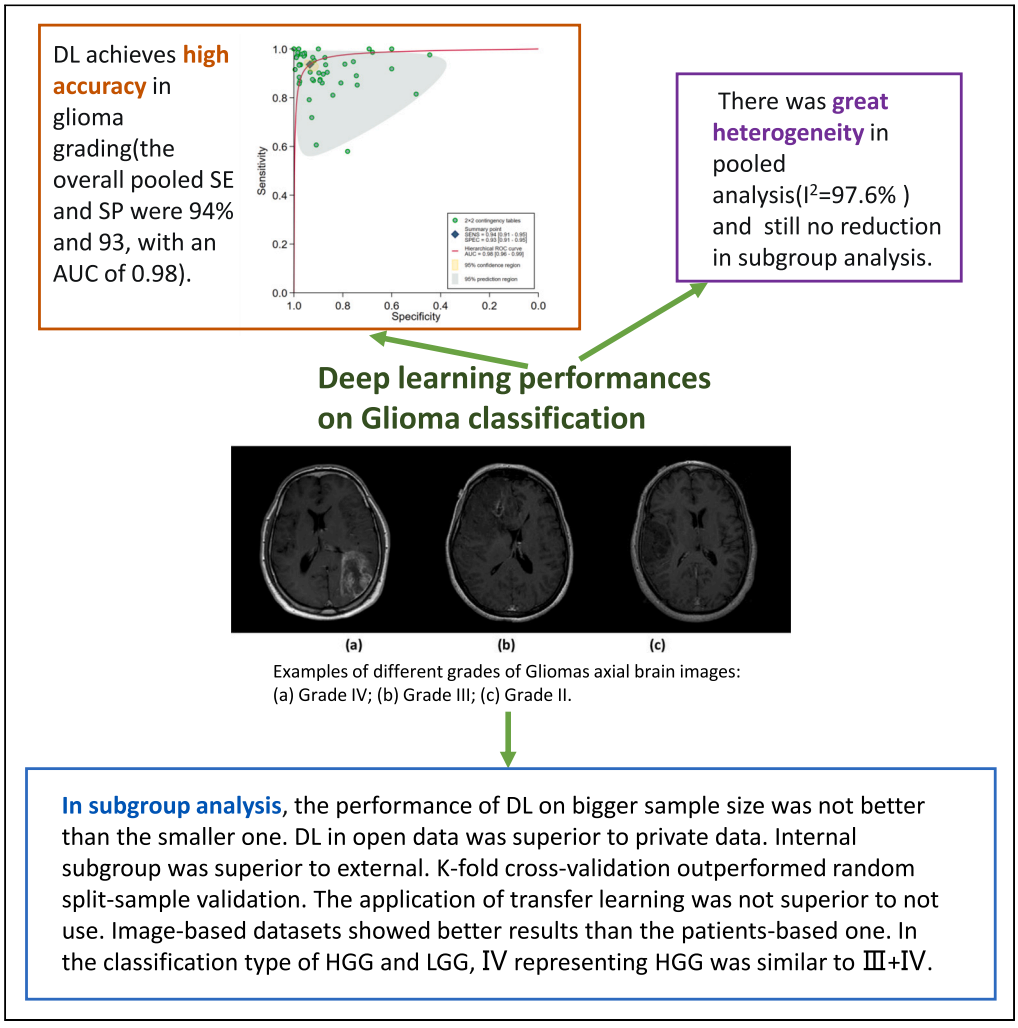
iScience

CellPress
OPEN ACCESS

Article

Performance of deep learning algorithms to distinguish high-grade glioma from low-grade glioma: A systematic review and meta-analysis



Wanyi Sun, Cheng Song, Chao Tang, Chenghao Pan, Peng Xue, Jinhu Fan, Youlin Qiao

xuepeng_pumc@foxmail.com (P.X.)
fanjh@cicams.ac.cn (J.F.)
qiaoy@cicams.ac.cn (Y.Q.)

Highlights

It is aimed to evaluate deep learning performance in glioma classification

We found DL achieved high accuracy in glioma grading

Heterogeneity was found in pooled analysis and still remained in subgroup analysis

In different subgroups, the performance of DL suggested major limitations in this field

Sun et al., iScience 26, 106815
June 16, 2023
https://doi.org/10.1016/j.isci.2023.106815

iScience

CellPress
OPEN ACCESS

Article

Performance of deep learning algorithms to distinguish high-grade glioma from low-grade glioma: A systematic review and meta-analysis

Wanyi Sun,[1,4] Cheng Song,[2,4] Chao Tang,[3] Chenghao Pan,[1] Peng Xue,[2,*] Jinhu Fan,[1,5,*] and Youlin Qiao[2,*]

SUMMARY

This study aims to evaluate deep learning (DL) performance in differentiating low- and high-grade glioma. Search online database for studies continuously published from 1st January 2015 until 16th August 2022. The random-effects model was used for synthesis, based on pooled sensitivity (SE), specificity (SP), and area under the curve (AUC). Heterogeneity was estimated using the Higgins inconsistency index (I^2). 33 were ultimately included in the meta-analysis. The overall pooled SE and SP were 94% and 93%, with an AUC of 0.98. There was great heterogeneity in this field. Our evidence-based study shows DL achieves high accuracy in glioma grading. Subgroup analysis reveals several limitations in this field: 1) Diagnostic trials require standard method for data merging for AI; 2) small sample size; 3) poor-quality image preprocessing; 4) not standard algorithm development; 5) not standard data report; 6) different definition of HGG and LGG; and 7) poor extrapolation.

INTRODUCTION

Glioma originates in the glial cells surrounding and supporting neurons in the brain and is the most common type of malignant brain tumor, representing approximately 80% of all cases.[1] The estimated annual incidence of glioma is in the range of 6 out of 100,000 worldwide.[2] Although relatively rare compared to other malignant tumors, glioblastoma, the most common and deadliest form of glioma, results in a remarkably high mortality rate. The median overall survival is only approximately 19 months regardless of care.[3] The World Health Organization (WHO) categorizes glioma into 4 subtypes—grades I to IV based on their aggressiveness.[4] Clinically, gliomas are normally grouped into low-grade glioma (LGG) and high-grade glioma (HGG).

Accurate categorization of LGG and HGG is indispensable to determining the treatment option and the prognosis of patients. Histopathological characterization following biopsy is a routine procedure to diagnose and grade glioma in clinical practice. However, the procedure is expertise-demanding, workforce-intensive, and time-consuming.[5] To fill in this gap, state-of-the-art medical imaging techniques, especially magnetic resonance imaging (MRI), are widely applied to identify and classify glioma non-invasively, yet both inter- and intraoperator variability cannot be fully avoided. The interpretation of medical images is also highly dependent on the experience and skills of clinicians.

To overcome the aforementioned drawbacks, deep learning (DL), a subset of artificial intelligence (AI), has shown great promise in the automatic classification of medical images.[6,7] For instance, the recent advancement of DL algorithms has rendered Food and Drug Administration (FDA) approves a few diagnosis tools for clinical practice.[8] In our context, numerous independent studies have investigated the performance of DL in glioma classification worldwide. To date; however, there is no systematic review and meta-analysis to assess the diagnostic performance of DL algorithms in grading glioma. This evidence-based study is expected to contribute to the further implementation of DL-based models in routine clinical practice.

RESULTS

Study selection and characteristics

Through literature research, we identified 1178 records. After filtering 166 duplicated records, 1012 records stayed. 901 records were excluded further after a title or abstract scanning, followed by filtering 64 records

[1]Department of Cancer Epidemiology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[3]Shenzhen Maternity & Child Healthcare Hospital, Shenzhen, China

[4]These authors contribute equally

[5]Lead contact

*Correspondence: xuepeng_pumc@foxmail.com (P.X.), fanjh@cicams.ac.cn (J.F.), qiaoy@cicams.ac.cn (Y.Q.)

https://doi.org/10.1016/j.isci.2023.106815

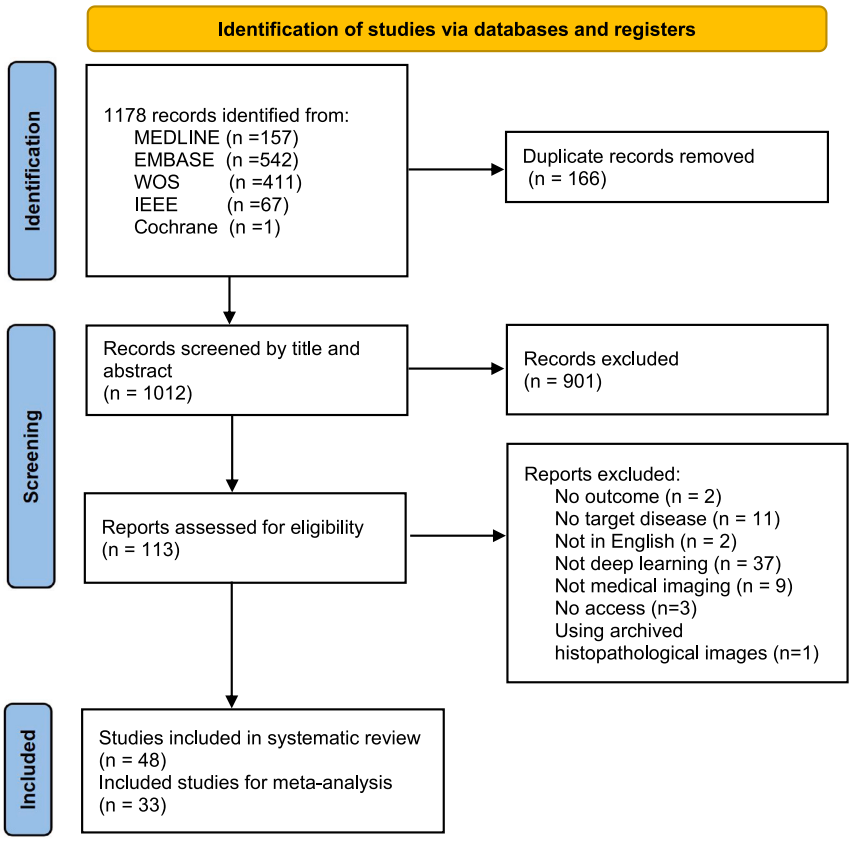


Figure 1. PRISMA flowchart of the study

The literature review and record screening processes followed PRISMA (preferred reporting items for systematic reviews and meta-analyses).

for no outcome, no target disease, no English article, etc. Finally, we included 49 articles that met our inclusion criteria for systematic review, among which 33 articles can fully provide data for meta-analysis (Figure 1).

We totally included 19102 patients. The gold standard was histopathology in all articles. Among the 48 included studies, only 33 studies were included in the meta-analysis due to unextractable or calculation errors in the contingency tables of 15 studies. 52% (17/33) were $\leq$ 130 sample size, 48% (16/33) were > 130, 21% (7/33) were private data, 79% (26/33) were open data, 27% (9/33) were k-fold cross-validation. 73% (24/33) were random split-sample validation, 73% (24/33) were not using the transfer learning, 27% (9/33) were using, 61% were based on image data (20/33), 39% were based on patient data (13/33), 39% (13/33) used grade IV to represent HGG, 61% (20/33) used III+IV, 82% (27/33) were based on internal validation, and 18% (6/33) were based on external (Tables 1, 2, and 3).

Pooled performance of DL algorithms

Among 33 articles with sufficient data, when considering all the records in line with our criteria (including all results in every study), the overall (54 contingency tables) pooled sensitivity (SE) and specificity (SP) were 94% (95% CI: 91–95%) and 93% (95% CI: 91–95%) (Figure 2), with the area under the curve (AUC) of 0.98 (95% CI: 0.96–0.99) for all DL algorithms (Figure 4A).

Considering the problem of reusing samples, we also used the highest accuracy as the criteria to select only one reported performance for each study. The pooled results of highest accuracy for SE and SP were 94% (95% CI: 90–96%) and 94% (95% CI: 90–96%) (Figure 3), with the AUC of 0.98 (95% CI: 0.96–0.99) (Figure 4B).

