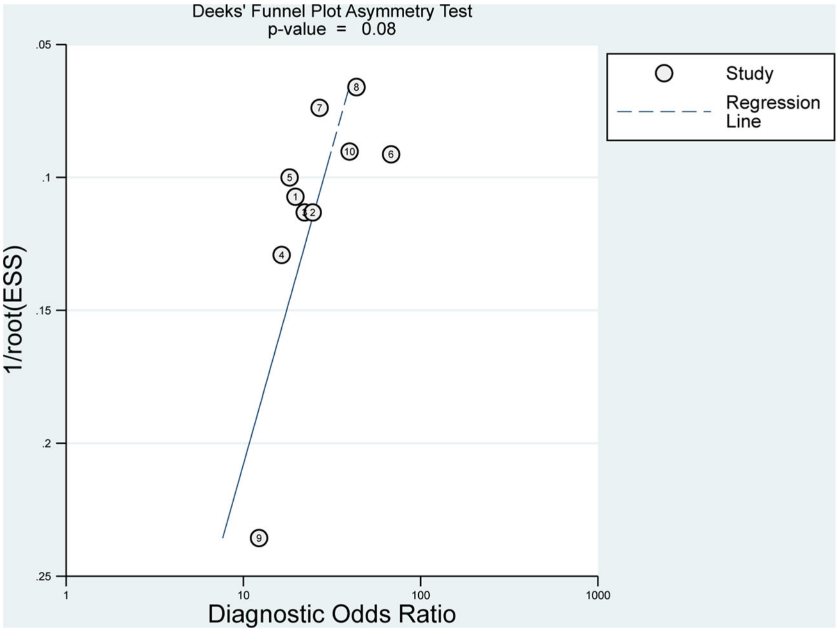


Fig. 5 Deeks' funnel plot. The *p*-value of 0.08 suggests that the likelihood of publication bias is low

are factors closely associated with the pathogenic process of glioma and thus should not be ignored in a diagnostic model [38]. An included study found that clinical features had a high weight in their radiomics model [4]. Halligan et al proposed that clinical imaging biomarkers should be incorporated into a multivariate model to evaluate their additive contribution to the overall outcome [38]. Similarly, one study revealed that clinical features were superior to radiomics in accurately predicting the invasiveness of lung adenocarcinoma (78.4% vs 73.0%). However, the combination of the two was outperformed either technique alone (83.8%). For instance, age is a risk factor for the development of high-grade glioma, but the contribution of age to the development of brain metastasis is not uniform [39, 40]. Age or other numerical clinical data can be readily quantified by radiomics models (i.e., age as a variable in an algorithm or function). However, radiologists can only provide qualitative diagnoses based on a broad "age group." Therefore, future radiomics analyses should

Table 4 Diagnostic performance of studies grouped by radiomics protocols. The sensitivities/specificities of the groups that only investigated diffusion MRI or did not use AI classifiers were regarded as studies with suspected data leakage

Protocol	Number of studies	Pooled diagnostic performance	
		Sensitivity (95% CI)	Specificity (95% CI)
MRI sequence			
Only T1WI/T2WI/T1CE	7	0.83 (0.80–0.86)	0.84 (0.80–0.87)
Only Diffusion MRI	***5***	0.89 (0.83–0.93)	0.90 (0.83–0.95)
Segmentation			
Manual	6	0.85 (0.82–0.88)	0.83 (0.78–0.87)
(Semi)Automatic	4	0.86 (0.82–0.90)	0.88 (0.83–0.92)
Region/volume of interest			
ROI	6	0.84 (0.81–0.87)	0.85 (0.82–0.88)
VOI	4	0.90 (0.86–0.93)	0.87 (0.82–0.91)
Investigated area			
Only tumoral	6	0.79 (0.75–0.84)	0.85 (0.80–0.88)
Including peritumoral	4	0.87 (0.83–0.90)	0.85 (0.80–0.89)
AI classifier			
Without AI classifier	***5***	0.87 (0.79–0.92)	0.78 (0.69–0.85)
With AI classifier	10	0.86 (0.83–0.88)	0.85 (0.82–0.88)

AI artificial intelligence, *CI* confidence interval, *ROI* region of interest, *T1CE* T1 contrast enhancement, *T1WI* T1 weighted image, *T2WI* T2 weighted image, *VOI* volume of interest incorporate clinical features to build more reliable models or add radiomics features to existing diagnostic models to verify their true diagnostic capabilities.

