cancers

Systematic Review

Machine Learning Applications for Differentiation of Glioma from Brain Metastasis—A Systematic Review

Leon Jekel [1,2,3], Waverly R. Brim [1,4], Marc von Reppert [1], Lawrence Staib [1], Gabriel Cassinelli Petersen [1], Sara Merkaj [1], Harry Subramanian [1], Tal Zeevi [1], Seyedmehdi Payabvash [1], Khaled Bousabarah [5], MingDe Lin [1,6], Jin Cui [1], Alexandria Brackett [7], Amit Mahajan [1], Antonio Omuro [8], Michele H. Johnson [1], Veronica L. Chiang [9,10], Ajay Malhotra [1], Björn Scheffler [2,3] and Mariam S. Aboian [1,*]

[1] Department of Radiology and Biomedical Imaging, Yale School of Medicine, 333 Cedar Street, P.O. Box 208042, New Haven, CT 06510, USA; leon.jekel@yale.edu (L.J.); waverly.brim@yale.edu (W.R.B.); marc.vonreppert@yale.edu (M.v.R.); lawrence.staib@yale.edu (L.S.); gabriel.cassinellipetersen@yale.edu (G.C.P.); sara.merkaj@yale.edu (S.M.); harry.subramanian@yale.edu (H.S.); tal.zeevi@yale.edu (T.Z.); sam.payabvash@yale.edu (S.P.); mingde.lin@yale.edu (M.L.); jcui527@gmail.com (J.C.); amit.mahajan@yale.edu (A.M.); michele.h.johnson@yale.edu (M.H.J.); ajay.malhotra@yale.edu (A.M.)

[2] DKFZ Division of Translational Neurooncology at the WTZ, German Cancer Consortium, DKTK Partner Site, University Hospital Essen, 45147 Essen, Germany; b.scheffler@dkfz-heidelberg.de

[3] German Cancer Research Center, 69120 Heidelberg, Germany

[4] Department of Computer Science, The Johns Hopkins Whiting School of Engineering, 3400 North Charles Street, Baltimore, MD 21218, USA

[5] Visage Imaging GmbH, Lepsiusstraße 70, 12163 Berlin, Germany; kbousabarah@visageimaging.com

[6] Visage Imaging Inc., 12625 High Bluff Dr, San Diego, CA 92130, USA

[7] Harvey Cushing/John Hay Whitney Medical Library, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA; alexandria.brackett@yale.edu

[8] Department of Neurology, Yale School of Medicine, 15 York St Ste LCI702, New Haven, CT 06510, USA; antonio.omuro@yale.edu

[9] Department of Neurosurgery, Yale University School of Medicine, New Haven, CT 06510, USA; veronica.chiang@yale.edu

[10] Department of Therapeutic Radiology, Yale University School of Medicine, New Haven, CT 06510, USA

* Correspondence: mariam.aboian@yale.edu

Simple Summary: We present a systematic review of published reports on machine learning (ML) applications for the differentiation of gliomas from brain metastases by summarizing study characteristics, strengths, and pitfalls. Based on these findings, we present recommendations for future research in this field.

Abstract: Glioma and brain metastasis can be difficult to distinguish on conventional magnetic resonance imaging (MRI) due to the similarity of imaging features in specific clinical circumstances. Multiple studies have investigated the use of machine learning (ML) models for non-invasive differentiation of glioma from brain metastasis. Many of the studies report promising classification results, however, to date, none have been implemented into clinical practice. After a screening of 12,470 studies, we included 29 eligible studies in our systematic review. From each study, we aggregated data on model design, development, and best classifiers, as well as quality of reporting according to the TRIPOD statement. In a subset of eligible studies, we conducted a meta-analysis of the reported AUC. It was found that data predominantly originated from single-center institutions (n = 25/29) and only two studies performed external validation. The median TRIPOD adherence was 0.48, indicating insufficient quality of reporting among surveyed studies. Our findings illustrate that despite promising classification results, reliable model assessment is limited by poor reporting of study design and lack of algorithm validation and generalizability. Therefore, adherence to quality guidelines and validation on outside datasets is critical for the clinical translation of ML for the differentiation of glioma and brain metastasis.

check for updates

Citation: Jekel, L.; Brim, W.R.; von Reppert, M.; Staib, L.; Cassinelli Petersen, G.; Merkaj, S.; Subramanian, H.; Zeevi, T.; Payabvash, S.; Bousabarah, K.; et al. Machine Learning Applications for Differentiation of Glioma from Brain Metastasis—A Systematic Review. *Cancers* **2022**, *14*, 1369. https://doi.org/10.3390/cancers14061369

Academic Editor: David Wong

Received: 19 January 2022
Accepted: 1 March 2022
Published: 8 March 2022

Keywords: systematic review; glioma; glioblastoma; brain metastasis; machine learning; artificial intelligence; reporting quality assessment

1. Introduction

Gliomas and brain metastases are the most common brain malignancies and account for a substantial proportion of cancer-related mortality [1]. Brain metastases (BM) occur in 2% of patients with cancer at the point of diagnosis, appear in 12.1% of patients with metastatic disease to any site [2], and occur with the site of primary tumor being unknown in up to 15% of patients at first presentation with cerebral metastasis [3]. Gliomas make up more than 30% of the overall tumors of the central nervous system (CNS) and account for 81% of total CNS malignancies [4]. Clinical management of gliomas and brain metastases varies immensely, thus requiring differential diagnosis early in the course of evaluation.

Magnetic resonance imaging (MRI) is currently regarded as the reference standard for evaluation of cerebral malignancies and their effects on the brain, therapeutic response, and overall disease progression [5]. Classic metastatic disease to the brain can be easily differentiated in clinical practice from gliomas in the setting of multiple metastases and imaging features of defined lesion borders and prominent surrounding edema. On the contrary, solitary parenchymal brain metastases, which are seen in approximately 30% of patients with CNS metastasis [5–7], can mimic the appearance of higher-grade gliomas, in particular glioblastoma (GBM), and complicate accurate diagnosis, especially when the primary site of cancer is unknown upon first discovery of metastasis. In conventional high-resolution MRI, definitive diagnosis of the lesion can be ambiguous. For example, rim-enhancing lesions in contrast-enhanced T1 sequences are most often found to be high-grade gliomas (40%), closely followed by brain metastases (30%) [7]. While recent research has shown added value of advanced MR imaging techniques, such as MR perfusion and spectroscopy, for tumor differentiation, these methods are not always implemented in standard clinical practice, and standard imaging protocols often feature conventional MRI only.

The increasing volume of medical imaging data and the exponential growth of computational power over the course of the last years has propelled the investigation of machine learning (ML) applications in neuroradiology, especially for tasks that require specialized expertise. ML algorithms can perform complex tasks without explicit programming, but instead by learning analytically through exposure to data with the subsequent ability to model complex associations, e.g., for image segmentation or image interpretation [8]. Several studies have developed predictive ML models for increased diagnostic performance in differentiation of cerebral metastatic disease from glioma, however, incorporation of these algorithms into clinical practice has not been achieved yet, as per the ACR Data Science Institute AI Central [9].