Table 1. Participant demographics for the 48 included studies (33 included in meta-analysis)

	Participants			
First author and year	Inclusion criteria	Exclusion criteria	Number of patients	Mean or median age (SD; range)
Yu et al. (2022)[9]	(1) Histopathologically confirmed and graded glioma according to the current WHO criteria, (2) images acquired before the operation, (3) data sequences including T1w imaging, T2w imaging, FLAIR imaging, and enhanced T1w imaging.	NR	560	NR(NR; NR)
van der Voort et al. (2022)[10]	Newly diagnosed with a glioma and when preoperative pre- and post-contrast T1w, T2w, and T2w-FLAIR scans were available	The absence of one (or more) of the required scans (T1, post-contrast T1, T2w, T2w-FLAIR)	1748	NR(NR; NR)
Danilov et al. (2022)[11]	NR	NR	707	NR(NR; NR)
Chen et al. (2022)[12]	Patients diagnosed with glioma after case diagnostic screening; patients with complete imaging and follow-up data; and patients with complete follow-up records	Patients with other malignant tumors at the same time; patients with other serious underlying diseases or with dysfunction of important organs such as the heart, lung, liver, and kidney; those who died of diseases or accidents other than glioma; and those who suffered from claustrophobia.	66	53.6(11.3; NR)
Tripathi et al. (2022)[13]	The images which contain tumor region	NR	322	NR(NR; NR)
Xiao et al. (2022)[14]	NR	NR	24	NR(NR; NR)
Wang et al. (2022)[15]	NR	NR	378	NR(NR; NR)
Tasci et al. (2022)	NR	NR	369	NR(NR; NR)
Li et al. (2022)[16]	(1) Pathologically diagnosed as diffuse gliomas; (2) high-quality preoperative T1w, T2w, and T1CE MR images were available; (3) age ≥ 18 years; (4) known IDH status (detected by immunohistochemistry or pyrosequencing); and (5) known 1p19q status (detected using fluorescence *in situ* hybridization) for LGGs.	NR	1016	47(NR; NR)
Khazaee et al. (2022)[17]	NR	NR	335	NR(NR; NR)
Jiang et al. (2021)[18]	NR	NR	620	NR(NR; NR)
He et al. (2021)[19]	NR	NR	499	NR(NR; NR)

(*Continued on next page*)

Table 1. *Continued*

First author and year	Participants			
	Inclusion criteria	Exclusion criteria	Number of patients	Mean or median age (SD; range)
Haq et al. (2021)[20]	NR	NR	351	NR(NR; NR)
Raghavendra et al. (2021)[21]	NR	NR	461	NR(NR; NR)
Chakrabarty et al. (2021)[22]	NR	NR	2105	57(NR; 47–65)
Yahyaoui et al. (2021)[23]	NR	NR	230	NR(NR; NR)
Yamashiro et al. (2021)[24]	NR	NR	285	NR(NR; NR)
Yao et al. (2021)[25]	(1) All cases accepted MRI scan, diagnosed by clinical imaging physicians and neurosurgeons strictly referring to MRI diagnostic criteria. (2) The patients were diagnosed as BG according to post operative pathological results. (3) The patient had no history of craniocerebral surgery and substantial brain injury. (4) Patients had clear consciousness, were able to communicate normally, and had no mental illness	(1) Cases diagnosed as having cerebral infarction, (2) cases with severe communication disorder or mental illness, (3) cases with intracranial hypertension and other characteristics of intracranial lesions, and (4) patients with liver and kidney dysfunction or allergy to contrast agents	60	55.82(4.18; 20–60)
Bezdan et al. (2021)[26]	NR	NR	NR	NR(NR; NR)
Shen et al. (2021)[27]	(1) Male or female; (2) suspected as malignant glioma on preoperative contrast enhancement MRI; (3) voluntarily signed informed consent of surgical treatment and additional specimen beyond what was needed for routine clinical diagnosis; and (4) no contraindication of ICG.	NR	23	NR(NR; NR)
Irmak et al. (2021),[28]	NR	NR	346	NR(NR; NR)
Al-Saffar et al. (2021)[29]	NR	NR	160	NR(NR; NR)
Hu et al. (2021)[30]	The cases in BraTs had clear MR images and tumor masks, and the cases in TCGA and HuaShan had pathological grading and IDH1 information	NR	800	NR(NR; NR)
Luo et al. (2021)[31]	These cases must contain complete imaging data together with histopathology	NR	655	NR(NR; mostly 18–60)

(*Continued on next page*)

Table 1. *Continued*

First author and year	Participants			
	Inclusion criteria	Exclusion criteria	Number of patients	Mean or median age (SD; range)
Decuyper et al. (2021)[32]	A histologically proven glioma of WHO grade II, III or IV, the availability of preoperative T1CE MRI together with a T2 and/or FLAIR sequence of sufficient quality and information on WHO grade, IDH mutation and 1p19q co-deletion status	NR	738	NR(NR; NR)
Gutta et al. (2021)[33]	NR	NR	237	NR(NR; NR)
Ozcan et al. (2021)[34]	NR	NR	104	NR(NR; NR)
Mzoughi et al. (2021)[35]	NR	NR	284	NR(NR; NR)
Koyuncu et al. (2020)[36]	NR	NR	285	NR(NR; NR)
Cinarer et al. (2020)[37]	NR	NR	121	NR(NR; NR)
Mzoughi et al. (2020)[35,38]	NR	NR	351	NR(NR; NR)
Zhuge et al. (2020)[39]	NR	NR	315	NR(NR; NR)
Naser et al. (2020)[40]	NR	NR	110	46(14; 20–75)
Alis et al. (2020)[41]	Diagnosed with whom grade I to IV according to surgical or biopsy-derived histopathological findings; >18 years of age; having preoperative or pre interventional brain MRI with T2W-FLAIR and contrast-enhanced T1W images	Motion or susceptibility artifacts on MRI; history of radio therapy or chemotherapy for prior brain tumor; residual or recurrent brain tumors; gliomas <1 cm in diameter; incomplete clinical data	181	58(NR; 27–78)
Hollon et al. (2020)[42]	(1) Male or female; (2) subjects undergoing CNS tumor resection at Michigan Medicine, New York Presbyterian/Columbia University Medical Center or the University of Miami Health System; (3) subject or durable power of attorney able to give informed consent; and (4) subjects in whom there was additional specimen beyond what was needed for routine clinical diagnosis.	(1) Poor quality of specimen on visual gross examination due to excessive blood, coagulation artifact, necrosis or ultrasonic damage or (2) specimen classified as out of distribution by the linear discriminant analysis classifier using the Mahalanobis distance-based confidence score.	693	NR(NR; NR)
Sharif et al. (2020)[43]	NR	NR	1211	NR(NR; NR)
Lo et al. (2019)[44]	NR	NR	130	NR(NR; NR)
Gonbadi et al. (2019)[45]	NR	NR	285	NR(NR; NR)

(*Continued on next page*)

Table 1. *Continued*

First author and year	Participants			
	Inclusion criteria	Exclusion criteria	Number of patients	Mean or median age (SD; range)
Ali et al. (2019)[46]	NR	NR	285	NR(NR; NR)
Sultan et al. (2019)[28]	NR	NR	73	NR(NR; NR)
Muneer et al. (2019)[47]	NR	NR	20	NR(NR; 30–60)
Anaraki et al. (2019)[48]	NR	NR	688	NR(NR; NR)
Sajjad et al. (2018)[49,50]	NR	NR	NR	NR(NR; NR)
Shahzadi et al. (2018)[49]	NR	NR	60	NR(NR; NR)
Yang et al. (2018)[51]	NR	NR	113	NR(NR; 10–87)
Al-Zurfi et al. (2018)[52]	NR	NR	30	NR(NR; NR)
Ge et al. (2018)[53]	NR	NR	285	NR(NR; NR)
Khawaldeh et al. (2018)[54]	NR	NR	109	NR(NR; 18–89)
Ye et al. (2017)[55]	NR	NR	274	NR(NR; NR)

SD = standard deviation, WHO = world health organization, T1w = T1 weighted, T2w = T2 weighted, FLAIR = fluid attenuated inversion recovery, NR = not reported, T1CE = T1 contrast-enhanced, MR = magnetic resonance, IDH = Isocitrate dehydrogenase, LGG = low-grade glioma, MRI = magnetic resonance imaging, BG = brain glioma, ICG = indocyanine green, CNS = central nervous system.

Table 2. Model training and validation for the 48 included studies(33 included in meta-analysis)

First author and year	Focus	Reference standard	Type of internal validation	External validation	DL versus clinician
Yu et al. (2022)[9]	Brain tumor	Histopathology	Random split-sample validation	No	No
van der Voort et al. (2022)[10]	Brain tumor	Histopathology	Random split-sample validation	Yes	No
Danilov et al. (2022)[11]	Brain tumor	Histopathology	Random split-sample validation	No	No
Chen et al. (2022)[12]	Brain tumor	Histopathology	NR	No	No
Tripathi et al. (2022)[13]	Brain tumor	Histopathology	Random split-sample validation	No	No
Xiao et al. (2022)[14]	Brain tumor	Histopathology	Random split-sample validation	No	No
Wang et al. (2022)[15]	Brain tumor	Histopathology	Random split-sample validation	No	No
Tasci et al. (2022)	Brain tumor	Histopathology	Random split-sample validation	No	No
Li et al. (2022)[16]	Brain tumor	Histopathology	Random split-sample validation	No	No
Khazaee et al. (2022)[17]	Brain tumor	Histopathology	Random split-sample validation	No	No
Jiang et al. (2021)[18]	Brain tumor	Histopathology	Random split-sample validation	No	No
He et al. (2021)[19]	Brain tumor	Histopathology	5-fold cross-validation	Yes	No
Haq et al. (2021)[20]	Brain tumor	Histopathology	Random split-sample validation	No	No
Raghavendra et al. (2021)[21]	Brain tumor	Histopathology	10-fold cross-validation	No	No
Chakrabarty et al. (2021)[22]	Brain tumor	Histopathology	Random split-sample validation	Yes	No
Yahyaoui et al. (2021)[23]	Brain tumor	Histopathology	Random split-sample validation	No	No
Yamashiro et al. (2021)[24]	Brain tumor	Histopathology	Random split-sample validation	No	No
Yao et al. (2021)[25]	Brain tumor	Histopathology	NR	No	No
Bezdan et al. (2021)[26]	Brain tumor	Histopathology	Random split-sample validation	No	No
Shen et al. (2021)[27]	Brain tumor	Histopathology	Random split-sample validation	No	Yes
Irmak et al. (2021)[28]	Brain tumor	Histopathology	5-fold cross-validation	No	No
Al-Saffar et al. (2021)[29]	Brain tumor	Histopathology	Random split-sample validation	No	No
Hu et al. (2021)[30]	Brain tumor	Histopathology	Random split-sample validation	No	No
Luo et al. (2021)[31]	Brain tumor	Histopathology	Random split-sample validation	Yes	No
Decuyper et al. (2021)[32]	Brain tumor	Histopathology	Random split-sample validation	Yes	No
Gutta et al. (2021)[33]	Brain tumor	Histopathology	Random split-sample validation	No	No
Ozcan et al. (2021)[34]	Brain tumor	Histopathology	5-fold cross-validation	No	No
Mzoughi et al. (2021)[35]	Brain tumor	Histopathology	Random split-sample validation	No	No
Koyuncu et al. (2020)[36]	Brain tumor	Histopathology	2-fold cross-validation	No	No
Cinarer et al. (2020)[37]	Brain tumor	Histopathology	Random split-sample validation	No	No
Mzoughi et al. (2020)[38]	Brain tumor	Histopathology	Random split-sample validation	No	No
Zhuge et al. (2020)[39]	Brain tumor	Histopathology	5-fold cross-validation	No	No
Naser et al. (2020)[40]	Brain tumor	Histopathology	5-fold cross-validation	No	No
Alis et al. (2020)[41]	Brain tumor	Histopathology	10-fold cross-validation	No	No
Hollon et al. (2020)[42]	Brain tumor	Histopathology	Random split-sample validation	Yes	Yes
Sharif et al. (2020)[43]	Brain tumor	Histopathology	Random split-sample validation	No	No
Lo et al. (2019)[44]	Brain tumor	Histopathology	10-fold cross-validation	No	No
Gonbadi et al. (2019)[45]	Brain tumor	Histopathology	Random split-sample validation	No	No
Ali et al. (2019)[46]	Brain tumor	Histopathology	Random split-sample validation	No	No
Sultan et al. (2019)[28]	Brain tumor	Histopathology	Random split-sample validation	No	No
Muneer et al. (2019)[47]	Brain tumor	Histopathology	Random split-sample validation	No	No
Anaraki et al. (2019)[48]	Brain tumor	Histopathology	Random split-sample validation	No	No
Sajjad et al. (2018)[50]	Brain tumor	Histopathology	Random split-sample validation	No	No
Shahzadi et al. (2018)[49]	Brain tumor	Histopathology	Random split-sample validation	No	No

(*Continued on next page*)

Table 2. *Continued*

First author and year	Focus	Reference standard	Type of internal validation	External validation	DL versus clinician
Yang et al. (2018)[51]	Brain tumor	Histopathology	5-fold cross-validation	No	No
Al-Zurfi et al. (2018)[52]	Brain tumor	Histopathology	Leave-one-out cross-validation	No	No
Ge et al. (2018)[53]	Brain tumor	Histopathology	Random split-sample validation	No	No
Khawaldeh et al. (2018)[54]	Brain tumor	Histopathology	Random split-sample validation	No	No
Ye et al. (2017)[55]	Brain tumor	Histopathology	Random split-sample validation	No	No

DLdeep learning, NR = not reported.