The lack of standardized quality control and reporting throughout the entire workflow limits the application of radiomics. For instance, at each step of a radiomics study, it is essential to keep validation/testing data completely independent or invisible before validation/testing for a generalizable predictive model [41]. Studies using the entire dataset for both model training and evaluation without any validation/testing [7, 8, 12, 15, 16] were with potential biases because they were unable to demonstrate the replicability of the model on unknown data. Two studies [13, 14] used the entire dataset for feature selection before cross-validation, rather than within each fold of the cross-validation. Under such circumstances, the features included in the model are determined by the entire data set (including those used for validation/testing), resulting in an overestimation of the diagnostic performance [41]. In addition, non-standard and incomplete reports, such as the absence of specifically selected features, made it impossible to validate the resultant model. Detailed and accurate reporting is necessary for the application and reproduction of a method. In light of this, we recommend the development of guidelines or checklists for this approach, such as checklists for artificial intelligence in medical imaging (CLAIM) [42] or rules for predictive modeling of individual differences in neuroimaging [41].

The diagnostic performance of the radiomics model varied with the strategies employed. For instance, the model performed better when using diffusion MRI. It provides functional information on the invasion of white matter [43, 44] which can be revealed by radiomics. Second, features from the peritumoral area can be heterogeneous because peritumoral edema of metastases is considered “pure vasogenic”, which is due to parenchymal compression, whereas the edema of high-grade glioma is caused by infiltration of malignant cells [45, 46]. Therefore, peritumoral area features provide more comprehensive information for the radiomics model. Third, even though the inclusion of multiple slices of the tumor, for instance, VOI generates large data, it increases the risk of over-fitting, implying the need for a robust feature selection algorithm. Fourth, automatic segmentation techniques can segregate the tumors based on MRI-derived biomarkers, which may facilitate the building of the radiomics model [47]. Finally, although the AI classifier did not show prominent diagnostic performance in our evaluation, it is undeniably a future research direction and development trend. Most of the included studies employed machine learning or deep learning in feature selection [24–26, 28] or classification [4, 13, 14, 24–31] but the best performing AI classifiers in each study were different. To date, there is no universal classifier because the performance of a classifier can be affected by the characteristics of the sample [48]. Identifying a uniform and robust classifier for specific medical problems has always been challenging [49].

Regarding limitations, first, given that all the studies included in the meta-analysis were retrospective in design, they suffered inherent selection bias and data loss. Therefore, prospective multicenter trials are needed to validate our findings. Second, there was no heterogeneity among the included studies, which generated insignificances in the comparison of different radiomics protocols. Third, given the small sample size, the effects of the various combinations of different parameters captured at each step of radiomics were not explored. Fourth, different radiomics feature was not considered one of the variables in our comparison due to their heterogeneousness. Fifth, the diagnostic performances of radiomics investigating diffusion MRI or without AI classifier were obtained from studies with possible data leakage, relatively lacking reliability and generalizability.

Conclusively, our meta-analysis demonstrated the potential of radiomics in differentiating high-grade gliomas from brain metastasis. How to improve the performance of a radiomics model was also explored. However, more studies are needed to further improve the application and reliability of radiomics.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-08828-x.

Funding This study was funded by (1) the National Natural Science Foundation of China (81971574), (2) the Natural Science Foundation of Guangdong Province in China (2021A1515011350), (3) the Science and Technology Program of Guangzhou in China (202002030268, 201904010422, 202102010025, 202102010031), (4) the GuangDong Basic and Applied Basic Research Foundation (2021A1515220060), (5) the Special Fund for the Construction of High-level Key Clinical Specialty (Medical Imaging) in Guangzhou, (6) Guangzhou Key Laboratory of Molecular Imaging and Clinical Translational Medicine.

Declarations

Guarantor The scientific guarantor of this publication is Dr. Ruimeng Yang and Shengsheng Lai.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was not required for this study because the article type is a systematic review and meta-analysis.

Ethical approval Ethical approval was not required for this study because the article type is a systematic review and meta-analysis.