We present a literature review, as well as summary reports on average model performance to identify the most promising approaches reported in the current body of literature. We aim to detect shared pitfalls to clinical implementation in the field of differentiation of brain metastases and gliomas. Through our assessment of quality of reporting, we aim to systematically analyze the shortcomings that are common in the literature and, ultimately, formulate recommendations for researchers engaging in this growing field. We aim to identify, to which degree the opacity of reporting and lack of standardization in algorithm development prevail in the literature, which, as pointed out previously [10,11], are detrimental to clinical translation and FDA clearance of ML models.

2. Materials and Methods

2.1. Database Search

This systematic review was registered with the International Prospective Register of Systematic Reviews (PROSPERO) under the registration number CRD42020209938 and conducted according to the guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy) [12] statement. To collect all relevant original research on the applications of AI in neuro-oncology, database searches of Ovid Embase, OVID MEDLINE, Cochrane trials (CENTRAL), and Web of Science—Core Collection were conducted by a clinical librarian (A.B.) in September 2020, January 2021, and September 2021, respectively. The search strategy included both keywords and controlled vocabulary combining the terms for: "artificial intelligence", "machine learning", "deep learning", "radiomics", "magnetic resonance imaging", "glioma", as well as related terms. The search strategy was independently reviewed by a second institutional librarian. All publications identified by the search were subjected to a screening in Covidence software (Veritas Health Innovation Ltd. Melbourne, Australia). The search identified 12,470 candidate articles (Figure 1). 152 duplicates were removed and screening of the remaining 12,318 articles was conducted by a neuroradiology attending (M.A.), radiology resident (H.S.) and three graduate students (L.J., W.B. and M.v.R.). The board-certified neuroradiology attending (M.A.) resolved ambiguous screening conflicts. The abstract review further excluded 10,995 articles that lacked pertinence to neuro-oncology and ML. A total of 1323 articles were reviewed at full-text level and evaluated for eligibility for inclusion in the review. For reviewer-independent uniformity in screening, 8 exclusion criteria were predefined: (1) Abstract-only; (2) No application of ML reported; (3) Not an original article; (4) Not published in English; (5) No investigation of glioma/glioblastoma; (6) Unrelated to MRI, magnetic resonance spectroscopy (MRS), and positron emission tomography (PET) imaging; (7) No human research subjects; (8) Duplicates. Due to fulfilment of at least one of those criteria, 437 additional studies were excluded. 886 eligible full text studies were then reviewed by either a radiology resident (H.S.) or a graduate student (L.J., W.B. and G.C.P.), in addition to a second review by the board-certified neuroradiologist (M.A.). Twenty-nine studies that were identified to specifically investigate the differentiation of glioma from brain metastasis were then analyzed in the present study. The search strategy is provided in the Supplementary Materials (Figure S1).

2.2. Data Extraction and Aggregation

The data extraction was performed independently by two reviewers (L.J., W.B.) using predefined tables in Microsoft Excel. Disagreements were discussed in regular team meetings and were resolved with a supervising neuroradiology attending (M.A.) until consensus was reached. Data points compiled in this study include: article characteristics (title, author, publication year), data characteristics (data source, dataset size, types and number of tumors for training/testing/validation, model validation technique), class balancing (ratio of glioma to brain metastases), model characteristics (best performing ML classifier, classification task, type of features and imaging sequences used for classification, outcome measures for classifier performance) (Table 1), and reporting characteristics. Whenever referred to in this study, internal validation describes measures for assessment of quality and robustness of the model and its ability to predict outcomes on unseen data. Some studies report three-way partitioning of their dataset into training, validation, and testing sets. In this context, validation data serves to mathematically optimize and finetune model hyperparameters and should not be confused with overall internal validation.

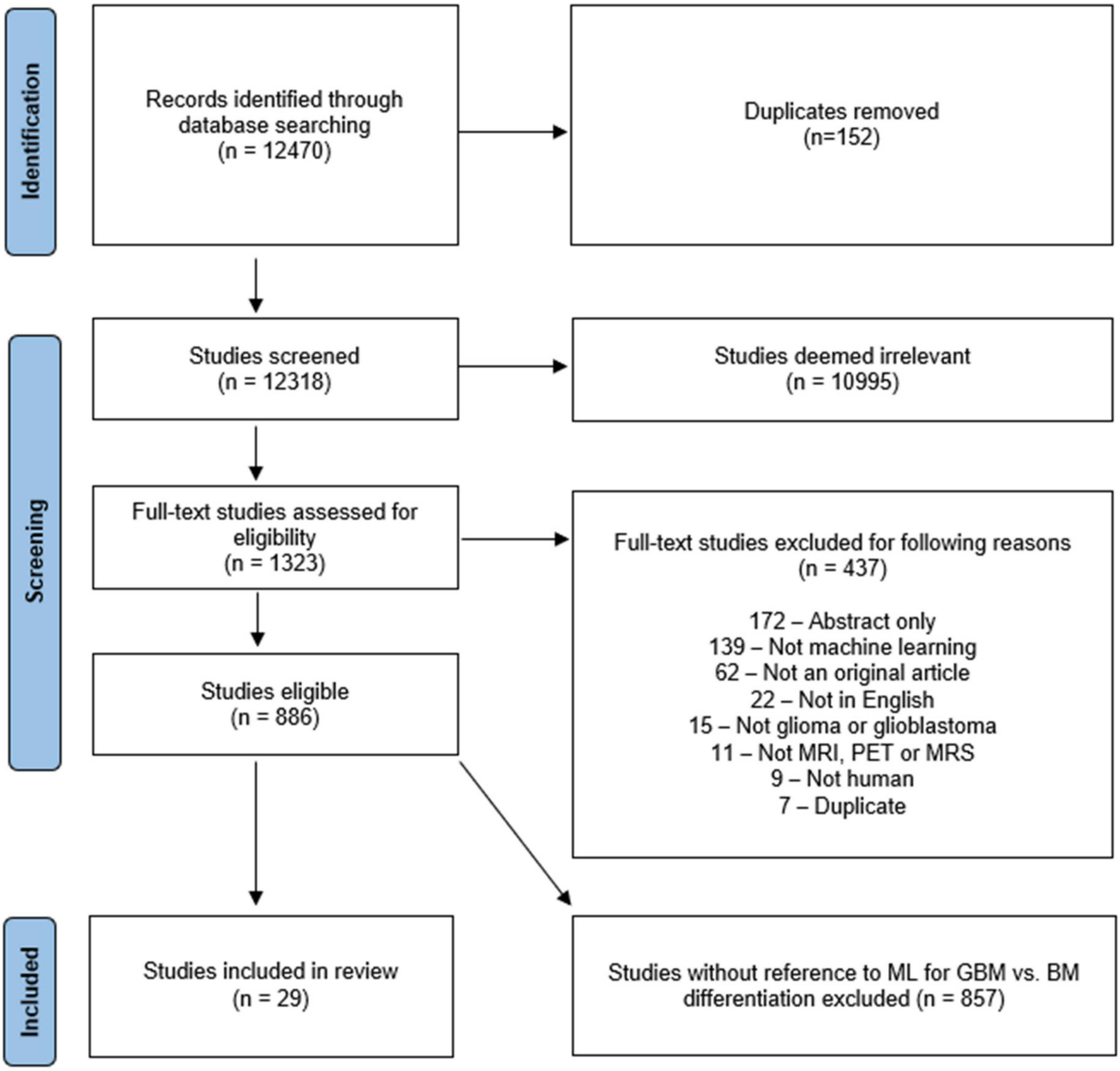
Identification
Records identified through database searching (n = 12470)
Duplicates removed (n=152)
Screening
Studies screened (n = 12318)
Studies deemed irrelevant (n = 10995)
Full-text studies assessed for eligibility (n = 1323)
Full-text studies excluded for following reasons (n = 437)
172 – Abstract only
139 – Not machine learning
62 – Not an original article
22 – Not in English
15 – Not glioma or glioblastoma
11 – Not MRI, PET or MRS
9 – Not human
7 – Duplicate
Studies eligible (n = 886)
Included
Studies included in review (n = 29)
Studies without reference to ML for GBM vs. BM differentiation excluded (n = 857)

