Online Figures and Tables

Online Figures

Online Figure 1: Search strategy used on bibliographic databases.

Online Figure 2: Forest plot of highest AUC of every study in validation and meta-analysis.

Online Figure 3: Performance of radiologists versus, and with, ML models.

Online Tables

Online Table 1: Custom-built data extraction form.

Online Table 2: TRIPOD items and adherence indices.

Online Table 3: Pipeline characteristics of individual studies.

Online Table 4: Summary of aims and ML performance per study.

Online Table 5: Results of PROBAST risk of bias (ROB) assessment per domain.

Embase <1974 to 2021 January 29>
1 exp Artificial Intelligence/ 45007
2 machine learning/ 37007
3 deep learning/ 12393
4 ((artificial* or machine* or deep*) adj3 (intelligence or learning)).tw,kw. 73708
5 AI.ti,ab. 39008
6 exp computer assisted diagnosis/ 1169775
7 computer* assist* diagnosis.tw,kw. 937
8 radiomics/ 1903
9 radiomic*.tw,kw. 4867
10 or/1-9 1305625
11 exp nuclear magnetic resonance imaging/ 1001848
12 (Magnetic Resonance Imag* or MR-Imag* or MR Imag or MRI* or NMR).tw,kw. 882389
13 11 or 12 1278307
14 exp glioma/ 139715
15 glioma*.tw,kw. 84577
16 (glial adj2 (tumor* or tumour*)).tw,kw. 3616
17 (glioblastoma* or astrocytoma* or astrocytic glioma* or astroglioma).tw,kw. 77347
18 or/14-17 165100
19 10 and 13 and 18 9560
20 limit 19 to yr="2020 - 2022" 771

Ovid MEDLINE(R) ALL <1946 to January 29, 2021>
1 exp Artificial Intelligence/ 106412
2 ((artificial* or machine* or deep*) adj3 (intelligence or learning)).tw,kw. 55350
3 AI.ti,ab. 28603
4 exp Image Interpretation, Computer-Assisted/ 551508
5 computer* assist* diagnosis.tw,kw. 626
6 radiomic*.tw,kw. 3204
7 or/1-6 706556
8 exp Magnetic Resonance Imaging/ 465045
9 (Magnetic Resonance Imag* or MR-Imag* or MR Imag or MRI*).tw,kw. 435821
10 8 or 9 625065
11 exp Glioma/ 85314
12 glioma*.tw,kw. 60258
13 (glial adj2 (tumor or tumour)).tw,kw. 831
14 (glioblastoma* or astrocytoma* or astrocytic glioma* or astroglioma).tw,kw. 51784
15 or/11-14 115828
16 7 and 10 and 15 4493
17 limit 16 to yr="2020 - 2021" 260

Cochrane CENTRAL (trials)
ID Search Hits
#1 MeSH descriptor: [Artificial Intelligence] explode all trees 1040
#2 (artificial* OR machine* OR deep*) AND (intelligence OR learning) 3131
#3 AI 7937

#4 MeSH descriptor: [Image Processing, Computer-Assisted] explode all trees 3582
#5 computer* assist* diagnosis 6489
#6 radiomic* 210
#7 #1 OR #2 OR #3 OR #4 OR #5 OR #6 20843
#8 MeSH descriptor: [D008279] explode all trees 0
#9 Magnetic Resonance Imag* OR MR-Imag* OR MR Imag OR MRI* OR NMR 36332
#10 #8 OR #9 36332
#11 MeSH descriptor: [Glioma] explode all trees 1197
#12 glioma* 1792
#13 (glial AND (tumor OR tumour)) 70
#14 glioblastoma* OR astrocytoma* OR astrocytic glioma* OR astroglioma 2432
#15 #11 OR #12 OR #13 OR #14 3580
#16 #7 AND #10 AND #15 with Publication Year from 2020 to 2021, in Trials 2

Web of Science
13
235
#12
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=2020-2021

12
711
#11 AND #7 AND #6
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

11
132,043
#10 OR #9 OR #8
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

10
75,253
TS=(glioblastoma* or astrocytoma* or astrocytic glioma* or astroglioma)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

9
2,783
TS=(glial NEAR/2 (tumor* or tumour*))
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

8
91,055
TS=(glioma*)

Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

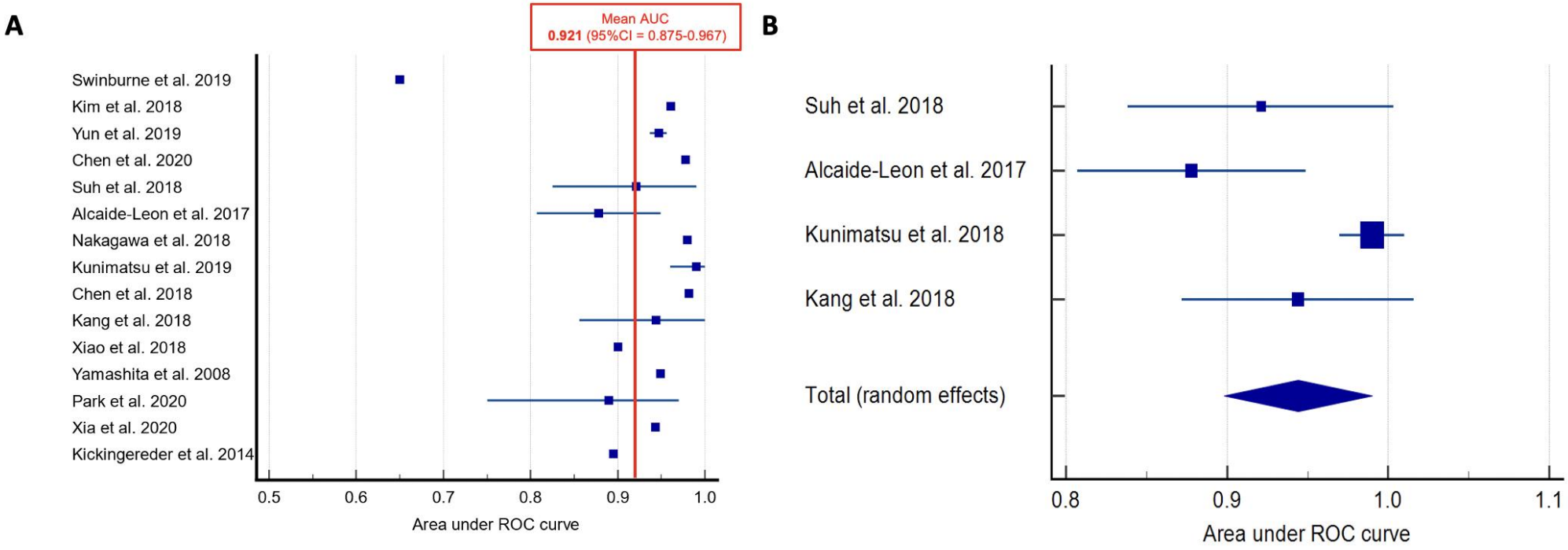
7
1,046,719
TS=(Magnetic Resonance Imag* or MR-Imag* or MR Imag or MRI* or NMR)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

6
331,970
#5 OR #4 OR #3 OR #2 OR #1
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

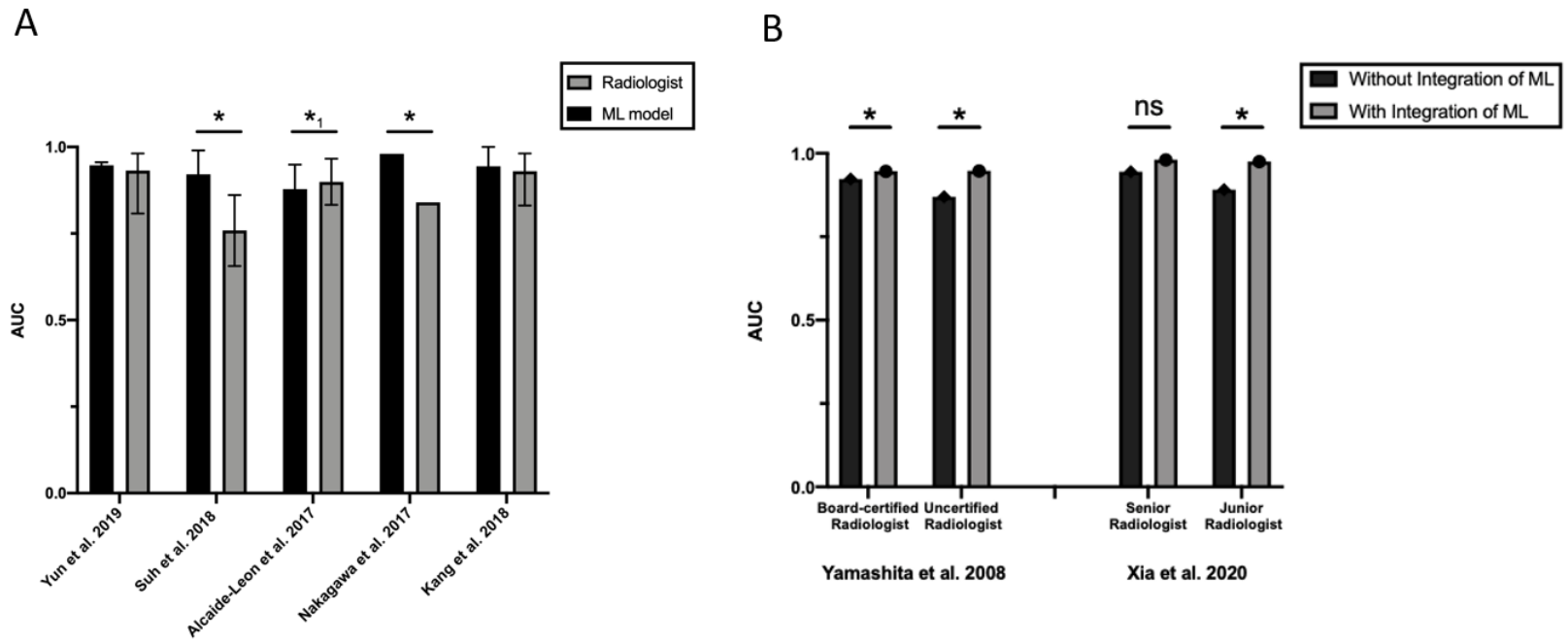
5
4,619
TS=(radiomic*)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
# 4
6,845
TS=(computer* assist* diagnosis)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

3
51,251
AB=(AI)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years
# 2
11,596
TI=(AI)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years
# 1
276,118
TS=((artificial* or machine* or deep*) NEAR/3 (intelligence or learning))
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
Timespan=All years

Online Figure 1. Search strategy used on bibliographic databases.