Methodology
- retrospective
- systemic review and meta-analysis
- performed at one institution

References

1. Louis DN, Perry A, Reifenberger G et al (2016) The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol 131:803–820
2. Giese A, Westphal M (2001) Treatment of malignant glioma: a problem beyond the margins of resection. J Cancer Res Clin Oncol 127:217–225
3. Pruitt AA (2017) Epidemiology, treatment, and complications of central nervous system metastases. Continuum (Minneap Minn) 23: 1580–1600
4. Artzi M, Bressler I, Ben Bashat D (2019) Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis. J Magn Reson Imaging 50:519–528
5. Kuo MD, Jamshidi N (2014) Behind the numbers:decoding molecular phenotypes with radiogenomics–guiding principles and technical considerations. Radiology 270:320–325
6. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762
7. Vamvakas A, Tsougos I, Arikidis N et al (2018) Exploiting morphology and texture of 3D tumor models in DTI for differentiating glioblastoma multiforme from solitary metastasis. Biomed Signal Process Control 43:159–173
8. Zhang G, Chen X, Zhang S et al (2019) Discrimination between solitary brain metastasis and glioblastoma multiforme by using ADC-based texture analysis: a comparison of two different ROI placements. Acad Radiol 26:1466–1472
9. McInnes MDF, Moher D, Thombs BD et al (2018) Preferred Reporting Items for a Systematic Review and Meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 319:388–396
10. Whiting PF, Rutjes AW, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155:529–536
11. Marasini D, Quatto P, Ripamonti E (2016) Assessing the inter-rater agreement for ordinal data through weighted indexes. Stat Methods Med Res 25:2611–2633
12. Skogen K, Schulz A, Helseth E, Ganeshan B, Dormagen JB, Server A (2019) Texture analysis on diffusion tensor imaging: discriminating glioblastoma from single brain metastasis. Acta Radiol 60:356–366
13. Yang G, Jones TL, Barrick TR, Howe FA (2014) Discrimination between glioblastoma multiforme and solitary metastasis using morphological features derived from the p:q tensor decomposition of diffusion tensor imaging. NMR Biomed 27:1103–1111
14. Yang G, Jones TL, Howe FA, Barrick TR (2016) Morphometric model for discrimination between glioblastoma multiforme and solitary metastasis using three-dimensional shape analysis. Magn Reson Med 75:2505–2516
15. Csutak C, Stefan PA, Lenghel LM et al (2020) Differentiating high-grade gliomas from brain metastases at magnetic resonance: the role of texture analysis of the peritumoral zone. Brain Sci 10:638
16. Petrujkic K, Milosevic N, Rajkovic N et al (2019) Computational quantitative MR image features - a potential useful tool in differentiating glioblastoma from solitary brain metastasis. Eur J Radiol 119:108634
17. Higgins JPT, Thomas J, Chandler J et al (2021) Cochrane Handbook for Systematic Reviews of Interventions, Version 6.2. Available via https://training.cochrane.org/handbook/current/chapter-10#section-10-10-2. Accessed 17 October 2021
18. Deville WL, Buntinx F, Bouter LM et al (2002) Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2:9
19. Blanchet L, Krooshof PW, Postma GJ et al (2011) Discrimination between metastasis and glioblastoma multiforme based on morphometric analysis of MR images. AJNR Am J Neuroradiol 32: 67–73
20. Fang K, Wang Z, Li Z et al (2021) Convolutional neural network for accelerating the computation of the extended Tofts model in dynamic contrast-enhanced magnetic resonance imaging. J Magn Reson Imaging 53:1898–1910
21. Sachdeva J, Kumar V, Gupta I, Khandelwal N, Ahuja CK (2013) Segmentation, feature extraction, and multiclass brain tumor classification. J Digit Imaging 26:1141–1150
22. Bathla G, Priya S, Liu Y et al (2021) Radiomics-based differentiation between glioblastoma and primary central nervous system lymphoma: a comparison of diagnostic performance across different MRI sequences and machine learning techniques. Eur Radiol 31: 8703–8713
23. Sartoretti E, Sartoretti T, Wyss M et al (2021) Amide proton transfer weighted (APTw) imaging based radiomics allows for the differentiation of gliomas from metastases. Sci Rep 11:5506
24. Bae S, An C, Ahn SS et al (2020) Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation. Sci Rep 10:12110
25. Liu Z, Jiang Z, Meng L et al (2021) Handcrafted and deep learning-based radiomic models can distinguish GBM from brain metastasis. J Oncol 2021:5518717
26. Ortiz-Ramon R, Ruiz-Espana S, Molla-Olmos E, Moratal D (2020) Glioblastomas and brain metastases differentiation following an MRI texture analysis-based radiomics approach. Phys Med 76: 44–54
27. Priya S, Liu Y, Ward C et al (2021) Machine learning based differentiation of glioblastoma from brain metastasis using MRI derived radiomics. Sci Rep 11:10478
28. Qian Z, Li Y, Wang Y et al (2019) Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers. Cancer Lett 451:128–135
29. Shin I, Kim H, Ahn SS et al (2021) Development and validation of a deep learning-based model to distinguish glioblastoma from solitary brain metastasis using conventional MR images. AJNR Am J Neuroradiol 42:838–844
30. Swinburne NC, Schefflein J, Sakai Y et al (2019) Machine learning for semi-automated classification of glioblastoma, brain metastasis and central nervous system lymphoma using magnetic resonance advanced imaging. Ann Transl Med 7:232
31. Tateishi M, Nakaura T, Kitajima M et al (2020) An initial experience of machine learning based on multi-sequence texture parameters in magnetic resonance imaging to differentiate glioblastoma from brain metastases. J Neurol Sci 410:116514
32. Jiang R, Du FZ, He C, Gu M, Ke ZW, Li JH (2014) The value of diffusion tensor imaging in differentiating high-grade gliomas from brain metastases: a systematic review and meta-analysis. PLoS One 9:e112550
33. Suh CH, Kim HS, Jung SC, Choi CG, Kim SJ (2018) Perfusion MRI as a diagnostic biomarker for differentiating glioma from brain metastasis: a systematic review and meta-analysis. Eur Radiol 28: 3819–3831
34. Suh CH, Kim HS, Jung SC, Kim SJ (2018) Diffusion-weighted imaging and diffusion tensor imaging for differentiating high-grade glioma from solitary brain metastasis: a systematic review and meta-analysis. AJNR Am J Neuroradiol 39:1208–1214

35. Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441–446
36. Pinto Dos Santos D, Dietzel M, Baessler B (2021) A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol 31:1–4
37. Biomarkers Definitions Working G (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89–95
38. Halligan S, Menu Y, Mallett S (2021) Why did European Radiology reject my radiomic biomarker paper? How to correctly evaluate imaging biomarkers in a clinical setting. Eur Radiol 31: 9361–9368
39. Cagney DN, Martin AM, Catalano PJ et al (2017) Incidence and prognosis of patients with brain metastases at diagnosis of systemic malignancy: a population-based study. Neuro Oncol 19:1511–1521
40. Ostrom QT, Gittleman H, Stetson L, Virk S, Barnholtz-Sloan JS (2018) Epidemiology of Intracranial Gliomas. Prog Neurol Surg 30: 1–11
41. Scheinost D, Noble S, Horien C et al (2019) Ten simple rules for predictive modeling of individual differences in neuroimaging. Neuroimage 193:35–45
42. Mongan J, Moy L, Kahn CE Jr (2020) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2:e200029
43. Lee SK (2012) Diffusion tensor and perfusion imaging of brain tumors in high-field MR imaging. Neuroimaging Clin N Am 22: 123–134, ix
44. Wang W, Steward CE, Desmond PM (2009) Diffusion tensor imaging in glioblastoma multiforme and brain metastases: the role of p, q, L, and fractional anisotropy. AJNR Am J Neuroradiol 30:203–208
45. Caravan I, Ciortea CA, Contis A, Lebovici A (2018) Diagnostic value of apparent diffusion coefficient in differentiating between high-grade gliomas and brain metastases. Acta Radiol 59:599–605
46. Lee EJ, terBrugge K, Mikulis D et al (2011) Diagnostic value of peritumoral minimum apparent diffusion coefficient for differentiation of glioblastoma multiforme from solitary metastatic lesions. AJR Am J Roentgenol 196:71–76
47. Jones TL, Byrnes TJ, Yang G, Howe FA, Bell BA, Barrick TR (2015) Brain tumor classification using the diffusion tensor image segmentation (D-SEG) technique. Neuro Oncol 17:466–476
48. Briand B, Ducharme GR, Parache V, Mercat-Rommens C (2009) A similarity measure to assess the stability of classification trees. Comput Stat Data Anal 53:1208–1217
49. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts H (2015) Machine learning methods for quantitative radiomic biomarkers. Sci Rep 5:13087

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