Figure 1. Characterization of search strategy using PRISMA. This flowchart represents the search and screening workflow and the eligibility criteria applied to the studies. BM = brain metastasis.

Table 1. Overview of study characteristics and best performing classifier from each study. Abbreviations: GBM = Glioblastoma; MET = Brain metastasis; PCNSL = Primary central nervous system lymphoma; MEN = Meningioma; MED = Medulloblastoma; CV = Cross-validation; LOOCV = Leave-One-Out cross-validation; ML = Machine learning; DL = Deep learning; T1CE = contrast-enhanced T1-weighted sequence; DWI = Diffusion weighted imaging; DTI = Diffusion tensor imaging; PWI = Perfusion weighted imaging; rCBV = relative cerebral blood volume; FLAIR = Fluid-attenuated inversion recovery; TE = Time to echo; AUC = Area under the receiver operating characteristic curve; ADC = Apparent diffusion coefficient; LASSO = Least absolute shrinkage and selection operator; SVM = Support vector machine; MLP = Multilayer perceptron; NNW = Neural networks; LogReg = Logistic Regression; DNN = Deep neural network; LDA = Linear discriminant analysis; NB = Naïve Bayes; VFI = Voting feature intervals; KNN = k-nearest neighbors; PNN = Probabilistic neural networks; RF = Random Forest; RBF = Radial basis function kernel; n/a = not available.

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Swinburne et al., 2019 [13]	26	9	9	1.000	no	yes	GBM vs. MET vs. PCNSL	8 (PCSNL)	LOOCV			no	single-center	ML	SVM, MLP	Pathology	Perfusion	MLP (Ktrans on T1CE mask) Accuracy: 83.3% AUC: 0.83
Park et al., 2020 [14]	276	137	59	2.322	no	yes	GBM vs. MET vs. PCNSL	80 (PCSNL)	216 (109 GBM, 58 PCNSL, 49 MET)	60 (28 GBM, 22 CNSL, 10 MET)		no	multi-center	DL	CNN	Pathology	Perfusion (Temporal Patterns of Time-Signal Intensity Curves from DSC)	CNN (DSC, FLAIR, T1CE)—internally validated AUC: 0.95 Sensitivity: 0.9 Specificity: 0.857
Shrot et al., 2019 [15]	141	41	38	1.079	no	yes	GBM vs. MET vs. PCNSL vs. MEN	12 (PCSNL), 50 (Meningioma)	LOOCV			no	single-center	ML	Decision tree (SVM)	Pathology	Morphology, Diffusion, Perfusion	Binary hierarchical tree with SVM classifier (T1, T1c, T2, FLAIR, DTI, DSC) Sensitivity: 0.974 Specificity: 0.969

Table 1. *Cont.*

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Yamashita et al., 2008 [16]	126	95	19	5.000	multiple	no	Glioma vs. MET vs. PCNSL	12 (PCSNL)	LOOCV			no	not specified	ML	3-layered NNW	Pathology	Clinical, Qualitative/Semantic imaging features	ANN AUC: 0.946 board-certified radiologists without ANN: Accuracy: 87.9% AUC: 0.923 Sensitivity: 0.808 Specificity: 0.903 board-certified radiologists with ANN: Accuracy: 91.5% AUC: 0.946 Sensitivity: 0.868 Specificity: 0.931
Blanchet et al., 2011 [17]	33	18	15	1.200	solitaty	yes	GBM vs. MET		LOOCV			no	single-center	ML	k-means clustering	Pathology	Shape	k-means clustering (T1, T2) Accuracy: 93.9%
Tsolaki et al., 2013 [18]	49	35	14	2.500	solitary	yes	GBM vs. MET		10-fold CV			no	single-center	ML	SVM, Naive Bayes, KNN	Pathology	Spectroscopy	SVM (MRS: NAA; rCBV)—peritumoral Accuracy: 98% Sensitivity: 0.98 Specificity: 0.99 SVM (MRS: NAA/Cr; rCBV)—intratumoral Accuracy: 95% Sensitivity: 0.94 Specificity: 0.95
Yang et al., 2014 [19]	48	30	18	1.667	solitary	yes	GBM vs. MET		LOOCV			no	single-center	ML	QDA, NB, SVM, KNN, NNW (MLP architecture)	Pathology	Shape, Diffusion	Neural Network (DTI) Accuracy: 97.92% AUC: 0.975 Sensitivity: 100% Specificity: 96.55%

Table 1. *Cont.*

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Tateishi et al., 2020 [20]	127	73	53	1.377	multiple, largest selected for classification	yes	GBM vs. MET		5-fold CV			no	single-center	ML	SVM	Pathology, clinical history of path-proven primary cancer	Texture	SVM (T1CE, T2, ADC) AUC: 0.92
Abidin et al., 2019 [21]	52	35	17	2.059	solitary	yes	GBM vs. MET		stratified 10-fold CV			no	single-center	ML	AdaBoost	Pathology	First-order statistics, Texture, Higher-order-features: Topology (Minkowski functionals), Wavelet-transformed, Local Binary Patterns (LBP)	AdaBoost (Local Binary Pattern, T1CE) AUC: 0.846
Bae et al., 2020 [22]	248	159	89	1.787	solitary	yes	GBM vs. MET		166 (109 GBM, 57 MET)	82 (50 GBM, 32 MET)		yes	single-center	ML and DL	DNN, AdaBoost, (L-SVM, LDA)	Pathology	DL extracted (DL) Shape, First-order statistics, Texture (traditional ML)	Deep Neural Network (T1CE)—internal AUC: 0.986 Deep Neural Network (T1CE)—external AUC: 0.956 Accuracy: 89% Sensitivity: 0.906 pecificity: 0.88