Online Figure 2: A) Forest plot showing the highest AUC of every study in validation. The center dot represents the highest AUC in a validation set reported by the authors in the study, and the whiskers the 95%CI (in cases where it was reported). The mean AUC among all the studies was 0.921 (95%CI = 0.875-0.967), and is represented by the vertical red line. **B) Forest plot of the meta-analysis.** The first four rows show the AUC and 95% CI achieved by the studies included in the meta-analysis. The last row shows the pooled AUC of 0.944 (95%CI = 0.918-0.98) calculated using a random-effects model.



Online Figure 3: A) Comparison of the performance of ML models with radiologists. Plotted are the best performing ML model alongside the best performing radiologist in the same validation set. The top of the bar represents the mean AUC, and the whiskers the 95%CI (in cases where it was reported). **B) Performance of radiologists before and after incorporation of ML pipeline in their decision process.** Both studies compared experienced radiologists with their more novice counterparts. Xia et al. compared the performance of a single junior versus senior radiologist. Yamashita et al. reported and compared the average performance of several board-certified and uncertified radiologists.

*= Significant difference ($p<0.05$) was reported by the authors of the study.

$*_1$= Alcaide-Leon et al. reported that ML was significantly non inferior to radiologists.

ns= difference of mean AUC was reported as being not significant ($p>0.05$).

Sheet 1 (Pipeline characteristics of individual studies):

Category	Information	Explanation
General information	**Title**	Enter title of study.
	Author	Enter name of first author.
	Year	Enter year of publication.
Dataset characteristics	**Total number of patients**	Enter the total number (#) of patients used in the study, both for training and validation.
	Fraction of patients used for training	Number of patients used for training divided by total # of patients used in the study. If only k-fold-cross-validation, specify k.
	Fraction of patients used testing	Number of patients used for testing divided by total # of patients used in the study. If only k-fold-cross-validation, specify k.
	Was external validation used?	Enter either "Yes" or "No". If "Yes", then specify if geographical or temporal external validation.
	Glioma/PCNSL ratio	= #Patients with glioma / #Patients with PCNSL.
	Immune status of PCNSL patients	Specify whether included PCNSL patients are immunosuppressed (IS) or immunocompetent (IC). If information not available note "Not specified".
	Source of data	- private single center, - private multi-center, - public dataset (e.g BraTS, TCIA).
Tumor and Ground-Truth	**Tumor types studied**	List the different tumor types studied. Specify the type of glioma.
Tumor and Ground-Truth Machine Learning characteristics	**Ground-truth diagnosis**	Specify the method for ground-truth diagnosis of glioma and PCNSL, and what proportion of patients were diagnosed with that method: (E.g.: "Histopathology (100%)")
	Overall type of Machine Learning	Specify whether

		- classical Machine Learning - Deep Learning - both
Machine Learning characteristics	**Supervision**	Specify whether - Unsupervised Learning - Supervised Learning
	Classification algorithm	Enter the classifier algorithm used (E.g.: Logistical Regression, Random Forest etc.).
	Type of features used	Specify the type of features by the algorithm. (E.g.: "First order and texture features").
	Number of features used in the final model	Specify the final number of features used in the final model.
Imaging characteristics	**Overall imaging technique**	Specify whether - MRI - CT - PET
Imaging characteristics	**If MRI, which field strength?**	Specify whether - 1.5 T - 3 T - 7 T
	If MRI, which sequence was used for features?	Specify whether - T1c+ - T1 - T2 - FLAIR - DWI/ADC - DSC - DCE - ASL - MRS
	If PET, which tracer?	- Specify the tracer used

Sheet 2 (Model performance metrics with examples from Chen et al. 2020)

Author, Year	Classifier + Feature selection method	Sequence of features	Prediction of…	Training, internal validation, external validation?	AUC (95%CI if available)	Accuracy (95%CI if available)	Sensitivtiy (95%CI if available)	Specificity (95%CI if available)
Chen, 2020	LDA + RF	T1c+,	GBM vs PCNSL	Training	0.97	0.968	0.935	0.99
Chen, 2020	LDA + RF	T1c+,	GBM vs PCNSL	Internal validation	0.964	0.957	0.906	0.99

Online Table 1: Custom built data extraction form. The custom-built form that we used for data extraction was developed in Microsoft Excel and consisted of two sheets. The first sheet collects the information on the pipeline characteristics of the individual studies, while the second sheet is useful for extracting individual model performance metrics for every test of every developed model by the researchers.

Section - Topic	Item number	Explanation	Adherence index(%)
Title	Item 1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	0
Abstract	Item 2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	0
Background and Objectives	Item 3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	78.3
	Item 3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	100
Methods – Source of data	Item 4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	82.6
	Item 4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	78.3
Methods – Participants	Item 5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centers.	60.9
	Item 5b	Describe eligibility criteria for participants.	65.2
	Item 5c	Give details of treatments received, if relevant.	87
Methods – Outcome	Item 6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	100
	Item 6b	Report any actions to blind assessment of the outcome to be predicted.	100
	Item 7a	Clearly define all predictors used in developing or validating the	87

Methods – Predictors		multivariable prediction model, including how and when they were measured.	
	Item 7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	0
Methods – Sample size	Item 8	Explain how the study size was arrived at.	47.8
Methods – Missing data	Item 9	Describe how missing data were handled (e.g., complete case analysis, single imputation, multiple imputation) with details of any imputation method.	34.8
Methods – Statistical analysis methods	Item 10a	Describe how predictors were handled in the analyses.	56.5
	Item 10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	13
	Item 10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	0
Methods –Risk groups	Item 11	Provide details on how risk groups were created, if done.	Not applicable in any study.
Results – Participants	Item 13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	73.9
	Item 13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	0
Results – Model development	Item 14a	Specify the number of participants and outcome events in each analysis.	100
	Item 14b	If done, report the unadjusted association between each candidate predictor and outcome.	22.2 (Applicable in 21 studies)