Heterogeneity analysis

The overall (54 contingency tables) pooled analysis showed I^2 = 97.6% in SE and I^2 = 96.7% in SP. Besides, the highest accuracy pooled analysis indicated I^2 = 98.31% and 96.32% in SE and SP, respectively.

To explore the causes of heterogeneity, we applied meta-regression containing suspectable variables. Including:1) sample size; 2) data sharing; 3) type of internal validation; 4) transfer learning applied; 5) data unit; 6) classification; and 7) type of validation. Among the first 5 variables, data sharing showed no statistical significance (p = 0.39 in SE, p = 0.91 in SP), but the rest 4 variables showed significance at least in one of SE or SP, which indicated heterogeneity. As for the classification and type of validation, both of them had statistical significance in SE and SP, revealing heterogeneity (Table S1).

Subgroup analysis

All variables included in the meta-regression were divided into 2 groups for subgroup analysis.

Sample size: In ≤ 130 subgroup, the SE was 96% (95% CI: 92–98%), SP was 92% (95% CI:85–96%) and AUC was 0.98 (95% CI: 0.97–0.99), while in > 130 subgroup, they were 91% (95% CI: 85–95%), 94% (95% CI: 90–97%), and 0.98 (95% CI: 0.96–0.99), respectively. Heterogeneity still existed in two subgroups (≤ 130: I^2 = 74.08% in SE and 80.42% in SP; > 130: I^2 = 99.05% in SE and 97.73% in SP) (Figures S1 and S18).

Data sharing: In private data subgroup, the SE was 88% (95% CI: 73–95%), SP was 82% (95% CI:74–88%) and AUC was 0.89 (95% CI: 0.86–0.92), while in open data subgroup, they were 95% (95% CI: 92–97%), 96% (95% CI: 92–97%), and 0.99 (95% CI: 0.97–0.99), respectively. Heterogeneity still existed in two subgroups (private data: I^2 = 98.71% in SE and 96.84% in SP; open data: I^2 = 94.31% in SE and 91.89% in SP) (Figures S2 and S9).

Type of internal validation: In k-fold cross-validation subgroup, the SE was 96% (95% CI: 90–99%), SP was 97% (95% CI:89–99%), and AUC was 0.99 (95% CI: 0.98–1.00), while in random split-sample validation subgroup, they were 92% (95% CI: 88–95%), 92% (95% CI: 88–95%), and 0.97 (95% CI: 0.95–0.98), respectively. Heterogeneity still existed in two subgroups (k-fold cross-validation: I^2 = 98.50% in SE and 97.92% in SP; random split-sample validation: I^2 = 98.27% in SE and 95.92% in SP) (Figures S3 and S10).

Transfer learning applied: In no applied subgroup, the SE was 94% (95% CI: 90–97%), SP was 93% (95% CI:88–96%), and AUC was 0.98 (95% CI: 0.96–0.99), while in applied subgroup, they were 93% (95% CI: 85–97%), 95% (95% CI: 93–97%), and 0.98 (95% CI: 0.96–0.99), respectively. Heterogeneity still existed in two subgroups (no applied: I^2 = 98.70% in SE and 97.12% in SP; applied: I^2 = 95.56% in SE and 90.05% in SP) (Figures S4 and S11).

Data unit: In image subgroup, the SE was 95% (95% CI: 90–97%), SP was 96% (95% CI:92–98%), and AUC was 0.99 (95% CI: 0.97–0.99), while in patient subgroup, they were 92% (95% CI: 87–95%), 88% (95% CI: 81–92%), and 0.96 (95% CI: 0.94–0.97), respectively. Heterogeneity still existed in two subgroups (image: I^2 = 99.21% in SE and 98.21% in SP; patient: I^2 = 86.03% in SE and 78.39% in SP) (Figures S5 and S12).

Classification: In grade IV represented HGG subgroup, the SE was 94% (95% CI: 91–96%), SP was 93% (95% CI:89–96%) and AUC was 0.98 (95% CI: 0.96–0.99), while in grade III+IV represented HGG subgroup, they were 93% (95% CI: 89–96%), 94% (95% CI: 89–96%) and 0.98 (95% CI: 0.96–0.99), respectively. Heterogeneity

Table 3. Indicator, algorithm, and data source for the 48 included studies(33 included in meta-analysis)

	Indicator definition			Algorithm		Data source				
							Number of training/ internal/external			
First author and year	Device	Exclusion of poor-quality imaging	Heatmap provided	Algorithm architecture	Transfer learning applied	Source of data	Images	Cases	Data range	Open access data
Yu et al. (2022),[9]	MRI	Yes	No	3D U-Net	No	Retrospective study, data from BraTS 2019 and the PACS system of Henan Provincial People's Hospital	NR/NR/NR	448/112/NR	2012–2020	Yes
van der Voort et al. (2022),[10]	MRI	NR	No	CNN	No	Retrospective study, data rom 4 in-house datasets and 5 publicly available datasets	6032/NR/960	1508/NR/240	NR	Yes
Danilov et al. (2022),[11]	MRI	NR	No	DenseNet, Resnest200e	No	Retrospective study, data from N.N. Burdenko Neurosurgery Center, Russia	15957/1773/NR	636/71/NR	2009–2018	No
Chen et al. (2022)[12]	MRI	NR	Yes	CNN	No	Prospective study, data from The First People's Hospital of Lianyungang	NR/NR/NR	NR/NR/NR	2019.03–2020.03	No
Tripathi et al. (2022),[13]	MRI	Yes	No	Residual networks	Yes	Retrospective study, data from TCIA	6653/739/NR	NR/NR/NR	NR	Yes
Xiao et al. (2022)[14]	Near-infrared fluorescence imaging	NR	Yes	DLS-DARTS	Yes	Prospective study, data from Beijing Tiantan Hospital, Capital Medical University	952//NR	N163R/NR/NR	NR	No
Wang et al. (2022)[15]	MRI	NR	No	3D CNN	No	Retrospective study, data from MICCAI 2020 CPM-Radpath Challenge	NR/NR/NR	305/73/NR	NR	Yes
Tasci et al. (2022)	MRI	NR	No	Xception, IncResNetv2, EfficientNet	Yes	Retrospective study, data from BraTS 2020	17830/4457/NR	NR/NR/NR	NR	Yes
Li et al. (2022),[16]	MRI	Yes	No	2.5D DCNN	No	Retrospective study, data from Beijing Tiantan Hospital	NR/NR/NR	780/236/NR	2014.09–2018.04	No
Khazaee et al. (2022)[17]	MRI	NR	No	EfficientNetB0	No	Retrospective study, data from BraTS 2019	21523/5381/NR	NR/NR/NR	NR	Yes

(*Continued on next page*)

Table 3. *Continued*

First author and year	Indicator definition: Device	Indicator definition: Exclusion of poor-quality imaging	Indicator definition: Heatmap provided	Algorithm: Algorithm architecture	Algorithm: Transfer learning applied	Data source: Source of data	Data source: Number of training/internal/external: Images	Data source: Number of training/internal/external: Cases	Data source: Data range	Data source: Open access data
Jiang et al. (2021)[18]	MRI	Yes	No	SE-ResNeXt	Yes	Retrospective study, data from BraTS 2017 and 2019	21700/9300/NR	NR/NR/NR	NR	Yes
He et al. (2021),[19]	MRI	NR	No	HOMIF	No	Retrospective study, data from TCIA and BraTS 2017	NR/NR/NR	172/42/166 and 228/57/95	NR	Yes
Haq et al. (2021),[20]	MRI	Yes	No	GoogleNet	No	Retrospective study, data from BraTS 2018	$20^{1}/_{8}4$/NR	NR/NR/NR	NR	Yes
Raghavendra et al. (2021),[21]	MRI	NR	No	VGG-16	No	Retrospective study, data from TCIA	1600/800/NR	NR/NR/NR	NR	Yes
Chakrabarty et al. (2021),[22]	MRI	NR	No	3D-CNN	No	Retrospective study, data from Washing University School of Medicine, BraTS 2018, BraTS 2019, TCIA, and TCGA	415/108/348	415/108/348	2001.02–2019.10	Yes
Yahyaoui et al. (2021),[23]	MRI	NR	No	3D-CNN	No	Retrospective study, data from BraTS 2015 and 2019	190/40/NR	NR/NR/NR	NR	Yes
Yamashiro et al. (2021),[24]	MRI	NR	No	3D-CNN	No	Retrospective study, data from BraTS 2018	6602/46/NR	NR/NR/NR	NR	Yes
Yao et al. (2021)[25]	MRI	NR	Yes	VGG-16	Yes	Prospective study, data from Hunan Cancer Hospital	NR/NR/NR	NR/NR/NR	2019.07–2020.02	No
Bezdan et al. (2021)[26]	MRI	NR	No	CNN-HEHO	Yes	Retrospective study, data from three datasets in TCIA	7200/800/NR	NR/NR/NR	NR	Yes
Shen et al. (2021),[27]	Fluorescent imaging	Yes	Yes	DCNN	Yes	Prospective study, data from Beijing Tiantan Hospital, Capital Medical University	636/296/NR	NR/NR/NR	2019.03–2020.4	No
Irmak et al. (2021),[28]	MRI	NR	No	CNN	No	Retrospective study, data from TCIA	3656/914/NR	NR/NR/NR	NR	Yes
Al-Saffar et al. (2021)[29]	MRI	NR	No	MLP&SVM	No	Retrospective study, data from TCIA	NR/NR/NR	NR/NR/NR	NR	Yes
Hu et al. (2021)[30]	MRI	NR	No	3D U-Net	No	Retrospective study, data from BraTS 2017, TCGA, and HuaShan Hospital	NR/NR/NR	533/267/NR	2001–2018	Yes

(*Continued on next page*)

Table 3. *Continued*

First author and year	Indicator definition: Device	Indicator definition: Exclusion of poor-quality imaging	Indicator definition: Heatmap provided	Algorithm: Algorithm architecture	Algorithm: Transfer learning applied	Data source: Source of data	Data source: Number of training/internal/external: Images	Data source: Number of training/internal/external: Cases	Data source: Data range	Data source: Open access data
Luo et al. (2021),[31]	MRI	NR	Yes	3D U-Net	No	Retrospective study, data from two hospitals including Huashan Hospital and Shanghai International Medical Center	NR/NR/NR	188/411/56	2010–2017	No
Decuyper et al. (2021),[33]	MRI	Yes	No	3D U-Net	No	Retrospective study, data from TCGA, BraTS 2019, and from Ghent University Hospital	NR/NR/NR	528/100/110	NR	Yes
Gutta et al. (2021),[34]	MRI	Yes	No	CNN	No	Retrospective study, data from the Keck Medical Center of the University of the Southern California	560/100/NR	NR/NR/NR	2007.05–2019.01	No
Ozcan et al. (2021)[35]	MRI	NR	No	CNN, AlexNet, GoogLeNet, SqueezeNet	Yes	Retrospective study, data from Amasya University	NR/NR/NR	83/21/NR	2016.12–2019.10	Yes
Mzoughi et al. (2021)[35]	MRI	Yes	No	3D-CNN	No	Retrospective study, data from BraTS 2018	NR/NR/NR	227/57/NR	NR	No
Koyuncu et al. (2020),[37]	MRI	NR	No	GM-CPSO-NN	No	Retrospective study, data from BraTS 2017	NR/NR/NR	143/142/NR	NR	Yes
Cinarer et al. (2020),[38]	MRI	NR	Yes	CNN	No	Retrospective study, data from TCIA	NR/NR/NR	95/26/NR	NR	Yes
Mzoughi et al. (2020)[38]	MRI	Yes	No	3D-CNN	No	Retrospective study, data from BraTS 2018	NR/NR/NR	284/67/NR	NR	Yes
Zhuge et al. (2020),[40]	MRI	Yes	No	2D R-CNN, 3DConvNet	No	Retrospective study, data from BraTS 2018 data and TCIA	NR/NR/NR	252/63/NR	NR	Yes
Naser et al. (2020),[41]	MRI	Yes	No	VGG-16	Yes	Retrospective study, data is available at TCIA	652/163/NR	86/22/NR	NR	Yes
Alis et al. (2020),[42]	MRI	Yes	No	MLP	No	Retrospective study, Istanbul Mehmet Akif Ersoy Thoracic and Cardiovascular Surgery Training and Research Hospital, Turkey	NR/NR/NR	121/60/NR	2013.01–2019.01	No

(*Continued on next page*)