Table 1. *Cont.*

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Artzi et al., 2019 [23]	439	212	227	0.934	solitary	yes	GBM vs. MET vs. MET-subtypes		5-fold CV			no	single-center	ML	SVM, KNN, decision trees, ensemble classifiers	Pathology	Clinical features, Qualitative/semantic imaging features, Morphology, First-order statistics, Texture, Higher-order features: Wavelet features, Bagof-features	SVM (T1CE) Accuracy: 89% AUC: 0.96 Sensitivity: 0.86 Specificity: 0.85
Yang et al., 2016 [24]	48	30	18	1.667	solitary	yes	GBM vs. MET		LOOCV			no	single-center	ML	SVM	Pathology	Shape	SVM (DTI, Cluster 1 & 4) Accuracy: 95.83% AUC: 0.983 Sensitivity: 0.9444 Specificity: 0.9667
Dong et al., 2020 [25]	120	60	60	1.000	solitary	n/a	Glioma vs. MET		84 (42 GBM, 42 MET)	36 (18 GBM, 18 MET)		no	single-center	ML	NNW, DT, NB, KNN, SVM	Radiological	Shape, First-order statistics, Texture	Naive Bayes (T1, T1CE, T2) Accuracy: 60% Sensitivity: 0.45 Specificity: 0.75 Combined(LOG) [Decision Tree, SVM, NNW, kNN, NB] Accuracy: 64% Sensitivity: 0.5 Specificity: 0.73 Agreement of all 5 classifier: Accuracy: 94% Sensitivity: 1 Specificity: 0.89

Table 1. *Cont.*

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Meier et al., 2020 [26]	109	25	84	0.298	231 lesions in 109 patients	yes	GBM vs. MET		stratified 3-fold CV			no	single-center	ML	SVM	Pathology	Qualitative/ Semantic imaging features	SVM (Qualitative image features) F1-Score: 0.865
Georgiadis et al., 2008 [27]	67	21	19	1.105	no	no	Glioma vs. MET vs. MEN	27 (Meningioma)	external cross-validation (ECV) with 3-fold split			no	single-center	ML	PNN, LSFT-PNN, SVM-RBF, ANN, Cubic LSFT-PNN, Quardratic LSFT-PNN	Radiological	Texture	ANN (T1)—Primary tumors vs. Secondary tumors (MET + Meningioma) Accuracy: 100%
Tsolaki et al., 2015 [28]	126	80	22	3.636	solitary	no	Glioma vs. MET vs. MEN	24 (Meningioma)	10-fold cross validation			no	single-center	ML	SVM, Naïve Bayes, k-NN, LDA	Pathology	Spectroscopy, Diffusion, Perfusion	SVM (DWI/DTI/PWI/short TE)—peritumoral Accuracy: 98% SVM (DWI/DTI/PWI/short TE)—intratumoral Accuracy: 96%
Zacharaki et al., 2009 [27]	98	74	24	3.083	no	no	Glioma vs. MET vs. MEN	4 (Meningioma)	LOOCV			no	single-center	ML	SVM, k-NN, LDA	Pathology	Shape, First-order statistics, Texture	SVM (FLAIR, T2, T1ce, rCBV, T1) Accuracy: 84.7% AUC: 0.882 Sensitivity: 0.882 Specificity: 0.865

Table 1. *Cont.*

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Zacharaki et al., 2011 [29]	97	73	23	3.174	no	no	Glioma vs. MET vs. MEN		LOOCV			no	single-center	ML	VFI, KNN, Naive Bayes	Pathology	Clinical, Shape, First-order	kNN with wrapper evaluator Accuracy: 96.91%
Svolos et al., 2013 [30]	115	73	18	4.056	solitary	no	Glioma vs. MET vs. MEN	24 (atypical Meningioma)	10-fold cross validation			no	single-center	ML	SVM	Pathology	Diffusion, Perfusion	SVM (HGG Grade 4 vs. MET) (ADC, FA, rCBV)—peritumoral Accuracy: 96% Sensitivity: 0.98 Specificity: 0.94
Sachdeva et al., 2016 [31]	428	177	66	2.682	no	no	Glioma vs. MET vs. MEN vs. MED	97 (Meningioma), 88 (Medulloblastoma)	40% training, 10% testing, 50% validation	40% training, 10% testing, 50% validation	40% training, 10% testing, 50% validation	no	public dataset (PGIMER and SPL datasets)	ML	GA, GA-SVM, GA-ANN	Radiological	First-order statistics, Texture	GA-ANN—no binary classification Accuracy: 94% (imputed)
Payabvash et al., 2020 [32]	248	99	65	1.523	no	no	Glioma vs. MET vs. MED vs. Hemangioblastoma vs. Ependymoma	Hemangioblastoma (n = 44), Ependymoma (n = 27), Medulloblastoma (n = 26).	10-fold cross validaiton			no	single center	ML	NB, RF, NN, SVM	Pathology	Clinical (Age), Qualitative/Semantic imaging features, Diffusion	Random Forest—MET vs. All primary tumors Accuracy: 83% AUC: 0.82 Sensitivity: 55.6 Specificity: 92.6 PPV: 73.9
Qin et al., 2019 [33]	42	24	18	1.333	solitary	yes	GBM vs. MET		5-fold cross validation			no	single center	ML	Decision trees, LDA, LogReg, linear SVM, KNN	Pathology	First-order, Second-order (Energy)	kNN Accuracy: 92.9%

Table 1. *Cont.*

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Chen et al., 2019 [34]	134	n/a	n/a		no	yes	GBM vs. MET		80%	20%		no	single center	ML	LDA, SVM, RF, KNN, Gaussian NB	Pathology	Texture	LogReg + Distance correlation Accuracy: 79% AUC: 0.8 Sensitivity: 0.8 Specificity: 0.71
Ortiz-Ramón et al., 2020 [35]	100	50	50	1.000	no	yes	GBM vs. MET		nested cross-validation			no	single center	ML	random forest (RF), support vector machine (SVM) with linear kernel, k-nearest neighbors (KNN), naïve Bayes (NB) and multilayer perceptron (MLP)	Radiological	Texture	MLP Accuracy: 81% AUC: 0.91 Sensitivity: 0.91 Specificity: 0.8

Table 1. *Cont.*

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Shin et al., 2021 [36]	741	482	259	1.861	solitary	yes	GBM vs. MET		450	48	100	143	multi-center	DL	CNN (2D)	Pathology	DL extracted	CNN (2D; T1CE, T2)—internal Accuracy: 89% AUC: 0.889 Sensitivity: 0.939 Precision: 0.852 CNN—external Accuracy: 85.9% AUC: 0.835 Sensitivity: 0.889 Precision: 0.907
Priya et al., 2021 [37]	120	60	60	1.000	no	yes	GBM vs. MET		nested cross-validation			no	single center	ML	Linear (LASSO, Elastic Net) and logistic regression, NNW, SVM-MLP, RF, AdaBoost	Clinico-Radiological	Shape, First-order statistics, Texture	LASSO (T1, T1CE, T2, FLAIR, ADC) Accuracy: 89.2% AUC: 0.953 Sensitivity: 0.887 Specificity: 0.897
de Causans et al., 2021 [38]	180	92	88	1.045	multiple, largest selected for classification	yes	GBM vs. MET		143 (71 GBM, 72 BM)	nested cross-validation (10 repeated 5-fold CV)	37 (21, 16)	no	multi-center	ML	LogReg (Yeo-Johnson scaling features)	Pathology	Shape, First-order statistics, Texture	LogReg (T1CE) Accuracy: 80% Sensitivity: 0.75 Specificity: 0.86