Results – Model specification	Item 15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	4.3
	Item 15b	Explain how to use the prediction model.	17.4
Results – Model performance	Item 16	Report performance measures (with confidence intervals) for the prediction model.	0
Discussion – Limitations	Item 18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	91.3
Discussion – Interpretation	Item 19b	Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence.	100
Discussion - Implications	Item 20	Discuss the potential clinical use of the model and implications for future research.	78.3
Other information	Item 22	Give the source of funding and the role of the funders for the present study.	13

Online Table 2: TRIPOD items for a development model, as described by Collins et al. These items are made up of several elements. An item is scored with the score of 1 if all elements pertaining to it are reported. If one element is not reported, then the whole item is scored as 0. The adherence index is calculated by the number of times the item was fully reported divided by the number of studies. Item 21 is not reported ("Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets."), as it should not be included in overall scoring.

Of important note: It is important that the reader recognizes that some elements in items with an overall adherence index of 0% were correctly reported but never all of them, leading to the overall low adherence. If the reader wishes to receive a breakdown for one of the TRIPOD elements, please reach out to the corresponding author and we will happily provide the information.

	Dataset characteristics							Tumor and Ground-truth		ML characteristics				
Study	**Number of patients included[1]**	**Source of data**	**Glioma / PCNSL case ratio**	**Immune status of PCNSL patients**	**% used for training**	**% used for validation**	**Was an external validation done?**	**Tumor types studied**	**Gold-standard for diagnosis (%)**	**ML or DL ?**	**Algorithms studied**	**Type of features used**	**Number of features used for ML model**	**MRI *field strength* and sequences performed on patients**
Swinburne et al. 2019 [34]	17	Single center	1.13	Not specified	Leave-one-out-cross-validation (LOOCV)		No	GBM, CNSL, Metastasis	Pathology (100%)	ML and DL	SVM, MLP	Perfusion and Diffusion metrics	1 feature per unique experiment (14 in total)	*3 T* T1c+, T2, FLAIR, DWI, DSC, DCE
Kim et al. 2018 [28]	143	Multi-center	1.32	Not specified	0.6	0.4	Yes (Geo)	GBM, typical and atypical PCNSL	Pathology (100%)	ML	SVM, LogReg, RF	Shape, First-Order (FO), Texture matrices (TM)	15	*3 T* T1c+, T2, FLAIR, DWI
Yun et al. 2019 [41]	195	Multi-center	1.57	Not specified	0.75	0.25	Yes (Geo)	GBM, PCNSL	Pathology (100%)	ML and DL	SVM, RF, GLM, MLP, CNN	FO, TM, Wavelet transformed (WT)	10	*3 T* T1c+, DWI
Chen et al. 2020 [23]	138	Single center	1.23	Not specified	0.8	0.2	No	GBM, PCNSL	Pathology (100%)	ML	SVM, LogReg, LDA	Shape, FO, TM	Median 8 (1-16 per selection method)	*3 T* T1c+
Suh et al. 2018 [33]	77	Single center	0.43	Not specified	10-fold-cross-validation (10-fold-x-validation)		No	GBM (typical and atypical), PCNSL	Pathology (100%)	ML	RF	Shape, FO, TM, WT	80	*3 T* T1c+, T2, FLAIR, DWI
Alcaide-Leon et al. 2017 [21]	106	Single center	2.03	32 IC 2 IS	10-fold-x-validation		No	WHO III glioma, GBM, PCNSL	Pathology (100%)	ML	SVM	FO, TM	Feature number after selection not specified (153 before selection)	*1.5 T and 3 T* T1c+
Nakagawa et al. 2017 [30]	70	Not specified	1.8	Not specified	10-fold-x-validation		No	GBM, PCNSL	Pathology (100%)	ML	XGBoost, uvLogReg	FO, TM	48	*3 T* T1c+, T2, DWI DSC
Kunimatsu et al. 2019 [29]	76	Single center	2.62	Only IC	0.79	0.21	Yes (Temp)	GBM, PCNSL	Pathology (100%)	ML	SVM	FO, TM	4	*3 T* T1c+

Chen et al. 2018 [24]	96	Not specified	2.2	Not specified	0.67	0.33	No	GBM, PCNSL	Pathology (100%)	ML	SVM	Scale-invariant-feature-transform (SIFT)	496	*Not reported* T1c+
Kang et al. 2018 [26]	196	Multi-center	1.3	Only IC	0.55	0.45	Yes (Geo)	GBM, PCNSL (typical, atypical)	Pathology (100%)	ML	NB, RF, LDA, DT, kNN, AdaBoost	Shape, FO, TM, WT	55	*3 T* T1c+, FLAIR, DWI, DSC
Shrot et al. 2019 [32]	53	Single center	3.4	Not specified	LOOCV		No	GBM, PCNSL, Metastais, Meningioma	Pathology (100%)	ML	Binary tree with SVM in nodes	Intensity, Morphology, Diffusion and Perfusion metrics	20	*1.5 T and 3 T* T1c+, T2, FLAIR, DWI, DSC
Xiao et al. 2018 [36]	82	Single center	2.73	Only IC	10-fold-x-validation		No	GBM, PCNSL	Pathology (100%)	ML	SVM, LogReg, RF, NB	FO, TM	3	*1.5 T and 3 T* T1c+, T2
Yamashita et al. 2008 [39]	107	Single center	7.92	Not specified	LOOCV		No	LGG, HGG, PCNSL, Metastasis	Pathology (100%)	DL	ANN (MLP)	Clinical, MR features (such as oedema, hemorrhage etc.)	15	*1.5 T* T1c+, T2
Yamasaki et al. 2013 [37]	40	Not specified	1	Not specified	0.05-0.95	0.05-0.95	No	GBM (typical, atypical), PCNSL (typical, atypical)	Pathology (100%)	ML	SVM	Luminance histogram range, ADC value	2	*Not reported* T1c+, DWI
Park et al. 2020 [31]	259	Multi-center	1.74	Not specified	0.83	0.17	Yes (Geo)	GBM, PCNSL, Metastasis	Pathology (100%)	DL	CNN	Temporal Patterns of Time-Signal Intensity Curves from DSC	9	*3 T* T1c+, T2, FLAIR, DSC
Xia et al. 2020 [35]	240	Single center	1.16	Not specified	Cross-vendor (cv): 0.621 Mixed vendor (mv): 0.8	cv: 0-379 mv: 0.2	Yes (Temp)	GBM, PCNSL	Pathology (100%)	ML	LogReg, GLM	Shape, FO, TM, WT	16	*3 T* T1c+, FLAIR, DWI