Table 3. *Continued*

	Indicator definition			Algorithm		Data source				
First author and year	Device	Exclusion of poor-quality imaging	Heatmap provided	Algorithm architecture	Transfer learning applied	Source of data	Number of training/internal/external Images	Number of training/internal/external Cases	Data range	Open access data
Hollon et al. (2020),[43]	Stimulated Raman histology imaging	Yes	Yes	Inception-ResNet-v2	No	Prospective study, data from University of Michigan, Columbia University, and University of Miami	NR/NR/NR	NR/NR/74	2015.06–2019.02	No
Sharif et al. (2020)[43]	MRI	NR	No	Inception V3	No	Retrospective study, data from BraTS 2013, 2015, 2017 and 2018	NR/NR/NR	708/428/NR	NR	Yes
Lo et al. (2019),[45]	MRI	NR	No	AlexNet	Yes	Retrospective study, data from TCIA	117/13/NR	NR/NR/NR	NR	Yes
Gonbadi et al. (2019),[46]	MRI	NR	No	CNN	No	Retrospective study, data from BraTS 2017	$20^{5}/_{8}0$/NR	NR/NR/NR	NR	Yes
Ali et al. (2019),[47]	MRI	NR	No	DCGAN	No	Retrospective study, data from BraTS 2017	5220/1305/NR	NR/NR/NR	NR	Yes
Sultan et al. (2019),[28]	MRI	Yes	No	CNN	No	Retrospective study, data from the Repository of Molecular Brain Neoplasia Data, TCIA	439/77/NR	NR/NR/NR	NR	Yes
Muneer et al. (2019),[48]	MRI	Yes	No	VGG-19	Yes	Retrospective study, data from Government Medical College, India	389/168/NR and 553/228/NR	NR/NR/NR	NR	Yes
Anaraki et al. (2019),[50]	MRI	Yes	No	CNN+GA	Yes	Retrospective study, data from TCIA and Hazrat-e Rasool General Hospital at Tehran, Iran	6500/1500/NR	NR/NR/NR	NR	Yes
Sajjad et al. (2018),[49]	MRI	Yes	No	VGG-19	Yes	Retrospective study, data from Radiopaedia	$8^{1}/_{3}0$/NR and 2722/908/NR	NR/NR/NR	NR	Yes
Shahzadi et al. (2018)[49]	MRI	NR	No	CNN-LSTM	Yes	Retrospective study, data from BraTS 2015	NR/NR/NR	48/12/NR	NR	Yes
Yang et al. (2018)[51]	MRI	NR	No	AlexaNet and GoogLeNet	Yes	Retrospective study, data from Tangdu Hospital of the Fourth Military Medical College	694/173/NR	90/23/NR	NR	No
Al-Zurfi et al. (2018)[52]	MRI	NR	Yes	DINN	No	Retrospective study, data from TCIA	NR/NR/NR	29/1/NR	NR	Yes

(*Continued on next page*)

Table 3. *Continued*

First author and year	Indicator definition			Algorithm		Data source				
	Device	Exclusion of poor-quality imaging	Heatmap provided	Algorithm architecture	Transfer learning applied	Source of data	Number of training/ internal/external		Data range	Open access data
							Images	Cases		
Ge et al. (2018),[54]	MRI	NR	No	Multistream CNN fusion network	No	Retrospective study, data from BraTS 2017 competition	864/216/NR	NR/NR/NR	NR	Yes
Khawaldeh et al. (2018),[55]	MRI	NR	No	AlexNet	No	Retrospective study, data from TCIA	2627/448/NR	NR/NR/NR	NR	Yes
Ye et al. (2017),[56]	MRI	NR	No	3D CNN with GMU fusion	Yes	Retrospective study, data from BraTS 2015	NR/NR/NR	$24\frac{1}{3}$3/NR	NR	Yes

MRI = magnetic resonance imaging, BraTS = brain tumor segmentation, PACS = picture archiving and communication system, NR = not reported, CNN = convolutional neural network, TCIA = the cancer imaging archive, DLS-DARTS = double-learnable-stem differentiable architecture search, MICCAI = medical image computing and computer assisted interventions, CPM-Radpath = computational precision medicine: radiology-pathology, DCNN = deep convolutional neural network, HOMIF = hierarchical-order multimodal interaction fusion network, VGG = visual geometry group network, TCGA = the cancer genome atlas, HEHO = hybridized elephant herding optimization, MLP&SVM = multi-layer perceptron and support vector machine, GM-CPSO-NN = Gauss-map-based chaotic particle-swarm optimization neural network, R-CNN = residual convolutional neural network, DCGAN = deep convolutional generative adversarial networks, GA = genetic algorithms, LSTM = long short term memory, DINN = deep iteration matrix of neural network, GMU = gated multimodal unit.

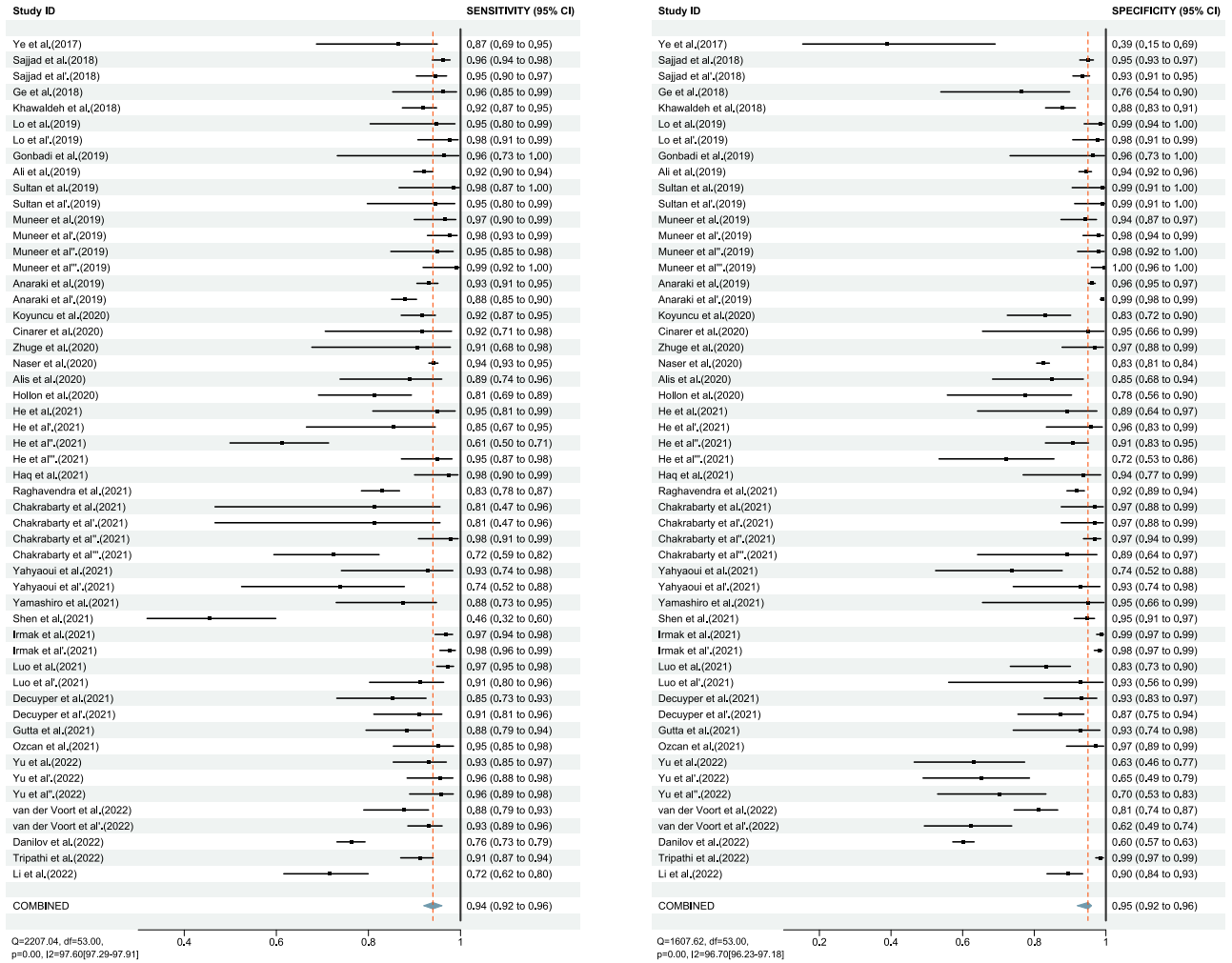


Figure 2. Forest plot of the pooled performance of deep learning (DL) algorithms, based on all 54 tables in 33 studies.

still existed in two subgroups (IV: I^2 = 85.92% in SE and 92.73% in SP; III+IV: I^2 = 98.60% in SE and 97.41% in SP) (Figures S6 and S13).

Type of validation: In internal subgroup, the SE was 94% (95% CI: 91–96%), SP was 94% (95% CI:92–96%) and AUC was 0.98 (95% CI: 0.97–0.99), while in external subgroup, they were 92% (95% CI: 88–95%), 82% (95% CI: 58–94%) and 0.94 (95% CI: 0.91–0.96), respectively. Heterogeneity still existed in two subgroups (internal: I^2 = 97.83% in SE and 95.78% in SP; external: I^2 = 48.29% in SE and 93.30% in SP) (Figures S7 and S14).

Publication bias evaluation

In the overall pooled analysis, the p value of Deeks' funnel plot was 0.873. In the highest accuracy pooled analysis, which value was 0.493. Neither of these analyses indicated publication bias (Figure S15).

Quality assessment

The quality of the total 48 included studies was assessed using QUADAS-2 and a summary of the risk of bias and applicability concerns for 48 studies was provided in Figure S16. The detailed results were also supplied in the Figure S17. In the patient selection domain of risk of bias, 35 studies were deemed high or unclear risk due to unreported inclusion and exclusion criteria, or unknown patient enrollment procedure. For index test, 35 studies were considered at an unclear risk because of a lack of pre-specified

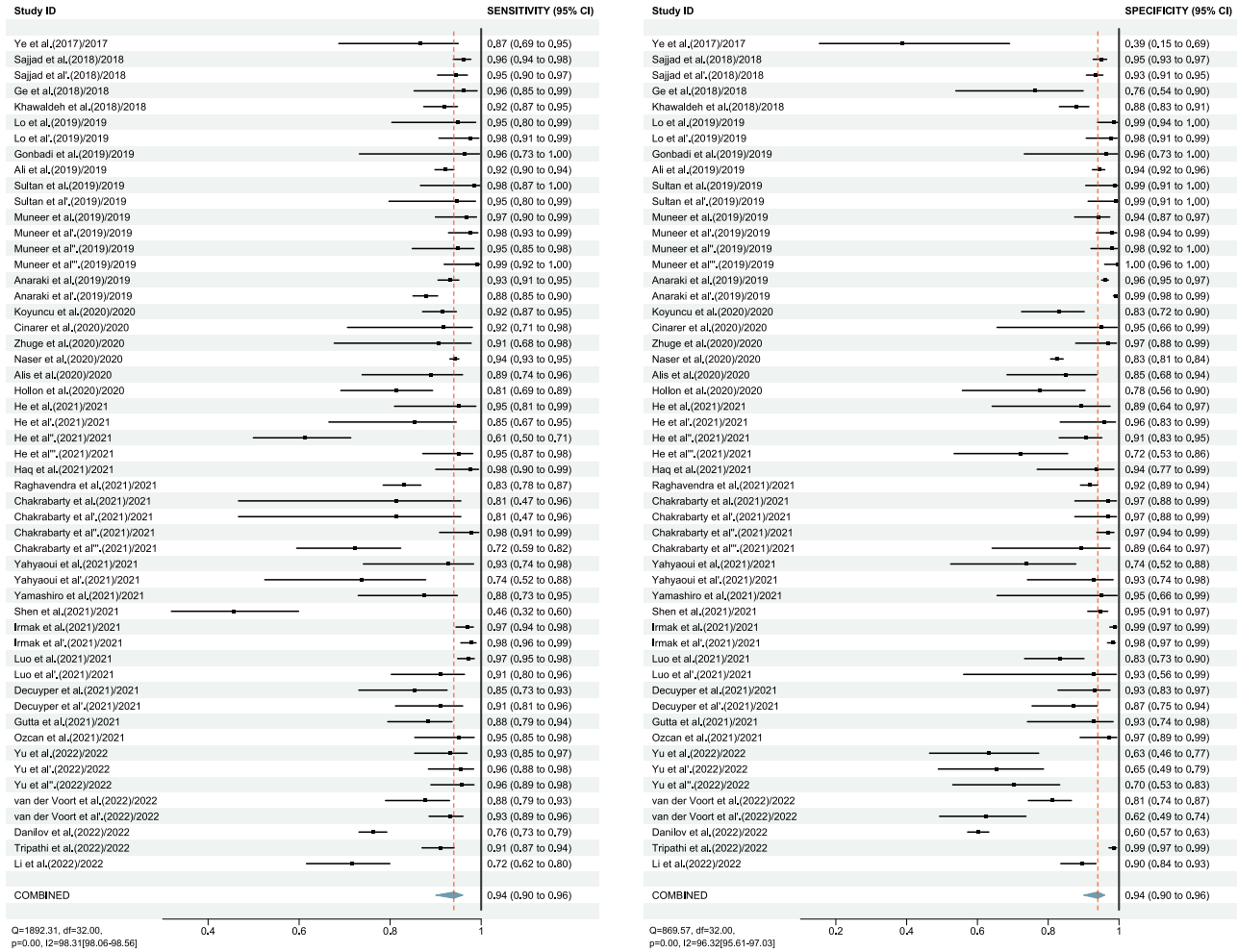


Figure 3. Forest plot of the pooled performance of deep learning (DL) algorithms, reporting the highest accuracy of 33 studies.

thresholds. No risk of bias was observed in the reference standard domain while the bias of flow and timing was unclear for 9 studies due to the following exclusion of patients for further analysis.