Table 1. *Cont.*

Paper	Total Patient Number	Number of Glioma Patients	Number of BM Patients	Ratio of Glioma/met Patients	Solitary BM Only	GBM Only	Tumor Types Studied	Number of Additional Tumors	Number of Patients (Training)	Number of Patients (Validation)	Testing	External Validation	Source of Data	ML Method	Algorithms Used for Classification	Gold Standard for Accuracy	Extracted Feature Types	Best Performing Classifier
Liu et al., 2021 [39]	268	140	128	1.094	solitary	yes	GBM vs. MET		208 (110 GBM, 98 BM)	10-fold cross validation	60 (30, 30)	no	single center	ML	RF, DT, LogReg, AdaBoost, Gaussian processing, SVM	Pathology	Shape, First-order statistics, Texture, Higher-order: Wavelet-transformed, Laplace of Gaussian	Random Forest (Boruta selection) (T1CE) Accuracy: 85% AUC: 0.97 Sensitivity: 0.84 Specificity: 0.93
Samani et al., 2021 [40]	136	86	50	1.720	no, 3 patients with multifocal metastasis	yes	GBM vs. MET		108 (66 GBM, 40 BM)	5-fold cross validation	30 (20, 10)	no	single center	DL	2D CNN	Pathology	Diffusion	CNN (2D, DTI, FW-VP map)—patch wise Accuracy: 85% AUC: 0.9 Sensitivity: 0.87 Specificity: 0.81 CNN—majority vote, subject-wise: Accuracy: 93%

2.3. Assessment of Quality of Reporting

The present systematic review also includes a thorough assessment of reporting quality. We conducted a TRIPOD [41] adherence evaluation for model development studies, a reporting guideline with 22 main items and 65 adherence elements in total. The TRIPOD scores were evaluated by two reviewers and, in case of disagreement, consensus by discussion was made. For calculation of TRIPOD adherence scores, we followed the appraisal guidelines [42]. Average TRIPOD adherence score and average degree of item satisfaction, indicated as adherence index (ADI) ranging from 0 to 1, were determined using Microsoft Excel.

2.4. Statistical Analysis

Reporting quality was compared in study cohorts published before and during/after 2019. The difference in median TRIPOD adherence score was examined for statistical significance using the Mann-Whitney U test in MedCalc version 20.019 (MedCalc Software bv, Ostend, Belgium; https://www.medcalc.org (accessed on 1 March 2022); 2021). Differences in mean AUC between study subsets were tested for statistical significance via Student's t-test using GraphPad Prism version 8.3.4. All eligible studies that reported AUC (area under the ROC curve) and standard error of mean (SEM) or 95% confidence intervals (CI), were subjected to a meta-analysis using MedCalc. In one study, SEM was imputed from reported standard deviation (SD) of the mean and sample size. These metrics were subjected to a random effects model. Results of this quantitative analysis were then illustrated in a forest plot [43]. Heterogeneity within the analyzed studies was then examined via the Higgins I^2-test [44].

3. Results

3.1. Study Selection

This systematic review identified 29 eligible studies as part of the literature on ML models for differentiation of gliomas from brain metastases (Figure 1). The studies were published between 2008 and 2021. Yamashita et al. [16] presented a predictive model that used a threshold for differentiation of glioma from brain metastasis. This threshold was determined by shape evaluation performed by an unsupervised ML algorithm and was therefore included in this systematic review.

3.2. Study Characteristics

3.2.1. Datasets

The investigated articles predominantly made use of local single-center hospital datasets (n = 25). Three studies were conducted on hospital data from multiple datasets. In one study, the source of data was not specified and is unclear [16] (Figure 2).

3.2.2. Dataset Composition

The total numbers of subjects described by the studies were overall small, averaging at 154.10 ± 147.25 (mean ± SD) patients. Deep learning (DL) studies used larger datasets, averaging at 350.25 ± 99.00 (mean ± SD). Datasets were found to exhibit a glioma-to-BM ratio of 1.80 (±1.10):1 (Figure 3). While every study presented a ML model for differentiation between gliomas and brain metastasis, classification tasks and inclusion criteria for subjects varied. Ten studies included further tumor entities, such as atypical meningioma [30] or primary CNS lymphoma [13] as prediction classes. While 15 studies specified inclusion of patients with only solitary brain metastasis, this was not explicitly mentioned in nine studies. Five studies explicitly stated the inclusion of BM patients with multiple lesions. Twenty studies reported the exclusive investigation of GBM. The rest of the studies (n = 9) also included lower-grade and/or atypical glioma patients. Tsolaki et al. [28] performed intra-class subgroup analyses and compared classification accuracy for differentiation of brain metastases against different glioma grades. Artzi et al. [23] and Meier et al. [26] differentiated between multiple types of brain metastases for further subtype analyses of brain tumor etiology.

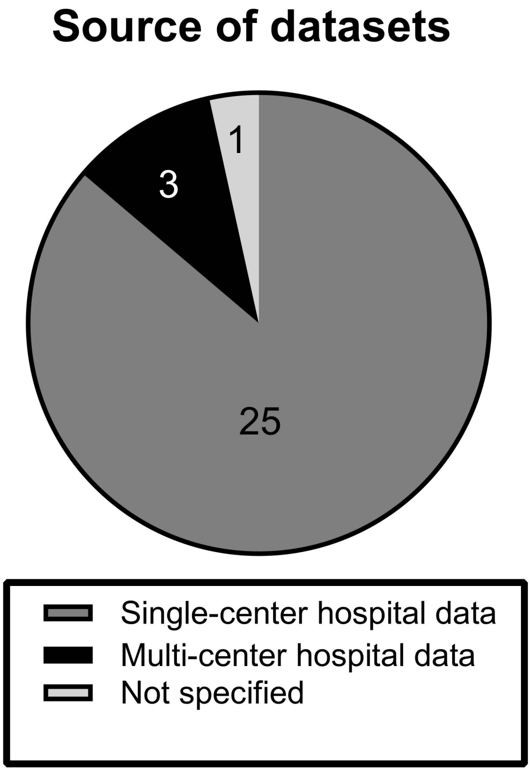

Figure 2. Source of datasets. The chart displays the distribution of types of datasets, from which MRI scans were derived, among the different studies. Note how the majority (89%) of studies trained and validated on single-center data.