Bao et al. 2019 [22]	20	Single center	1.22	Not specified	100	0	No	Non-hemorrhagic GBM and PCNSL	Pathology (100%)	ML	LogReg	ADC-, and CBV derived metrics	2	*3 T* T1, T1c+, T2, FLAIR, DWI, DSC
Eisenhut et al. 2020 [25]	74	Single center	1	Not specified	100	0	No	GBM, PCNSL	Pathology (100%)	ML	LogReg	ADC-, and CBV derived metrics	5	*1.5 T and 3 T* T1, T1c+, T2, SWI, FLAIR, DWI, DSC
Kickingereder et al. 2014 [27]	47	Single center	1.47	Only IC	LOOCV		No	Atypical GBM, PCNSL	Pathology (100%)	ML	LogReg	ADC-, CBV- and SWI-derived metrics	3	*3 T* T1, T1c+, T2, SWI, DWI, DSC
Wang et al. 2011 [43]	42	Single center	1.65	IC and IS	LOOCV		No	GBM, PCNSL	Pathology (100%)	ML	DT	CBV-, and DTI- derived metrics	5	*3 T* T1, T1c+, FLAIR, DTI, DSC
Zhou et al. 2018 [42]	92	Not specified	1.3	Only IC	100	0	No	GBM, PCNSL	Pathology (100%)	ML	LogReg	18F-FDG PET derived metrics	2	*No MRI performed* 18F-FDG PET/CT
Yamashita et al. 2016 [38]	50	Not specified	1.94	Not specified	100	0	No	GBM, PCNSL	Pathology (96%) and clinico - radiological data (4%)	ML	LogReg	IVIM, ADC- and 18F-FDG PET derived metrics	2	*3 T* T1, T1c+, 18F-FDG PET/CT, IVIM
Yamashita et al. 2013 [40]	56	Not specified	1.94	Not specified	100	0	No	GBM. PCNSL	Pathology (93%) and clinico - radiological data (7%)	ML	LogReg	ASL-, ADC- and 18F-FDG PET derived metrics	2	*3 T* T1, T1c+, T2, FLAIR, 18F-FDG PET/CT, DWI, ASL

[1]= Patients with tumors other than PCNSL or gliomas were not counted

Online Table 3. Pipeline characteristics of individual studies. Abbreviations: SVM=support vector machines; Geo = Geographical External Validation; MLP = Multilayer Perceptron Neural Network; LogReg = Logistic Regression; RF = Random Forests; GLM = Generalized Linear Model; CNN = Convolutional Neural Network; LDA = Linear Discriminant Analysis; XGBoost = eXtreme Gradient Boosting; uvLogReg = univariate Logistic Regression; Temp = Temporal External Validation; NB = Naïve Bayes; DT = Decision Tree; kNN = k-nearest neighbors; ANN = Artificial Neural Network; AdaBoost = Adaptive Boosting; IC= Immuno-competent; IS= Immuno-suppressed; DWI = Diffusion Weighted Imaging; DTI = Diffusion Tensor Imaging; DSC = Dynamic Susceptibility Contrast-enhanced imaging; ASL = Arterial Spin Labeling imaging; ADC = Apparent Diffusion Coefficient; CBV = Cerebral Blood Volume; SWI = Susceptibility Weighted Imaging; IVIM = Intravoxel Incoherent Motion MR imaging.