In the applicability concerns, 25 studies were considered at high or unclear applicability in the patient selection domain, 11 studies at unclear applicability in the index test domain, while low applicability concerns were observed for all studies in the reference standard domain.

Table 4 introduced supplemental figures, tables, and other information.

DISCUSSION

Up to now, previous systematic reviews and meta-analyses on AI applied to glioma focused on the following topics: prediction of AI on the molecular classification of glioma,[57–59] prediction the prognosis,[60] differential diagnosis between glioma and other brain tumors,[61,62] glioma image segmentation,[63] and grading of glioma.[64–66] As for the grading of glioma, two studies pointed out the current obstacles of AI deployment,[64,65] and one study conducted a meta-analysis on machine learning (ML) of grading.[66] However, though DL showed sufficient superiority in other cancers, such as cervical cancer and breast cancer,[67] it still remained vacant in grading glioma. Moreover, it is notable that glioma grading combined with AI exhibits some traits that are not present in other cancers, such as extensive use of public databases, and the importance of classification for prognosis.[22,66] Glioma grading based on open databases accounted for 26 of 33 in our study, which indicated a large number of applications for open-access data. Besides,

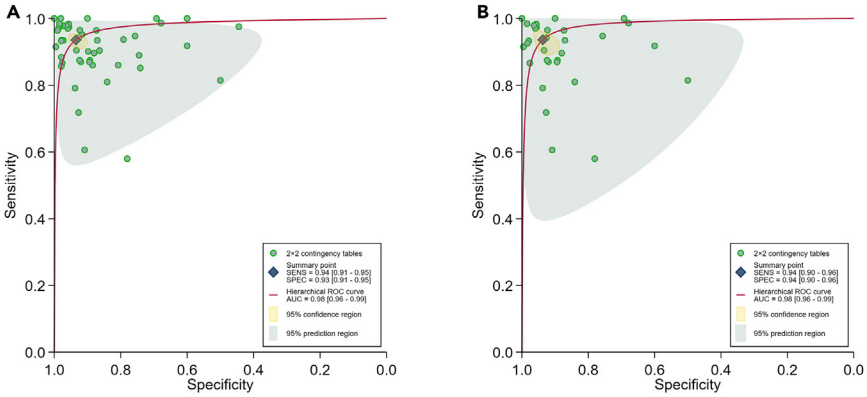


Figure 4. Overall pooled performance of deep learning (DL) algorithms on glioma grading

(A) Hierarchical summary receiver operating characteristic (HSROC) curve of all contingency tables (54 tables).

(B) HSROC curve reporting the highest accuracy (33 tables).

tumor grading is critical in glioma progression and prognosis; glioblastoma multiforme (grade IV) has a 5-year survival rate of less than 5%[68] while the survival rate of 15-year for grade II glioma is 86%.[68,69] Compared with traditional diagnostic methods, DL has advantages such as shorter diagnostic time, labor saving, and the ability to improve cancer screening in low-resource areas.[68] Thus, DL performance on glioma grading is worth lots of attention.

In our study the SE and SP were 94% (95% CI: 90–96%) and 94% (95% CI: 90–96%), respectively. A pertinent systematic review and meta-analysis focused on ML, which pooled 5 studies and showed the pooled SE and SP were 96% (95% CI: 93–98%) and 90% (95% CI: 85–94%).[66] From above results we can't differentiate the superiority of DL over ML. The potential explanation is that DL outperforms ML when the sample size is huge.[70] However, in our study, the median sample size was 130, which indicated that most of the eligible studies belonged to small sample data. Thus, DL performance might be hindered by data size limitations. Moreover, DL automatically extracts image features while ML mainly relies on images whose features have been extracted before, usually by clinicians or other experts.[64] This trait of DL makes it strongly hinge on the quality of images. In our study, only 16 of 33 studies excluded poor-quality images before processing. However, since DL algorithm after exclusion of poor-quality images will hardly present the real clinical setting; therefore, DL models should limit the exclusion of images.

It was noteworthy that we assessed DL from two different criteria: one used all available contingency tables; the other used only one contingency table reporting the highest accuracy from each article. The pooled results (SE:94% (95% CI: 91–95%), SP:93% (95% CI: 91–95%), and AUC:0.98 (95% CI: 0.96–0.99)) were modest worse than those in highest accuracy (SE:94% (95% CI: 90–96%), SP: 94% (95% CI: 90–96%), and AUC: 0.98 (95% CI: 0.96–0.99)). Besides, when the sample size increased, the confidence interval narrowed, which explained the phenomenon that CIs of overall datasets was narrower than the highest accuracy datasets. By the repeating use of samples in the overall analysis, it factitiously added the sample volume of duplicated articles. The phenomenon of single article containing multiple DL algorithms is commonplace in the oncology field,[9,22,71] which requests further meta-analysis of diagnostic evaluations to be equipped with the method to merge multiple sets within each study. Such an approach has already been used in clinical trials, but still remains vacant in diagnostic trials.[72]

As for subgroup analysis, in sample size ($\leq$ 130 or > 130) results, we didn't find the expected results that the bigger sample size group performed better than a smaller one. In views of the forest plot and original data, we could see that the > 130 sample size group contained narrower confidence intervals than the $\leq$ 130 group, but still incorporated poor results such as Shen et al. with 296 images (only 60.6% SE),[27] and Danilov et al. with 1773 images (only 58% SE and 78% SP).[11] Thus, the heterogeneity was still high in the > 130 sample size group, whereas it was decreased in the $\leq$ 130 group ($\leq$ 130: I^2 = 74.08% in SE and 80.42% in SP; > 130: I^2 = 99.05% in SE and 97.73% in SP). Our study implied that data quality varied enormously in the glioma classification area, which inevitably hindered us from drawing the conclusion of DL. Also, further study should embrace big data, which are the exact field DL experts in.[73]

Table 4. introduction of supplementary

Title	Introduction
Data S1	Search strategies for different databases
Figure S1	HSROC curves of different sample sizes
Figure S2	HSROC curves of open access data or not
Figure S3	HSROC curves of different internal validation types
Figure S4	HSROC curves of using transfer learning or not
Figure S5	HSROC curves of different data units
Figure S6	HSROC curves of glioma classification types
Figure S7	HSROC curves of validation types
FigureS8	Forest plot of different sample sizes
Figure S9	Forest plot of open access data or not
Figure S10	Forest plot of internal validation type
Figure S11	Forest plot of using transfer learning or not
Figure S12	Forest plot of data unit
Figure S13	Forest plot of glioma classification types
Figure S14	Forest plot of validation types
Figure S15	Funnel plot
Figure S16	QUADAS-2 summary plot
Figure S17	QUADAS-2 plot for each detailed item
Table S1	Meta regression result

In data sharing (open data/private data) subgroup analysis, we found that DL in open data performed superior than private data (SE: 95% vs. 88%; SP: 96% vs. 82%; AUC: 0.99 vs. 0.89). Glioma open-access databases, such as The Cancer Imaging Archive (TCIA)[74] and Brats[75] were used in 37 out of 49 studies included in the systematic review. Besides, these open databases were the standard databases in MICCAI (an AI competition held by Medical Image Computing and Computer Assisted Intervention Society), which is the top academic competition and play a cardinal role in AI. In these databases, the images were labeled and quality-checked by experienced clinical specialists and had been processed with standardization. By contrast to private data, the images of open data were of higher quality, which led to better results in the subgroup. Here, our results once again emphasized the great importance of data preprocessing. Besides, some recent efforts, which devoted toward standardization of preprocessing, preprocess datasets the same way as it is done for Brats. Thus, the data processed from these tools can be used alongside Brats data.[76] In this advanced field, there are not any regulations to ensure uniformity and high quality of preprocessing. Recently, the US Food & Drugs Administration (FDA) has approved serial available AI/ML-based medical devices and algorithms to standardize the process of AI tool development, which means that developers of algorithms go through rigorous evaluation before they launch their program.[8]

As for the type of internal validation, k-fold cross-validation outperformed random split-sample validation (SE: 96% vs. 92%; SP: 97% vs. 92%; AUC 0.99 vs. 0.97). K-fold cross-validation fits in small samples data and can conduct parameter tuning through multiple times of training and testing sets segmentation in the same database.[77] Therefore, it can improve the efficiency of data utilization. However, random split-sample validation only carries out cross-validation through one training set and test set segmentation, which has a large uncertainty, hardly to achieve true randomization.[63] In our study, since only 27% of the included research used k-fold cross-validation, we appeal for more k-fold cross-validation to be used in this field in future.

Another DL-related item was transfer learning, but in this study, we couldn't tell the superiority of transfer learning. Transfer learning enables a previously trained model used in another domain. Therefore, it skips the effort required to collect training data.[78] An article indicated that due to differences in demographic characteristics, transfer learning used on underrepresented patients might exert a negative influence on AI integration with oncology.[79] In this study, except for transfer learning used for open data, there were many studies using it from open data on private data with the discrepancy in the patient characteristics

with previous studies. For example, Shen et al., based on patients from a Chinese hospital, transferred the DL method from another study using the Brats database containing patients of the USA,[27] which indicated poor SE of 60.6%. Therefore, our result doesn't mean that transfer learning isn't suitable in this domain, since transfer learning from different population might create dissatisfying results due to populations but not the AI algorithm itself. Further, we hope that in the future when researchers apply transfer learning, they are supposed to take population heterogeneity into consideration.

With suspect to using images or patients as data unit, we concluded that image-based dataset showed better SE (95% vs. 92%) and SP (96% vs. 88%). It is noteworthy that whether the study reported image number or patient number, the AI process is still based on the image. Therefore, articles only reporting patient number were somehow without preciseness. Especially in contingency tables, if the article only provided patient numbers instead of image numbers, we inevitably underestimated the sample size of these studies, since every patient usually generates more than 1 image. Therefore, to obtain high-quality results, articles should exactly report not only patient numbers but also image numbers, and better report the image-contained results in contingency tables.

In the classification of HGG and LGG, 56% contingency tables (30/54) included in our study defined IV grade as HGG, such as Luo et al.[31] and Decuyper et al.,[32] while others defined III and IV(24/54) grades as HGG, such as Danilov et al.[11] and Li et al.[16] Here, our study found that IV represented HGG in classification was similar to III+IV(SE: 94% vs. 93%; SP:93% vs. 94%; AUC 0.98 vs. 0.98). In WHO glioma classification, diffuse glioma is defined as WHO grade II, anaplastic, or in case of 1p/19q-non-codeleted tumor as grade III and glioblastoma as grade IV.[80] In the image diagnosis, glioblastoma has the most invasive feature, which can be distinguished from diffuse astrocytoma and anaplastic astrocytoma.[81] However, distinguishing anaplastic astrocytoma and diffuse astrocytoma features in the images is another story. If researchers deem III+IV as HGG, which is to distinguish grade III from grade II, it cannot be easy to achieve.[82] Recent studies indicated that molecular profiling differences existing between these two grades might be used in classification. Important molecular diagnostic markers, such as isocitrate dehydrogenase (IDH) mutation,[83] 1p/19q co-deletion[84] and O-6-methylguanine-DNA methyltransferase promoter methylation,[85] had been included into guideline since WHO glioma classification 2016.[80] Therefore, in the future, DL algorithms evaluation in image-based glioma grading should also take molecular diagnostic markers into consideration, especially in the distinguishment of grade IIand III glioma.

As for internal validation or external validation, in our study, internal subgroup was superior to external subgroup (SE: 94% vs. 92%; SP: 94% vs. 82%; AUC 0.98 vs. 0.94). Internal validation is that in the validation phase, the testing set is separated from the original dataset, whereas external validation is that using a completely independent dataset out of the original one.[86] Though DL performed inferiorly in the external group, we still appeal to further studies to apply external instead of internal validation. One of the major limitations of including studies is that the majority of them didn't implement external validation, which made them hard to be generalized and reproduced. DL development is supposed to consider data extrapolation. DL algorithm should generalize to the real-world usage, which means not only exerts well in online database but also can show acceptable quality in clinical practice, such as being auxiliary with hospital clinician or commune healthcare worker. Xian et al. used DL in near-infrared fluorescence imaging to help intraoperative diagnosis,[86] which requested not only accuracy but also celerity. Besides, as for the application of AI technology in low-resource areas, the acceptance ability of healthcare workers also needs to be considered.[87] Therefore, in order to facilitate the practical application and promotion of DL merging with glioma classification, in addition to algorithm optimization, time of DL diagnosis, maneuverability, protection of patient information, etc, should also be under rigorous design.