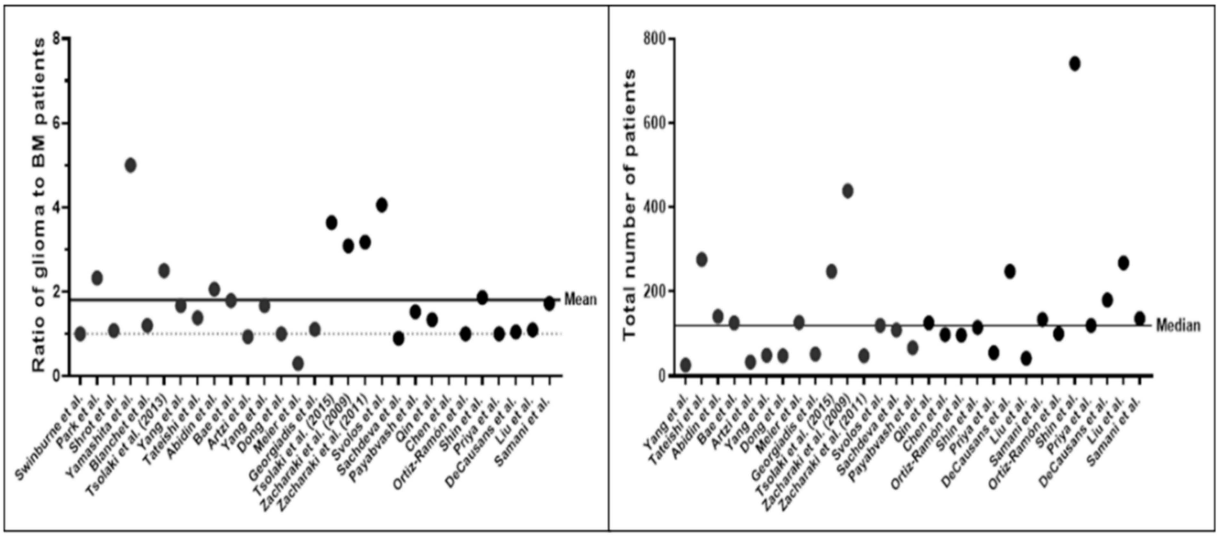

Figure 3. Class distribution of gliomas and brain metastases (**left**) and total number of patient studies (**right**) in each study. The panel on the left-hand side shows the ratio of glioma and brain metastasis patients among the different datasets. The dotted line indicates equal class distribution, i.e. class balance. The right-hand panel indicates the total number of patients across all studies. Note that most studies were trained and validated on datasets with less than 200 patients.

3.2.3. Imaging Modalities and Features

Different MR sequences were used for tumor measurements and subsequent feature extraction. In 25 out of 29 studies, contrast-enhanced T1 (T1CE) sequences were utilized within the workflow described in the study for delineation of the region of interest (ROI), which is the standard-of-care sequence in brain metastases. Only two studies [19,24] did not derive their ROIs from conventional MR imaging, but via segmentation on diffusion tensor imaging. Imaging features from conventional MRI were employed for classification in 23 studies. Ten studies included diffusion-weighted imaging, and six studies included perfusion-weighted imaging, five of which specified the use of dynamic susceptibility contrast (DSC) imaging, whereas the remaining one provided no further information. Four studies included MR spectroscopy. Multiple studies showed that classification results upon investigation of the peritumoral area (outside of the contrast-enhancing border on T1CE) were superior to those yielded from the intratumoral portion [18,28,30]. Similarly, Samani et al. [40] showed high classification results from exploiting the DTI-derived free water volume fraction from the peritumoral microenvironment. These findings suggest a reflection of the disparate tumor biology and mode of growth, which is of an infiltrative nature in glioma and of an expansile nature in brain metastases.

The imaging features used for classification of gliomas from brain metastases were heterogeneous and included clinical, qualitative, and semantic imaging features, as well as shape features and radiomics features of first, second and higher orders (Figure S3). Second-order, i.e. textural quantitative imaging features, were found to be employed most frequently (n = 12).

3.2.4. Algorithms

We aimed to identify the most common ML algorithms in the investigated body of literature. Most studies (n = 19) evaluated multiple classifiers in their study. Support vector machines (SVM) (n = 21) and k-nearest neighbors algorithms (kNN) (n = 10) were the most frequently investigated classifiers. Deep learning (DL) techniques, such as deep neural networks (DNN), or convolutional neural networks (CNN), were leveraged less (n = 4), but were increasingly represented in more recent publications between 2020 and 2021. The 29 best performing classifiers drawn from each study showed a variety of different algorithms. Among those, SVM (support vector machine) and its variations were again represented the most (n = 9). Lesser used traditional ML algorithms were used in the rest of the studies and were grouped into non-DL neural networks (n = 4), logistic regression (n = 3), tree-based ensemble classifiers (n = 2), namely Random Forest and AdaBoost, k-nearest neighbors (n = 2), and others (n = 2). DL-based algorithms, namely CNN (n = 3) and a DNN (n = 1), were the best reported classifiers in four studies. Algorithm representation among all reported classifiers versus representation among the best performing classifiers is visualized in Figure S4a.

3.2.5. Model Validation

Internal validation measures were reported in every study. Cross-validation, particularly leave-one-out cross validation (LOOCV) (n = 9), was performed in 21 studies. Two studies [31,36] presented a three-way split of their dataset into training, validation, and testing set. External validation sets stem from a geographically distinct location and should ensure that the model generalizes well onto data from foreign populations. This was only conducted in two studies [22,36].

3.2.6. Classification Performance

Classification accuracy (n = 19) was the most reported model evaluation metric. One study [26] exclusively provided a F1-Score, the harmonic mean of precision and recall, of 0.865 without class balancing and 0.326 with class balancing, respectively. We aggregated the internally validated performance metrics from all studies that provided dichotomized models for classification between glioma and brain metastases (n = 26). Two studies provided evaluation metrics for differentiation of brain metastasis from the four individual glioma grades separately. For our data aggregation, we opted to only include the herein reported differentiation between "Grade 4 glioma and brain metastasis", which more precisely reflects the clinical diagnostic dilemma in differentiating high grade gliomas from solitary brain metastases. Note that the following syntheses do not meet the criteria of a meta-analysis as most studies failed to provide estimates of level of certainty, such as SEM or CI.

The pooled average of all studies reporting accuracy (n = 19) was 0.881 ± 0.085 (mean ± SD). The average AUC, reported in 17 studies, was found to be 0.916 ± 0.052, while average sensitivity (n = 16) and specificity (n = 15) were 0.868 ± 0.123 and 0.843 ± 0.235, respectively (Figure 4). A subgroup analysis of AUC (mean ± SD) of classifiers modeled on conventional MRI only (0.907 ± 0.061) vs. advanced MRI (0.930 ± 0.053) did not reach statistical significance (p = 0.437) and is represented in Figure S5. Among the 26 best performing classifiers, SVM reached the highest mean (± SD) AUC of 0.936 (±0.045). Mean AUC from the best classifiers grouped by different algorithm types (SVM, DL algorithms, neural networks, tree-based algorithms, and logistic regression) did not vary significantly at an alpha level of 0.05 and are displayed in Figure S4b. Note that these numbers do not represent the results from a meta-analysis, as variance estimates were scarcely reported, and, thus, should be appraised critically. Caution against inference to real-world data is strongly advised.