	Aim of the study	Performance		
Study		**Training dataset**	**Internal validation**	**External validation**
Swinburne et al. 2019 [34]	Classification with the help of ADC and perfusion derived metrics		**SVM**: AUC of 0.63; Accuracy of 58.8 % ***MLP****:* AUC of 0.65; Accuracy of 64.7 %	
Kim et al. 2018 [28]	Classification using conventional radiomic features		**SVM**: AUC of 0.987; Accuracy of 94.1 % (Sensitivity = 93.8 %, Specificity = 94.4 %) **Multivariate LogReg**: AUC of 0.991; Accuracy of 94.1 % (Sens. = 95.8 %, Spec. = 91.7 %) ***RF****:* AUC of 1; Accuracy of 100 % (Sensitivity and Specificity both 100 %)	**Geographical External Validation:** **SVM**: AUC of 0.947; Accuracy of 91.2% (Sensitivity= 93.1 %, Specificity = 89.29 %) ***Multivariate LogReg****:* AUC of 0.961; Accuracy of 87.7% (Sens.= 89.7%, Spec. = 85.7%) **RF**: AUC of 0.953; Accuracy of 84.2 % (Sensitivity = 96.6% and Specificity 71.43%)
Yun et al. 2019 [41]	Compare classification performance of i)radiomics features + conventional ML ii) radiomics features + MLP iii) radiologists, and an iv) End-to-End CNN classifier	**GLM boosting**: AUC ± 95%CI of 0.943 ± 0.927-0.978 (Acc. = 94.3%, Sens = 96.3 % and Spec= 92.3 %) **SVM**: AUC of 0.934 **RF**: AUC of 0.927 ***MLP***: AUC (95%CI) of 0.994 (0.994-0.995) (Sens = 100 % and Spec= 100 %) **CNN**: AUC (95%CI) of 0.973 (0.966-0.980) (Sens = 100% and Spec = 94.5%) **Radiologist**: AUC (95%CI) of 0.908 (0.755-0.949) (Sens = 83.9% and Spec = 97.8%)	**GLM boosting**: AUC (95%CI) of 0.931 (0.914-0.941) (Sens = 98.8 % and Spec= 92.3 %) ***MLP****:* AUC (95%CI) of 0.991 (0.987-0.984) (Sens = 100 % and Spec = 100 %) **CNN**: AUC (95%CI) of 0.879 (0.856-0.902) (Sens = 83.3% and Spec = 83.3%) **Radiologist**: AUC (95%CI) of 0.875 (0.653-0.940) (Sens = 83.3 % and Spec = 100 %)	**Geographical External Validation** **GLM boosting**: AUC (95%CI) of 0.811 (0.795-0.835) (Sens = 85.5 % and Spec= 78.9 %) ***MLP****:* AUC (95%CI) of 0.947 (0.937-0.956) (Acc. = 85.7%, Sens = 92.9% and Spec= 82.1%) **CNN**: AUC (95%CI) of 0.486 (0.468-0.503) (Sens = 100 % and Spec = 35.7 %) **Radiologist**: AUC (95%CI) of 0.932 (0.808-0.981) (Sens = 89.7 % and Spec = 96.4 %)
Chen et al. 2020 [23]	Classification using conventional radiomic features	***LDA****:* AUC of 0.992 and Accuracy of 99.3% (Sens. = 99.6% and Spec. = 99%) **SVM**: AUC of 0.957 and Accuracy of 96.2% (Sens. = 99.8% and Spec.= 93.4%) **Multivariate LogReg**: AUC of 0.959 and Accuracy of 98.8% (Sens. = 94.2% and 98.1%)	***LDA****:* AUC of 0.978 and Accuracy of 97.9% (Sens. = 98.2% and Spec. = 97.6%) **SVM**: AUC of 0.959 and Accuracy of 96.4% (Sens. = 99.7% and Spec.= 94.3%) **Multivariate LogReg**: AUC of 0.975 and Accuracy of 96.6% (Sens. = 97.5% and 96.4%)	
Suh et al. 2018 [33]	Classification of atypical GBM and PCNSL using conventional radiomic features and comparison to radiologists, and ADC 10th percentile		***RF****:* AUC (95%CI) of 0.921 (0.825-0.99), and Accuracy of 89.6% (Sens. = 91.3%, Spec.=88.9%) **Radiologists**: AUC ± 95%CI of 0.759 ± 0.656-0.861 **ADC_{p10}**[1,2]: AUC ± 95%CI of 0.684 ± 0.560-0.890	
Alcaide-Leon et al. 2017 [21]	Classification using conventional radiomic		**Radial-kernel SVM:** AUC (95%CI) of 0.878 (0.807-0.949)	