To improve DL algorithms combined with glioma, based on previous analysis of our research, we try to summarize limitations in this field: 1) diagnostic trials require standard method for data merging for AI; 2) small sample size; 3) poor-quality image preprocessing; 4) not standard algorithm development; 5) not standard data report; 6) different definition of HGG and LGG; and 7) poor extrapolation.

Moreover, we offered the following suggestions for further separate studies in AI development in glioma: 1) use open databases; 2) before disclosure, be approved by FDA or other authoritative institutions first; 3) embrace big data; 4) encourage the use of k-fold cross-validation; 5) consider the consistency of characteristics of the two studies populations when using transfer learning; 6) report the number of images in

contingency tables; and 7) encourage external validation. Besides, we expect diagnostic trials to provide normative guidelines for data fusion, and top institutes can convene specialists from regarding professions, such as clinicians, AI engineers, pathologists to standardize image preprocessing, and AI development.

Our study used meta-analysis to integrate articles about the performance of DL algorithms in image-based glioma grading. To the best of our knowledge, this is the first meta-analysis to explore the performance of DL in this field. When analyzing the full data, we considered both the full use of data and the selection of representative data (the highest accuracy) of an article, which might be a reference way in the absence of the standard of combining multiple sets of data in diagnostic tests. We further used meta-regression to explore the source of heterogeneity, which indicated that sample size, data sharing, type of internal validation, transfer learning applies, classification, and type of validation did play an important role in heterogeneity. In subgroup analysis, we find that DL displayed with distinguishment in different subgroups. In explanation of difference, we gave recommendations under which DL performs more superior. More importantly, we provided suggestions on how DL should be normalized in glioma grading in the future based on the dilemma of DL development that existed in our results.

However, there are still some limitations in this study. Our results showed high heterogeneity, which was not significantly reduced in subgroup analysis. The items used in subgroup analysis were proved to exert an impact on heterogeneity in meta-regression, and they were considered as possible heterogeneity sources in previous studies. Liu et al. conducted a pooled analysis to evaluate the performance of healthcare workers versus DL, which implied the separation of DL from clinician data in studies.[88] Besides, DL-related items also contained huge diversity, such as performing external validation or internal validation,[89] using open-access dataset or not,[90] the application of transfer learning or not, as well as the validation type.[91] Therefore, items included in this study were scientific and have been shown to contribute to heterogeneity. Moreover, high heterogeneity was common in studies of the convergence of AI and medicine, such as the DL study of breast and cervical cancer,[67] glioma segmentation,[63] gastrointestinal cancer classification, and prognostication[92] and so on. Admittedly, the reason why heterogeneity was not reduced might also be explained by other possible factors, such as prospective or retrospective studies and DL diagnoses or clinician diagnoses.[67] Due to the scarcity of articles containing prospective studies ($^2/_3$3) or with a comparison of DL versus clinician (3/33) in our study, we couldn't perform meta-regression on them. Another limitation is that in glioma classification, we failed to incorporate molecular information, which is becoming increasingly important, since it marks a more refined classification of patients and is critical for clinical treatment choice and prognosis.[64] In addition, the QUADAS-2 assessment was not tailored for AI-based studies, which resulted in risk of bias and applicability concerns.

In conclusion, though the SE, SP, and AUC of DL algorithms are high in glioma grading, we still couldn't prove the superiority of DL over ML. In the whole dataset pooled analysis, we considered both the full use of data and the selection of representative data (the highest accuracy) for each article. Our study suggested that the results were highly heterogeneous and sample size, data sharing, type of internal validation, transfer learning applies, classification, and type of validation were the possible reasons. In subgroup analysis, we didn't find the bigger sample size group displayed better than the smaller one. DL in open data appeared superior to private data. As for type of internal validation, k-fold cross-validation outperformed random split-sample validation. In transfer learning use, we couldn't tell the superiority of transfer learning appliance in comparison to not use. Image-based datasets showed better results than the patients-based ones. In the classification of HGG and LGG, our study indicated that IV represented HGG excelled III+IV. As for internal validation or external validation, in our study, the internal subgroup was superior to the external. Besides, from the perspective of the whole results of our study, we strongly recommend separate research:1) use open databases; 2) before disclosure, be approved by FDA or other authoritative institutions first; 3) embrace big data; 4) encourage the use of random split-sample validation; 5) consider the consistency of characteristics of the two studies populations when using transfer learning; 6) report the number of images in contingency tables; 7) include molecular typing results to assist diagnosis if grade III incorporated in HGG; 8) encourage external validation; and 9) incorporation of AI-based quality of reporting tools (such as Quadas-AI, Probast-AI or Tripod-AI). Moreover, we can't emphasize more on normalization of image extraction, preprocessing, and algorithm development in this field. However, since the heterogeneity still remained in subgroups, these recommendations should be considered cautiously.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Search strategy and eligibility criteria
  - Data extraction
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quality assessment
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106815.

ACKNOWLEDGMENTS

We appreciate Professor Yu Jiang for teaching meta-analysis courses. We also are grateful for involved articles researchers to offer detailed diagnostic test results. We would like to thank Peking Union Medical College Education Foundation (NO: B0202023F-11) for funding us.

AUTHOR CONTRIBUTIONS

W.Y.S. carried out data analysis and article writing. C.S. was responsible for research retrieval. C.T. and C.H.P. took part in figures visualization and table compiling. J.H.F. and P.X. offered the idea of this research and supervised this work. Y.L..Q reviewed the article and provided suggestions for revision.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: November 1, 2022
Revised: March 23, 2023
Accepted: May 2, 2023
Published: May 5, 2023

REFERENCES

1. Ostrom, Q.T., Bauchet, L., Davis, F.G., Deltour, I., Fisher, J.L., Langer, C.E., Pekmezci, M., Schwartzbaum, J.A., Turner, M.C., Walsh, K.M., et al. (2014). The epidemiology of glioma in adults: a "state of the science" review. Neuro Oncol. *16*, 896–913. https://doi.org/10.1093/neuonc/nou087.

2. Weller, M., van den Bent, M., Preusser, M., Le Rhun, E., Tonn, J.C., Minniti, G., Bendszus, M., Balana, C., Chinot, O., Dirven, L., et al. (2021). EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat. Rev. Clin. Oncol. *18*, 170–186. https://doi.org/10.1038/s41571-020-00447-z.

3. Di Carlo, D.T., Cagnazzo, F., Benedetto, N., Morganti, R., and Perrini, P. (2019). Multiple high-grade gliomas: epidemiology, management, and outcome. A systematic review and meta-analysis. Neurosurg. Rev. *42*, 263–275. https://doi.org/10.1007/s10143-017-0928-7.

4. Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., et al. (2021). The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro Oncol. *23*, 1231–1251. https://doi.org/10.1093/neuonc/noab106.

5. McGirt, M.J., Woodworth, G.F., Coon, A.L., Frazier, J.M., Amundson, E., Garonzik, I., Olivi, A., and Weingart, J.D. (2005). Independent predictors of morbidity after image-guided stereotactic brain biopsy: a risk assessment of 270 cases. J. Neurosurg. *102*, 897–901. https://doi.org/10.3171/jns.2005.102.5.0897.

6. Coiera, E. (2018). The fate of medicine in the time of AI. Lancet (London, England) *392*, 2331–2332. https://doi.org/10.1016/s0140-6736(18)31925-1.

7. Kleppe, A., Skrede, O.J., De Raedt, S., Liestøl, K., Kerr, D.J., and Danielsen, H.E. (2021). Designing deep learning studies in cancer diagnostics. Nat. Rev. Cancer *21*, 199–211. https://doi.org/10.1038/s41568-020-00327-9.

8. Benjamens, S., Dhunnoo, P., and Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit. Med. *3*, 118. https://doi.org/10.1038/s41746-020-00324-0.

9. Yu, X., Wu, Y., Bai, Y., Han, H., Chen, L., Gao, H., Wei, H., and Wang, M. (2022). A lightweight 3D UNet model for glioma grading. Phys. Med. Biol. *67*, 155006. https://doi.org/10.1088/1361-6560/ac7d33.

10. van der Voort, S.R., Incekara, F., Wijnenga, M.M.J., Kapsas, G., Gahrmann, R., Schouten, J.W., Nandoe Tewarie, R., Lycklama, G.J., De Witt Hamer, P.C., Eijgelaar, R.S., et al. (2023). Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. Neuro Oncol. *25*, 279–289. https://doi.org/10.1093/neuonc/noac166.

11. Danilov, G., Korolev, V., Shifrin, M., Ilyushin, E., Maloyan, N., Saada, D., Ishankulov, T., Afandiev, R., Shevchenko, A., Konakova, T., et al. (2022). Noninvasive glioma grading with deep learning: a pilot study. Stud. Health Technol. Inf. *290*, 675–678. https://doi.org/10.3233/shti220163.

12. Chen, Z., Li, N., Liu, C., and Yan, S. (2022). Deep convolutional neural network-based brain magnetic resonance imaging applied in glioma diagnosis and tumor region identification. Contrast Media Mol. Imaging *2022*, 4938587. https://doi.org/10.1155/2022/4938587.

13. Tripathi, P.C., and Bag, S. (2022). A computer-aided grading of glioma tumor using deep residual networks fusion. Comput. Methods Progr. Biomed. *215*, 106597. https://doi.org/10.1016/j.cmpb.2021.106597.

14. Xiao, A., Shen, B., Shi, X., Zhang, Z., Zhang, Z., Tian, J., Ji, N., and Hu, Z. (2022). Intraoperative glioma grading using neural architecture search and multi-modal imaging. IEEE Trans. Med. Imag. *41*, 2570–2581. https://doi.org/10.1109/tmi.2022.3166129.

15. Wang, X., Wang, R., Yang, S., Zhang, J., Wang, M., Zhong, D., Zhang, J., and Han, X. (2022). Combining radiology and pathology for automatic glioma classification. Front. Bioeng. Biotechnol. *10*, 841958. https://doi.org/10.3389/fbioe.2022.841958.

16. Li, Y., Wei, D., Liu, X., Fan, X., Wang, K., Li, S., Zhang, Z., Ma, K., Qian, T., Jiang, T., et al. (2022). Molecular subtyping of diffuse gliomas using magnetic resonance imaging: comparison and correlation between radiomics and deep learning. Eur. Radiol. *32*, 747–758. https://doi.org/10.1007/s00330-021-08237-6.

17. Khazaee, Z., Langarizadeh, M., and Shiri Ahmadabadi, M.E. (2022). Developing an artificial intelligence model for tumor grading and classification, based on MRI sequences of human brain gliomas. Int. J. Cancer Manag. *15*. https://doi.org/10.5812/ijcm.120638.

18. Linqi, J., Chunyu, N., and Jingyang, L. (2022). Glioma classification framework based on SE-ResNeXt network and its optimization. IET Image Process. *16*, 596–605. https://doi.org/10.1049/ipr2.12374.

19. He, M., Han, K., Zhang, Y., and Chen, W. (2021). Hierarchical-order multimodal interaction fusion network for grading gliomas. Phys. Med. Biol. *66*, 215016. https://doi.org/10.1088/1361-6560/ac30a1.

20. Haq, E.U., Jianjun, H., Li, K., Haq, H.U., and Zhang, T. (2021). An MRI-based deep learning approach for efficient classification of brain tumors. J. Ambient Intell. Hum. Comput. https://doi.org/10.1007/s12652-021-03535-9.

21. Raghavendra, U., Gudigar, A., Rao, T.N., Rajinikanth, V., Ciaccio, E.J., Yeong, C.H., Satapathy, S.C., Molinari, F., and Acharya, U.R. (2022). Feature-versus deep learning-based approaches for the automated detection of brain tumor with magnetic resonance images: a comparative study. Int. J. Imag. Syst. Technol. *32*, 501–516. https://doi.org/10.1002/ima.22646.

22. Chakrabarty, S., Sotiras, A., Milchenko, M., LaMontagne, P., Hileman, M., and Marcus, D. (2021). MRI-based identification and classification of major intracranial tumor types by using a 3D convolutional neural network: a retrospective multi-institutional analysis. Radiol. Artif. Intell. *3*, e200301. https://doi.org/10.1148/ryai.2021200301.

23. Yahyaoui, H., Ghazouani, F., and Farah, I.R. (2021). Deep Learning Guided by an Ontology for Medical Images Classification Using a Multimodal Fusion, *4–5*, pp. 1–6.