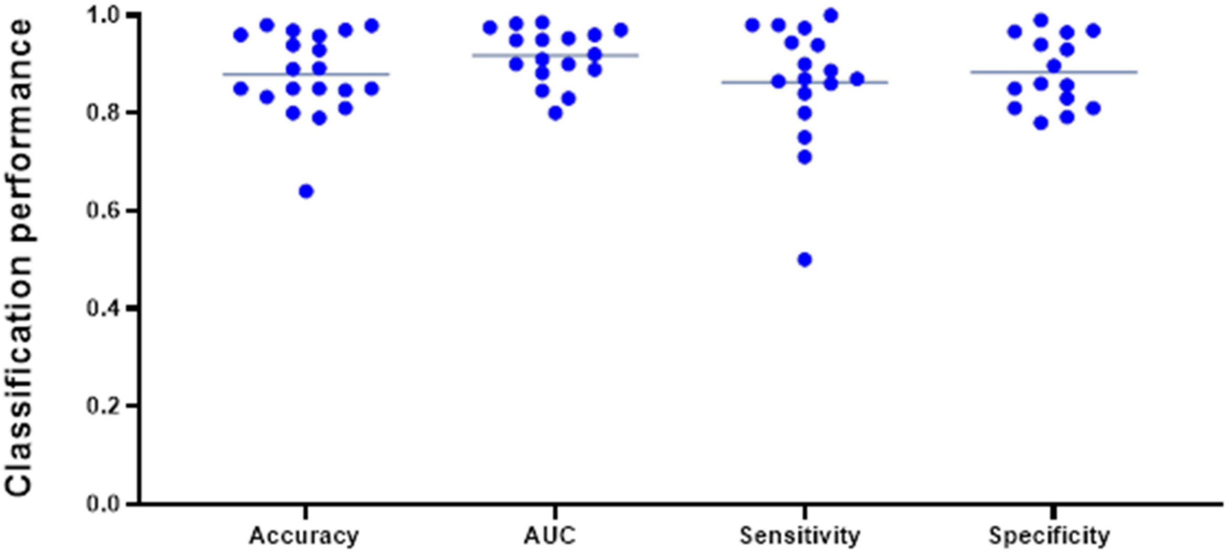

Figure 4. Most frequently reported performance metrics for the best performing classifier from each study included in this data aggregation (n = 26). The lines indicate the mean of the different metrics and reached an overall high level. Note that not all the above-mentioned evaluation metrics were indicated for every classifier and are displayed in different amounts in this plot.

Ensemble learning approaches were described in three studies. Dong et al. [25] presented an ensemble learning approach, where different traditional ML algorithms were combined for execution of a classification task. The ensemble classifier yielded a classification accuracy of 0.64, tying with the single best performing individual algorithm, a Naïve Bayes classifier, indicating that ensembling was unsuccessful. The same study also proposed an agreement pattern model, a voting ensemble, where only cases with unanimous class labels across all five different classifiers were analysed. This approach achieved a classification accuracy beyond 90%, significantly outperforming the other approaches—however, it was not stated in how many cases total agreement could be achieved. This approach likely favours easy to differentiate cases, thus bearing limited applicability in clinical practice, particularly for more ambiguous studies. Samani et al. [40] presented a similar voting ensemble method. The 2D CNN algorithm performed 16 × 16 voxel patch-wise classifications after training on about 6000 peritumoral patches from 113 training subjects. Training and cross-validation were also conducted patch-wise. Prediction of tumor type at patient level was performed in a holdout test set via majority voting of the individual subordinate patches per patient. While classification accuracy in patch-wise cross validation was 0.85, patient-wise classification after majority voting reached an accuracy of 0.93 in a holdout test set. Shin et al. [36], reporting on a 2D CNN, proceeded similarly, providing cross-validated predictions for 6617 axial sectional images, and concluding patient-wise predictions by majority voting. This validated model achieved an AUC of 0.889 on an internal holdout test set and 0.835 on an external validation set. Splitting imaging data from a subject into further divisions for model training, can serve as means of data augmentation to combat overfitting, a common phenomenon when DL is applied to small datasets.

Two studies reported externally validated classification results. Shin et al. [36] reported a convolutional neural network trained on T1CE- and T2-weighted masks with an AUC of 0.835 (95% CI 0.755–0.915). Bae et al. [22] reported a deep neural network trained on radiomics features retrieved from T1CE and peritumoral T2-weighted masks with an AUC of 0.956 (95% CI 0.918–0.990).

3.3. Meta-Analysis

Five studies were found to be eligible for a quantitative meta-analysis of effect estimates. The purpose of this analysis was to obtain robust assembled results on the performance of these ML classifiers. Three studies that reported internally validated AUC values, and two studies reporting externally validated AUC values were analyzed separately in a random effects model. In the internally validated models, an overall AUC of 0.913 (95% CI 0.902–0.925) was attained. Higgins I^2-test yielded a heterogeneity of 0.00%, however statistical significance ($p = 0.597$) was not given. The meta-analysis from two externally validated models reached an overall AUC of 0.907 (95% CI 0.826–0.988). Heterogeneity, again measured by Higgins I^2, was at a level of 86.32% at a significance level of $p < 0.01$. This indicates a high level of heterogeneity in the meta-analysis of externally validated classifiers. Forest plots of the meta-analyses are provided in the Supplementary Materials (Figure S2).

3.4. Quality of Reporting

Adherence to 29 TRIPOD items was assessed for each study in agreement of two reviewers. These included all TRIPOD items applicable to model development studies according to the official TRIPOD review guidelines, except for item 11 (risk groups), which was not applicable to any included study. Median TRIPOD adherence score was found to be 0.48, reflecting that 14 out of 29 TRIPOD items were fulfilled. TRIPOD adherence scores ranged from 0.17 (5/29) to 0.79 (23/29). The studies published in or after 2019 achieved higher TRIPOD scores on average ($p = 0.017$), with a median of 0.55 (95% CI: 0.49–0.61). On the other hand, the studies published before 2019 had a median TRIPOD score of 0.45 (95% CI: 0.40–0.49).

Across studies, average satisfaction per item was measured by what we labeled as adherence index (ADI), with values ranging from zero to one. Highest item satisfaction was detected for Background and Objectives (ADI of 96.1 and 100%, respectively), Study design (93.1%), Model development—Participants and outcomes (82.7%), and Limitations and Implications (both 86.2%). Lowest ADI, indicating low adherence to TRIPOD, were seen in the reporting of Title and Abstract (6.9 and 0%, respectively), Predictors—Blind assessment (3.4%), Participant characteristics (0%), and Full model specification and Model performance (both 10.3%) (Figure 5).

Outside of the framework of TRIPOD, we investigated the reporting on data availability. While six studies [14,32,34,37,38,40] explicitly mentioned data availability upon request, among all twenty-nine examined studies, only Liu et al. [39] provided the algorithm code and radiomics data on an open-source platform.