	features and comparison to radiologists		***Radiologists**:* AUC (95%CI) of 0.899 (0.833-0.966)	
Nakagawa et al. 2018 [30]	Classification using conventional radiomic features from and comparison to radiologists		***Multivariate XGBoost**:* AUC of 0.98 **Univariate LogReg on rCBV**: AUC of 0.86 **Radiologists:** AUC of 0.84	
Kunimatsu et al. 2019 [29]	Classification using conventional radiomic features		***Gaussian kernel SVM:*** AUC (95%CI) of 0.99 (0.96-1), and Accuracy of 80% **Linear kernel SVM:** AUC (95%CI) of 0.87 (0.77-0.95), and Accuracy of 70%	**Temporal External Validation** ***Linear and Gaussian SVM:*** Accuracy of 75%
Chen et al. 2018 [24]	Classification using SIFT features	***SVM**:* AUC of 0.991 and Accuracy of 95.3% (Sens.= 85% and Spec.= 100%)	***SVM**:* AUC of 0.982 and Accuracy of 90.6% (Sens.= 80% Spec.= 95.5%)	
Kang et al. 2018 [26]	Comparison of classification using radiomic features on conventional and diffusion MRI, and comparison to radiologists and 10^{th} percentile of ADC and 90^{th} percentile of CBV	**Radial SVM**: AUC of 0.968 on ADC features **Linear SVM**: AUC of 0.979 on ADC features ***RF**:* AUC of 0.983 on ADC features **LDA**: AUC of 0.982 on ADC features **DT**: AUC of 0.927 on T1c+ features **NB**: AUC of 0.955 on ADC features **kNN**: AUC of 0.968 on ADC features **AdaBoost**: 0.979 on ADC features	***RF*** with ADC radiomics: AUC (95%CI) of 0.984 (0.945-1) (Sens.= 80.9%, Spec.=100%) **LDA** on T1c+ radiomics: AUC (95%CI) of 0.968 (0.913-1) (Sens.= 85.7%, Spec.=95.2%) **ADC_{10}[1]**: AUC (95%CI) of 0.787 (0.633-0.898) (Sens.= 95.2% , Spec.= 57.1%) **CBV_{90}[1]** : AUC (95%CI) of 0.905 (0.774-0.973) (Sens. = 80.9%, Spec.= 90.5%) **Radiologists** : AUC (95%CI) of 0.908 (0.755-0.949) (Sens.= 83.9% Spec.=97.8%)	**Geographical External Validation** ***RF*** with ADC radiomics: AUC (95%CI) of 0.944 (0.856-1) (Acc.= 88.6%, Sens.= 85.7%, Spec.=75%) **LDA** on T1c+ radiomics: AUC (95%CI) of 0.819 (0.617-0.967) (Acc.= 78.6% Sens.= 71.4%, Spec.=82.1%) **ADC_{10}[1]**: AUC (95%CI) of 0.809 (0.683-0.901) (Sens.= 75.9% , Spec.= 82.1%) **Radiologists** : AUC (95%CI) of 0.930 (0.831-0.981) (Sens.= 89.7% Spec.=96.7%)
Shrot et al. 2019 [32]	Classification with morphological features; and diffusion, and perfusion metrics		***Binary hierarchical tree with SVM nodes**:* Accuracy for GBM of 95.7% and for PCNSL of 93.6%. Pairwise classification achieved for PCNSL vs GBM Sens.= 100% and Spec.= 100%	
Xiao et al. 2018 [36]	Classification using conventional radiomic features		***NB**:* AUC of 0.9, and Accuracy of 82% **SVM**: AUC of 0.87, and Accuracy of 88% **Trivariate LogReg**: AUC of 0.85, and Accuracy of 84%	
Yamashita et al. 2008 [39]	Classification using clinical and qualitative features, and comparison to Radiologists		***ANN (MLP)**:* AUC of 0.949 **Board-certified radiologsist**: Without ANN assistance → average AUC of 0.923 (Acc. = 87.9%, Sens.= 80.8%, Spec.= 90.3%). With ANN assistance → average AUC of 0.946	

			(Acc. = 91.5%, Sens.= 86.8%, Spec.= 93.1%) Difference was not sigificant **Precertification radiologists:** Without ANN assistance → average AUC of 0.87 (Acc. = 85.6%, Sens.= 75.6%, Spec.= 89%). sWith ANN assistance → average AUC of 0.947 (Acc. = 92.1%, Sens.= 87.5%, Spec.= 93.7%) Difference was sigificant	
Yamasaki et al. 2013 [37]	Classification using ADC and Luminance-range histogram (LRH) thresholding as features		***SVM using ADC and LRH**:* Accuracy = 95.4% **SVM** using only LRH: Accuracy = 83.3% **ADC thresholding** alone: Accuracy = 66% **LRH thresholding** alone: Accuracy = 75%	
Park et al. 2020 [31]	Classification using Time-Signal Intensity Patterns derived from DSC imaging by Autoencoder Neural Network	***CNN**:* AUC (95% CI) of 0.921 (0.860-0.951) (Sens.= 93.6%, Spec.= 81%)	***CNN**:* AUC (95% CI) of 0.93 (0.821-0.983) (Sens.= 85.7%, Spec.= 90.9%)	**Geographical External Validation** ***CNN**:* AUC (95% CI) of 0.89 (0.75-0.97) (Sens.= 95.2%, Spec.= 76.5%)
Xia et al. 2020 [35]	Classification using conventional radiomic features, and cross-MRI-vendor validation			**Temporal External Validation** Cross-vendor validation: **Single-sequence GLM**: AUC of 0.937 (Acc.= 89%, Sens. = 87%, Spec.= 0.911) **Multivariate LogReg**: AUC of 0.943 (Acc.= 91.2% , Sens.= 89.1% , Spec.= 93.3%) **Junior (Senior) Radiologist without assistance**: AUC of 0.891 (0.945), Acc.= 89% (94.5%), Sens.= 80.4% (91.3%), Spec.= 97.8% (97.8%) ***Junior (Senior) Radiologist with assistance:*** AUC of 0.975 (0.980), Acc.= 95.6% (95.6%), Sens.= 93.5% (93.5%), Spec.= 97.8% (97.8%) No significant difference to mixed-vendor validation
Bao et al. 2019 [22]	Classification using nCBV-, and ADC- derived metrics in non-hemorrhagic tumors	***Bivariate LogReg**:* AUC of 0.969 (Sens.= 88.9%, Spec.= 90.9%) **$nCBV_{mean}$[1,2] :** AUC of 0.869 (Sens.= 72.7%, Spec.= 88.9%) **ADC_{p25}[1]:** AUC of 0.838 (Sens.= 72.7%, Spec.= 88.9%)		
Eisenhut et al. 2020 [25]	Classification using CBV-, and ADC- derived features	***Bivariate LogReg:*** AUC of 1 (Acc. = 100% Sens.= 100%, Spec.= 100%)		