24. Yamashiro, H., Teramoto, A., Saito, K., and Fujita, H. (2021). Development of a fully automated glioma-grading pipeline using post-contrast T1-weighted images combined with cloud-based 3D convolutional neural network. Appl. Sci. *11*, 5118.

25. Yao, W., and Thomas, S. (2021). Deep learning-based magnetic resonance imaging image feature analysis for pathological classification of brain glioma. Sci. Program. *2021*, 1–9. https://doi.org/10.1155/2021/6778009.

26. Bezdan, T., Milosevic, S., Zivkovic, M., Bacanin, N., and Strumberger, I. (2021). Optimizing Convolutional Neural Network by Hybridized Elephant Herding Optimization Algorithm for Magnetic Resonance Image Classification of Glioma Brain Tumor Grade, *26–27*, pp. 171–176.

27. Shen, B., Zhang, Z., Shi, X., Cao, C., Zhang, Z., Hu, Z., Ji, N., and Tian, J. (2021). Real-time intraoperative glioma diagnosis using fluorescence imaging and deep convolutional neural networks. Eur. J. Nucl. Med. Mol. Imag. *48*, 3482–3492. https://doi.org/10.1007/s00259-021-05326-y.

28. Sultan, H.H., Salem, N.M., and Al-Atabany, W. (2019). Multi-classification of brain tumor images using deep neural network. IEEE Access *7*, 69215–69225. https://doi.org/10.1109/ACCESS.2019.2919122.

29. Al-Saffar, Z.A., and Yildirim, T. (2021). A hybrid approach based on multiple Eigenvalues selection (MES) for the automated grading of a brain tumor using MRI. Comput. Methods Progr. Biomed. *201*, 105945. https://doi.org/10.1016/j.cmpb.2021.105945.

30. Hu, Z., Zhuang, Q., Xiao, Y., Wu, G., Shi, Z., Chen, L., Wang, Y., and Yu, J. (2021). MIL normalization – prerequisites for accurate MRI radiomics analysis. Comput. Biol. Med. *133*, 104403. https://doi.org/10.1016/j.compbiomed.2021.104403.

31. Luo, H., Zhuang, Q., Wang, Y., Abudumijiti, A., Shi, K., Rominger, A., Chen, H., Yang, Z., Tran, V., Wu, G., et al. (2021). A novel image signature-based radiomics method to achieve precise diagnosis and prognostic stratification of gliomas. Lab. Invest. *101*, 450–462. https://doi.org/10.1038/s41374-020-0472-x.

32. Decuyper, M., Bonte, S., Deblaere, K., and Van Holen, R. (2021). Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma. Comput. Med. Imag. Graph. *88*, 101831. https://doi.org/10.1016/j.compmedimag.2020.101831.

33. Gutta, S., Acharya, J., Shiroishi, M.S., Hwang, D., and Nayak, K.S. (2021). Improved glioma grading using deep convolutional neural networks. AJNR. Am. J. Neuroradiol. *42*, 233–239. https://doi.org/10.3174/ajnr.A6882.

34. Özcan, H., Emiroğlu, B.G., Sabuncuoğlu, H., Özdoğan, S., Soyer, A., and Saygı, T. (2021). A comparative study for glioma classification using deep convolutional neural networks. Math. Biosci. Eng. *18*, 1550–1572. https://doi.org/10.3934/mbe.2021080.

35. Mzoughi, H., Njeh, I., Slima, M.B., Ben Hamida, A., Mhiri, C., and Mahfoudh, K.B. (2021). Towards a computer aided diagnosis (CAD) for brain MRI glioblastomas tumor exploration based on a deep convolutional neuronal networks (D-CNN) architectures. Multimed. Tool. Appl. *80*, 899–919. https://doi.org/10.1007/s11042-020-09786-6.

36. Koyuncu, H., Barstuğan, M., and Öziç, M.Ü. (2020). A comprehensive study of brain tumour discrimination using phase combinations, feature rankings, and hybridised classifiers. Med. Biol. Eng. Comput. *58*, 2971–2987. https://doi.org/10.1007/s11517-020-02273-y.

37. Çinarer, G., Emiroğlu, B.G., and Yurttakal, A.H. (2020). Prediction of glioma grades using deep learning with wavelet radiomic features. Appl. Sci. *10*, 6296.

38. Mzoughi, H., Njeh, I., Wali, A., Slima, M.B., BenHamida, A., Mhiri, C., and Mahfoudhe, K.B. (2020). Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification. J. Digit. Imag. *33*, 903–915. https://doi.org/10.1007/s10278-020-00347-9.

39. Zhuge, Y., Ning, H., Mathen, P., Cheng, J.Y., Krauze, A.V., Camphausen, K., and Miller, R.W. (2020). Automated glioma grading on conventional MRI images using deep convolutional neural networks. Med. Phys. *47*, 3044–3053. https://doi.org/10.1002/mp.14168.

40. Naser, M.A., and Deen, M.J. (2020). Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. Comput. Biol. Med. *121*, 103758. https://doi.org/10.1016/j.compbiomed.2020.103758.

41. Alis, D., Bagcilar, O., Senli, Y.D., Isler, C., Yergin, M., Kocer, N., Islak, C., and Kizilkilic, O. (2020). The diagnostic value of quantitative texture analysis of conventional MRI sequences using artificial neural networks in grading gliomas. Clin. Radiol. *75*, 351–357. https://doi.org/10.1016/j.crad.2019.12.008.

42. Hollon, T.C., Pandian, B., Adapa, A.R., Urias, E., Save, A.V., Khalsa, S.S.S., Eichberg, D.G., D'Amico, R.S., Farooq, Z.U., Lewis, S., et al. (2020). Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nat. Med. *26*, 52–58. https://doi.org/10.1038/s41591-019-0715-9.

43. Sharif, M.I., Li, J.P., Khan, M.A., and Saleem, M.A. (2020). Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images. Pattern Recogn. Lett. *129*, 181–189. https://doi.org/10.1016/j.patrec.2019.11.019.

44. Lo, C.-M., Chen, Y.-C., Weng, R.-C., and Hsieh, K.L.-C. (2019). Intelligent glioma grading based on deep transfer learning of MRI radiomic features. Appl. Sci. *9*, 4926.

45. Gonbadi, F.B., and Khotanlou, H. (2019). Glioma Brain Tumors Diagnosis and Classification in MR Images Based on Convolutional Neural Networks, *24–25*, pp. 1–5.

46. Ali, M.B., Gu, I.Y.-H., and Jakola, A.S. (2019). In Multi-stream Convolutional Autoencoder and 2D Generative Adversarial Network for Glioma Classification, G. Percannella, ed. Held in Cham, 2019//. M. Vento (Springer International Publishing), pp. 234–245.

47. Ahammed Muneer, K.V., Rajendran, V.R., and K, P.J. (2019). Glioma tumor grade identification using artificial intelligent techniques. J. Med. Syst. *43*, 113. https://doi.org/10.1007/s10916-019-1228-2.

48. Kabir Anaraki, A., Ayati, M., and Kazemi, F. (2019). Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. Biocybern. Biomed. Eng. *39*, 63–74. https://doi.org/10.1016/j.bbe.2018.10.004.

49. Shahzadi, I., Tang, T.B., Meriadeau, F., and Quyyum, A. (2018). CNN-LSTM: Cascaded Framework for Brain Tumour Classification, *3–6*, pp. 633–637. .

50. Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., and Baik, S.W. (2019). Multi-grade brain tumor classification using deep CNN with extensive data augmentation. Journal of Computational Science *30*, 174–182. https://doi.org/10.1016/j.jocs.2018.12.003.

51. Yang, Y., Yan, L.F., Zhang, X., Han, Y., Nan, H.Y., Hu, Y.C., Hu, B., Yan, S.L., Zhang, J., Cheng, D.L., et al. (2018). Glioma grading on conventional MR images: a deep learning study with transfer learning. Front. Neurosci. *12*, 804. https://doi.org/10.3389/fnins.2018.00804.

52. Al-Zurfi, A., Meziane, F., and Aspin, R. (2018). Automated Glioma Grading Based on an Efficient Ensemble Design of a Multiple Classifier System Using Deep Iteration Neural Networks Matrix, *6–7*, pp. 1–6.

53. Ge, C., Gu, I.Y., Jakola, A.S., and Yang, J. (2018). Deep learning and multi-sensor fusion for glioma classification using multistream 2D convolutional networks. Annual international conference of the IEEE engineering in medicine and biology society. IEEE engineering in medicine and biology society. Annual International Conference, 5894–5897. https://doi.org/10.1109/embc.2018.8513556.

54. Khawaldeh, S., Pervaiz, U., Rafiq, A., and Alkhawaldeh, R. (2017). Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks. Appl. Sci. *8*, 27.

55. Ye, F., Pu, J., Wang, J., Li, Y., and Zha, H. (2017). Glioma Grading Based on 3D Multimodal Convolutional Neural Network and Privileged Learning, *13–16*, pp. 759–763.

56. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

57. Zhao, J., Huang, Y., Song, Y., Xie, D., Hu, M., Qiu, H., and Chu, J. (2020). Diagnostic accuracy and potential covariates for machine learning to identify IDH mutations in glioma patients: evidence from a meta-analysis. Eur. Radiol. *30*, 4664–4674. https://doi.org/10.1007/s00330-020-06717-9.

58. van Kempen, E.J., Post, M., Mannil, M., Kusters, B., Ter Laan, M., Meijer, F.J.A., and Henssen, D. (2021). Accuracy of machine learning algorithms for the classification of molecular features of gliomas on MRI: a systematic literature review and meta-analysis. Cancers *13*, 2606. https://doi.org/10.3390/cancers13112606.

59. Jian, A., Jang, K., Manuguerra, M., Liu, S., Magnussen, J., and Di Ieva, A. (2021). Machine learning for the prediction of molecular markers in glioma on magnetic resonance imaging: a systematic review and meta-analysis. Neurosurgery *89*, 31–44. https://doi.org/10.1093/neuros/nyab103.

60. Zhou, Q., Xue, C., Ke, X., and Zhou, J. (2022). Treatment response and prognosis evaluation in high-grade glioma: an imaging review based on MRI. J. Magn. Reson. Imag. *56*, 325–340. https://doi.org/10.1002/jmri.28103.

61. Subramanian, H., Dey, R., Brim, W.R., Tillmanns, N., Cassinelli Petersen, G., Brackett, A., Mahajan, A., Johnson, M., Malhotra, A., and Aboian, M. (2021). Trends in development of novel machine learning methods for the identification of gliomas in datasets that include non-glioma images: a systematic review. Front. Oncol. *11*, 788819. https://doi.org/10.3389/fonc.2021.788819.

62. Nguyen, A.V., Blears, E.E., Ross, E., Lall, R.R., and Ortega-Barnett, J. (2018). Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis. Neurosurg. Focus *45*, E5. https://doi.org/10.3171/2018.8.Focus18325.

63. van Kempen, E.J., Post, M., Mannil, M., Witkam, R.L., Ter Laan, M., Patel, A., Meijer, F.J.A., and Henssen, D. (2021). Performance of machine learning algorithms for glioma segmentation of brain MRI: a systematic literature review and meta-analysis. Eur. Radiol. *31*, 9638–9653. https://doi.org/10.1007/s00330-021-08035-0.

64. Merkaj, S., Bahar, R.C., Zeevi, T., Lin, M., Ikuta, I., Bousabarah, K., Cassinelli Petersen, G.I., Staib, L., Payabvash, S., Mongan, J.T., et al. (2022). Machine learning tools for image-based glioma grading and the quality of their reporting: challenges and opportunities. Cancers *14*, 2623. https://doi.org/10.3390/cancers14112623.

65. Bahar, R.C., Merkaj, S., Cassinelli Petersen, G.I., Tillmanns, N., Subramanian, H., Brim, W.R., Zeevi, T., Staib, L., Kazarian, E., Lin, M., et al. (2022). Machine learning models for classifying high- and low-grade gliomas: a systematic review and quality of reporting analysis. Front. Oncol. *12*, 856231. https://doi.org/10.3389/fonc.2022.856231.

66. Sohn, C.K., and Bisdas, S. (2020). Diagnostic accuracy of machine learning-based radiomics in grading gliomas: systematic review and meta-analysis. Contrast Media Mol. Imaging *2020*, 2127062. https://doi.org/10.1155/2020/2127062.

67. Xue, P., Wang, J., Qin, D., Yan, H., Qu, Y., Seery, S., Jiang, Y., and Qiao, Y. (2022). Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. NPJ Digit. Med. *5*, 19. https://doi.org/10.1038/s41746-022-00559-z.

68. Erson-Omay, E.Z., Henegariu, O., Omay, S.B., Harmancı, A.S., Youngblood, M.W., Mishra-Gorur, K., Li, J., Özduman, K., Carrión-Grant, G., Clark, V.E., et al. (2017). Longitudinal analysis of treatment-induced genomic alterations in gliomas. Genome Med. *9*, 12. https://doi.org/10.1186/s13073-017-0401-9.