		rCBV[1,2]: AUC of 0.93 (Acc.= 86%) **ADC_{max}[1,2]**: AUC of 0.847 (Acc.=80%)		
Kickingereder et al. 2014 [27]	Classification of only atypical GBM vs typical PCNSL using combination of ADC-, rCBV-, and SWI-derived metrics.		**Trivariate LogReg on ADC, rCBV, and SWI**: Sensitivity of 96% for GBM (95% for PCNSL) **Univariate LogReg on ADC:** AUC of 0.895, and sensitivity of 82% for GBM (74% for PCNSL) **Univariate LogReg on rCBV**: AUC of 0.887, and sensitivity of 82% for GBM (74% for PCNSL)	
Wang et al. 2011 [43]	Classification GBM, PCNSL and Mets using a two-level decision tree with DTI- and DSC- derived metrics.		***DT:*** 1 .Layer (GBM vs Non-GBM) AUC of 0.938 (Acc.= 89.6%, Sens.= 89%, Spec.=93%) 2. Layer (PCNSL vs. Metastasis) AUC for 0.909 (Acc.= 81.6%, Sens.= 77%, Spec.= 94%) Overall Accuracy: GBM 84.6%, PCNSL 75%	
Zhou et al. 2018 [42]	Classification using 18F-FDG PET/CT	***Bivariate LogReg on SUV_{max} and T/N ratio:*** AUC of 0.923 (Sens.= 88.5%, Spec.= 82.7%) **SUV_{max} [1]**: AUC of 0.91 (Acc.= 84.6%, Sens.= 76.9%, Spec.= 92.3%)		
Yamashita et al. 2016 [38]	Classification using IVIM-, 18F-FDG PET-, and ADC_{min}- derived features	***Bivariate LogReg on f_{max} and D_{min}:*** AUC of 0.936 **D_{min} [1,2]:** AUC of 0.905 (Acc.= 83.3%, Sens.= 82.8%, Spec.= 84.6%) **SUV_{max} [1,3]:** AUC of 0.857 (Acc.= 88.1%, Sens.= 89.7%, Spec.= 84.6%) **ADC_{min} [1,2]:** AUC of 0.894 (Acc.= 83.3%, Sens.= 82.8%, Spec.= 84.6%)		
Yamashita et al. 2013 [40]	Classification using ASL-, ADC-, and DSC- derived features.	**Bivariate LogReg using rTBF and ADC_{min}:** AUC of 0.706 ***Absolute TBF* [1,2]:** AUC 0.888 (Acc.= 83%, Sens.= 83.3%, Spec.= 82.9%) **SUV_{max} [1,3]:** AUC of 0.848 (Acc.= 83.8%, Sens.= 92.3%, Spec.= 79.2%) **ADC_{min} [1,2]:** AUC of 0.768 (Acc.= 77.8%, Sens.= 92.8%, Spec.= 79.2%)		

[1]= Assessed by selecting a cut-off and performing an ROC analysis
[2]= Value was reported to be significantly higher in GBM compared to PCNSL ($p < 0.05$)
[3]= Value was reported to be significantly higher in PCNSL compared to GBM ($p < 0.05$)

Online Table 4: Summary of aims and ML performance per study. The aims of every individual study are detailed in the first column. The performance columns show the best performance of every ML algorithm for every study and for every type of test (training, internal, or external validation). Best performance was assessed in reference to the AUC. The best performing model overall is highlighted in cursive font. In cases where performance of several radiologists was assessed, we report the results of the best-performing radiologist only. Only exception to this is the study by Yamashita et al. 2008, as they provided an average AUC. If the classification performance of single parameters (e.g SUV_{max}, ADC_{min}) were assessed without using ML, but alongside a ML-algorithm, we also included those performance metrics for completeness.

	Domain 1 (Participants)[1]	Domain 2 (Predictors)	Domain 3 (Outcomes)	Domain 4 (Analysis) [2]	Overall[3]
High ROB	21.74% (n=5)	0%	0%	47.8% (n=11)	69.6% (n=16)
Unclear ROB	47.83% (n=11)	0%	0%	52.2% (n=12)	30.4% (n=7)
Low ROB	30.43% (n=7)	100% (n=23)	100% (n=23)	0%	0%

Online Table 5: Results of PROBAST risk of bias (ROB) assessment per domain. We performed a ROB assessment of the twenty-three studies included in our systematic review.

1 = The seven studies that were deemed to have a low ROB in Domain 1 did so by appropriately reflecting the target population of interest by including immunosuppressed PCNSL patients[23,45] or explicitly including participants with atypical variants of tumors.[28-30,35,39] Among the five studies that were judged to have a high ROB related to the selection of participants, four either excluded immunosuppressed patients or excluded patients with certain atypical features[24,31,38,44] and one included patients with "CNS lymphoma", not specifying if perhaps secondary CNSL were intermixed in the dataset.[36] These factors pose a risk factor for bias, since the studied population might differ from the one the model is likely to be used, and the results might therefore not be generalizable. Due to reporting deficiencies, the ROB in this domain could not be determined for eleven studies.

2 = The two main reasons for why studies were deemed to have a high ROB in Domain 4 were i) an inappropriate high ratio of features-per-participant, and ii) using an inappropriate method for handling missing data, particularly the complete-case method". Among the eleven studies with high ROB, five studies had an inappropriately high number of predictors- (features-)to-participants ratio and used the inappropriate complete-case method for handling missing data.[30,32,34,35,41] Furthermore, three other studies had only an inappropriate high predictor number to participants ratio[29,33,37], and other three only the inappropriate complete-case method for handling missing data.[24,26,45] For the remaining studies, the ROB in domain 4 could not be properly assessed since none reported any calibration measure or any information on the presence or missing data or the method to handle it.

3 = Overall, most studies were judged to have a high ROB, because of concerns in Domains 1 and 4. Due to the multiple reporting deficiencies (as assessed with the TRIPOD checklist), the remaining studies had an unclear ROB. Adherence to reporting standards is therefore strongly encouraged, since it is a prerequisite to successfully perform a risk of bias assessment.