69. Armstrong, G.T., Conklin, H.M., Huang, S., Srivastava, D., Sanford, R., Ellison, D.W., Merchant, T.E., Hudson, M.M., Hoehn, M.E., Robison, L.L., et al. (2011). Survival and long-term health and cognitive outcomes after low-grade glioma. Neuro Oncol. *13*, 223–234. https://doi.org/10.1093/neuonc/noq178.

70. Lee, J., Wang, N., Turk, S., Mohammed, S., Lobo, R., Kim, J., Liao, E., Camelo-Piragua, S., Kim, M., Junck, L., et al. (2020). Discriminating pseudoprogression and true progression in diffuse infiltrating glioma using multi-parametric MRI data through deep learning. Sci. Rep. *10*, 20331. https://doi.org/10.1038/s41598-020-77389-0.

71. Lu, C., Koyuncu, C., Corredor, G., Prasanna, P., Leo, P., Wang, X., Janowczyk, A., Bera, K., Lewis, J., Velcheti, V., and Madabhushi, A. (2021). Feature-driven local cell graph (FLocK): new computational pathology-based descriptors for prognosis of lung cancer and HPV status of oropharyngeal cancers. Med. Image Anal. *68*, 101903. https://doi.org/10.1016/j.media.2020.101903.

72. Woods, B.S., Hawkins, N., and Scott, D.A. (2010). Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: a tutorial. BMC Med. Res. Methodol. *10*, 54. https://doi.org/10.1186/1471-2288-10-54.

73. Koppe, G., Meyer-Lindenberg, A., and Durstewitz, D. (2021). Deep learning for small and big data in psychiatry. Neuropsychopharmacology *46*, 176–190. https://doi.org/10.1038/s41386-020-0767-z.

74. The Cancer Imaging Archive (TCIA). https://www.cancerimagingarchive.net/.

75. Brain Tumor Segmentation (BraTS). http://www.braintumorsegmentation.org/.

76. Kofler, F., Berger, C., Waldmannstetter, D., Lipkova, J., Ezhov, I., Tetteh, G., Kirschke, J., Zimmer, C., Wiestler, B., and Menze, B.H. (2020). BraTS toolkit: translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. Front. Neurosci. *14*, 125. https://doi.org/10.3389/fnins.2020.00125.

77. Moreno-Torres, J.G., Saez, J.A., and Herrera, F. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Transact. Neural Networks Learn. Syst. *23*, 1304–1312. https://doi.org/10.1109/tnnls.2012.2199516.

78. Ayana, G., Dese, K., and Choe, S.W. (2021). Transfer learning in breast cancer diagnoses via ultrasound imaging. Cancers *13*, 738. https://doi.org/10.3390/cancers13040738.

79. Byra, M., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., and Andre, M. (2019). Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med. Phys. *46*, 746–755. https://doi.org/10.1002/mp.13361.

80. Wesseling, P., and Capper, D. (2018). WHO 2016 Classification of gliomas. Neuropathol. Appl. Neurobiol. *44*, 139–150. https://doi.org/10.1111/nan.12432.

81. Weller, M., Wick, W., Aldape, K., Brada, M., Berger, M., Pfister, S.M., Nishikawa, R., Rosenthal, M., Wen, P.Y., Stupp, R., and Reifenberger, G. (2015). Glioma. Nat. Rev. Dis. Prim. *1*, 15017. https://doi.org/10.1038/nrdp.2015.17.

82. Weller, M., Weber, R.G., Willscher, E., Riehmer, V., Hentschel, B., Kreuz, M., Felsberg, J., Beyer, U., Löffler-Wirth, H., Kaulich, K., et al. (2015). Molecular classification of diffuse cerebral WHO grade II/III gliomas using genome- and transcriptome-wide profiling improves stratification of prognostically distinct patient groups. Acta Neuropathol. *129*, 679–693. https://doi.org/10.1007/s00401-015-1409-0.

83. Wick, W., Meisner, C., Hentschel, B., Platten, M., Schilling, A., Wiestler, B., Sabel, M.C., Koeppen, S., Ketter, R., Weiler, M., et al. (2013). Prognostic or predictive value of MGMT promoter methylation in gliomas depends on IDH1 mutation. Neurology *81*, 1515–1522. https://doi.org/10.1212/WNL.0b013e3182a95680.

84. Reuss, D.E., Sahm, F., Schrimpf, D., Wiestler, B., Capper, D., Koelsche, C., Schweizer, L., Korshunov, A., Jones, D.T.W., Hovestadt, V., et al. (2015). ATRX and IDH1-R132H immunohistochemistry with subsequent copy number analysis and IDH sequencing as a basis for an "integrated" diagnostic approach for adult astrocytoma, oligodendroglioma and glioblastoma. Acta Neuropathol. *129*, 133–146. https://doi.org/10.1007/s00401-014-1370-3.

85. Bady, P., Sciuscio, D., Diserens, A.C., Bloch, J., van den Bent, M.J., Marosi, C., Dietrich, P.Y., Weller, M., Mariani, L., Heppner, F.L., et al. (2012). MGMT methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. Acta Neuropathol. *124*, 547–560. https://doi.org/10.1007/s00401-012-1016-2.

86. Kim, D.W., Jang, H.Y., Ko, Y., Son, J.H., Kim, P.H., Kim, S.O., Lim, J.S., and Park, S.H. (2020). Inconsistency in the use of the term "validation" in studies reporting the performance of deep learning algorithms in providing diagnosis from medical imaging. PLoS One *15*, e0238908. https://doi.org/10.1371/journal.pone.0238908.

87. Cleaveland, S., Sharp, J., Abela-Ridder, B., Allan, K.J., Buza, J., Crump, J.A., Davis, A., Del Rio Vilas, V.J., de Glanville, W.A., Kazwala, R.R., et al. (2017). One Health contributions towards more effective and equitable approaches to health in low- and middle-income countries. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160168. https://doi.org/10.1098/rstb.2016.0168.

88. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet. Digit. Health *1*, e271–e297. https://doi.org/10.1016/s2589-7500(19)30123-2.

89. Moon, J.H., Hwang, H.W., Yu, Y., Kim, M.G., Donatelli, R.E., and Lee, S.J. (2020). How much deep learning is enough for automatic identification to be reliable? Angle Orthod. *90*, 823–830. https://doi.org/10.2319/021920-116.1.

90. Rudie, J.D., Rauschecker, A.M., Bryan, R.N., Davatzikos, C., and Mohan, S. (2019). Emerging applications of artificial intelligence in neuro-oncology. Radiology *290*, 607–618. https://doi.org/10.1148/radiol.2018181928.

91. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D.S.W., Karthikesalingam, A., King, D., Ashrafian, H., and Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit. Med. *4*, 65. https://doi.org/10.1038/s41746-021-00438-z.

92. Kuntz, S., Krieghoff-Henning, E., Kather, J.N., Jutzi, T., Höhn, J., Kiehl, L., Hekler, A., Alwers, E., von Kalle, C., Fröhling, S., et al. (2021). Gastrointestinal cancer classification and prognostication from histology using deep learning: systematic review. Eur. J. Cancer *155*, 200–215. https://doi.org/10.1016/j.ejca.2021.07.012.

93. Moses, L.E., Shapiro, D., and Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat. Med. *12*, 1293–1316. https://doi.org/10.1002/sim.4780121403.

94. Rutter, C.M., and Gatsonis, C.A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat. Med. *20*, 2865–2884. https://doi.org/10.1002/sim.942.

95. Macaskill, P. (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J. Clin. Epidemiol. *57*, 925–932. https://doi.org/10.1016/j.jclinepi.2003.12.019.

96. Whiting, P.F., Rutjes, A.W.S., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., Leeflang, M.M.G., Sterne, J.A.C., and Bossuyt, P.M.M.; QUADAS-2 Group (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. *155*, 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.

97. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Bmj *372*, n71. https://doi.org/10.1136/bmj.n71.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Original data	this paper	https://doi.org/10.57760/sciencedb.07885
Software and algorithms		
Original code	this paper	https://doi.org/10.57760/sciencedb.07886

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Wanyi Sun (sunwy1998@foxmail.com).

Materials availability

This study did not use any materials.

Data and code availability

- The original database is stored in the Science Data Bank to make publicly accessible. DOI is listed in the key resources table.
- The original code is stored in the Science Data Bank to make publicly accessible. DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Search strategy and eligibility criteria

The online database search consisted of Ovid-Medline, Embase, IEEE Xplore, Web of Science Core Collection, and the Cochrane Library for studies continuously published from 1st January 2015 until 16th August 2022. Keywords including 'glioma OR astrocytoma OR glioblastoma', 'DL OR AI NOT traditional ML, 'diagnosis OR classification OR grading', and 'performance OR (sensitivity and specificity) OR area under the curve' were used to explore pertinent studies. The detailed search strategies among these 5 databases were available in the supplemental information.

Any studies reporting the performance of DL algorithms on grading gliomas were included during the identification phase while only duplicates were removed. Further exclusion criteria were applied during the first round screening phase as follows: (1) studies published before January 1st 2015, when DL algorithms were not mature at this stage[56]; (2) non-clinical studies, reviews, letters, or comments; (3) studies to investigate the glioma segmentation; (4) investigations exclusively on brain tumor classifications and genetic or molecular subtypes of glioma, not related to glioma grading.

Eligibility assessment was then performed by two independent researchers (W.S. and C.S.) who had reviewed the titles and abstracts of all records, during which the full-text studies were reviewed and assessed in detail. Discrepancies were settled by a third senior researcher (P.X.). The studies using archived histopathological images, not in English, not using DL algorithms for classification, reporting no classifying outcomes, with no target disease, and no access were ruled out from the study while the remaining studies were included for systematic review and further meta-analysis.

Data extraction

Two independent researchers (W.S. and C.S.) reviewed the full-text articles (and supplemental information if available) and extracted study characteristics (patients' information, imaging modality, DL algorithms, etc.) and diagnostic performance of DL (true-positives, false-positives, true-negatives, and false-negatives) into a predetermined data extraction form. Conflicts were resolved through a team discussion and consensus. All classifications other than HGG and LGG were then converted into an exclusive binary classification of HGG and LGG to generate contingency tables for meta-analysis. The extracted data was used to calculate the pooled sensitivity, specificity, and area under the curve (AUC).

QUANTIFICATION AND STATISTICAL ANALYSIS

To assess the performance of DL algorithms to differentiate HGG from LGG, the definition of true positive (TP) was set for HGG while that of true negative (TN) was LGG. The included studies with inconsistent definition were redefined for our calculation. A hierarchical summary receiver operating characteristic (HSROC) curve with 95% confidence intervals (CI) and 95% prediction regions was employed to assess the overall performance of DL algorithms along with diagnostic parameters including pooled AUC, sensitivity, and specificity.[93–95] Given the inherent differences among the included studies, a bivariate random-effect model was implemented. Heterogeneity was estimated using the Higgins inconsistency index (I^2) statistic, of which 50% was defined as moderate and higher than 75% was defined as high respectively. Important variables affecting heterogeneity were assessed using meta-regression. The variables finally included in meta-regression analysis were: 1) sample size ($\leq$130/>130; 130 is the median of sample size); 2) data sharing (open data/private data); 3) type of internal validation (random split-sample validation/k-fold cross-validation); 4) transfer learning applied (applied/no applied); 5) data unit(image/patient); 6) classification (grade IV represented HGG/grade III+ IV represented HGG; According to the WHO classification standard and the actual classification of articles); 7) type of validation(internal/external). The first 5 variables did not differ in multiple DL performances in one study, so the meta-regression based on the highest accuracy pooled analysis was applied for these 5 factors. However, the rest 2 variables displayed diversely in one study, which requested the overall pooled analysis. Further subgroup analysis was performed by variables with statistically significant heterogeneity contribution. Meta-analysis was only conducted only when the number of studies was equal to or greater than three. Data analysis was conducted by STATA (version 15.1) software and the MIDAS module was used. The p value less than 0.05 was considered statistically significant. The original data and code were deposited at Science Data Bank and were publicly available (key resources table).

Quality assessment

The risk of bias and applicability concerns of the included studies were assessed by two researchers (W.S. and C.S.) using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool,[96] which allows for more transparent rating of bias and applicability of diagnostic accuracy studies. QUADAS-2 tool consists of four key domains: patient selection, index test, reference standard, flow and timing. Publication bias was assessed by a funnel plot.

ADDITIONAL RESOURCES

The study was registered with the open-access PROSPERO International prospective register of systematic reviews (CRD42022360385). The study was performed strictly following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement.[97] Both ethical approval and informed consent were not applicable since this study was a secondary analysis of publicly available data.