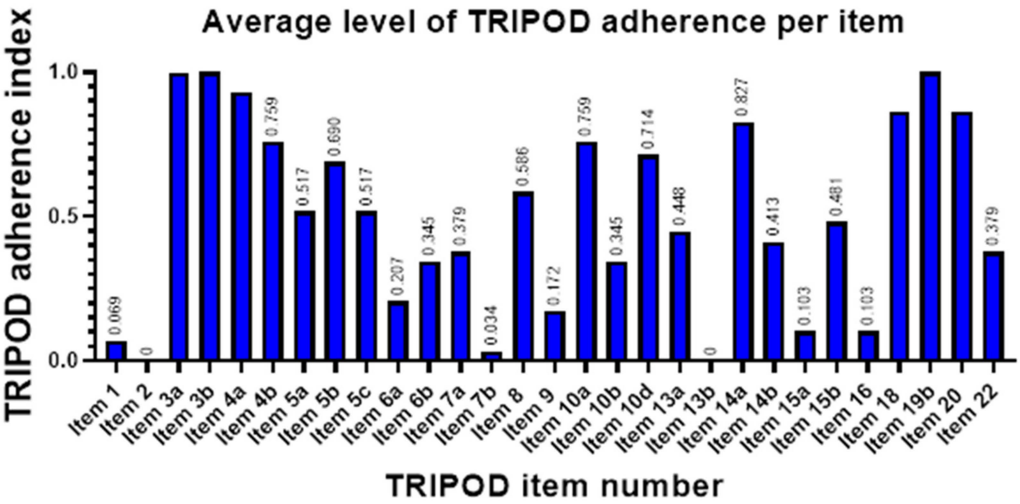

Figure 5. Bar graph of TRIPOD adherence index, a measure for degree of satisfaction for each TRIPOD item applicable in model development studies. Item 11 (Risk groups) was applicable in none of the studies at hand. Item 21 (Supplementary information) is not shown, as it is not included in overall scoring according to official guidelines. Item labels are as follows: 1—Title, 2—Abstract, 3a—Background, 3b—Objectives, 4a—Source of data (Study design), 4b—Source of data (Study dates), 5a—Participants (Study setting), 5b—Eligibility criteria, 5c—Participants (Treatments received), 6a—Outcome (Definition), 6b—Outcome (Blind assessment, 7a—Predictors (Definition), 7b—Predictors (Blind assessment), 8—Sample size, 9—Missing data, 10a—Statistical analysis (Predictors), 10b—Statistical analysis (Model development), 10d—Statistical analysis (Model evaluation), 13a—Flow of participants, 13b—Participant characteristics, 14a—Number of participants and outcomes, 14b—Model development (Predictor and outcome association), 15a—Full model specification, 15b—Model explanation, 16—Model performance, 18—Limitations, 19b—Interpretations, 20—Implications, 22—Funding. Note that items 15a (Full model specification) and 16 (Model performance) are among the TRIPOD items with lowest adherence in the surveyed studies despite having a paramount role for model reproducibility and successful translation to the clinic.

4. Discussion

Previous reports and systematic reviews have corroborated the potential benefit of machine learning for various applications in neuro-oncology, for instance, in prediction of tumor grade, molecular status, or differentiation of glioma from primary central nervous system lymphomas. Predictive ML models for the differentiation of gliomas from brain metastases have the potential to accurately and non-invasively provide preoperative diagnosis and, thus, to influence the strategy for individualized treatment. However, to our knowledge, no study has systematically reviewed the use of machine learning for this differential classification task yet.

Our systematic review was performed under PRISMA guidelines after a thorough search of four databases at three timepoints between September 2020 and September 2021, which resulted in the evaluation of 12,470 abstracts. We extracted information from 29 studies that reported the development of predictive ML models for the differentiation of glioma from brain metastasis.

Our study showed that most articles investigated SVM as classification algorithms and that SVM performed consistently well. As a traditional ML approach, it is a common algorithm in predictive modeling in neuroradiology, due to its simplicity and flexibility [45]. DL algorithms, such as CNN, were described in fewer publications that were more recent. These studies exhibited higher average sample sizes and reported external validation in two out of four instances. Bae et al. presented a deep neural network that outperformed classical ML algorithms, such as AdaBoost or SVM, with an externally validated AUC of 0.956. This aligns with previous studies, which have reported superior results from DL-based algorithms when compared to classical ML techniques for related classification or segmentation tasks [46,47]. DL is widely believed to be one of the major recent advances in the field of machine learning, largely thanks to the increasing availability of big data and growing computational capacities. DL algorithms have been leveraged for different tasks in neuro-oncology, such as acquisition, segmentation, and classification [48]. When compared to classical ML algorithms, which require extraction of handcrafted features, deep neural networks bear the ability to automatically extract relevant features for classification, a process referred to as representation learning [49]. While some studies suggest that DL can outperform classical ML techniques even in small datasets of 50 subjects [50], generally larger datasets are needed to account for the high number of weights within the complex architecture of deep neural networks [51]. As DL algorithms are prone to overfitting in small datasets, which is often the case for medical imaging datasets, validation in external datasets is even more important to preclude over-sensitization to institutional biases. If data is scarce, measures, such as transfer learning [52] or data augmentation should be explored to increase model robustness . Despite promising classification results in our reviewed DL studies, thus, caution is advised, as large, annotated, and high-quality datasets are necessary to prevent overfitting [53]. Ensemble learning methods were trialed in three studies [25,36,40]. Ensemble learning is based on the idea of combining multiple algorithms that were each either trained on different datasets (data diversity) or trained on the same data with differences in algorithm architecture (structural diversity), to generate a model that outperforms the individual classifiers [54]. The final outcome can be obtained either by mathematical fusion of the classifiers or by summing the individually predicted class labels via ensemble voting. This approach can be used to overcome typical challenges in ML, such as overfitting, and to mitigate phenomena, such as class imbalance or the "curse of dimensionality" [55]. Furthermore, we found that the average AUC from the best performing classifiers trained on conventional MRI sequences alone and those integrating information from advanced MRI did not differ significantly. However, we remark that the displayed outcome measures stemmed from internal validation, and hence do not allow inference and generalization to the whole population. In curated datasets, performance metrics are inherently dependent on different factors, such as number of training subjects or data quality that were reported heterogeneously across the displayed studies. This analysis could be repeated once there are enough studies that address this and perform external validation.

Different feature types were used for the classification between gliomas and brain metastases. Three studies compared classification performance using intra- versus peritumoral features and found that the latter achieved higher discrimination performance. This is in line with previously published findings: relative cerebral blood volume (rCBV) measurement has been shown to be a strong discriminator for glioma versus brain metastasis in the peritumoral edema [56,57]. Lu et al. [58] also showed the predictive utility of mean diffusivity from DTI in the peritumoral compartment for distinction between glioma and brain metastasis. Future research should further validate these results in machine learning studies.

Our findings show that there are several limitations that reduce overall model reproducibility, hence posing a barrier for clinical implementation. Most studies relied on single-center datasets, and validation on external imaging datasets was only performed in two studies. This is common in the ML literature on brain tumors because the generation of comprehensive annotated imaging datasets and data sharing are major limitations in the field [59]. Additionally, sample sizes were found to be consistently low at 152.3 ± 144.9. One of the problems of using small datasets is overfitting, where the model captures noise and inherent structures within the training data, which has important ramifications for model generalizability on heterogeneous and unseen data and can overestimate the accuracy measures of algorithms [60]. While there is no consensus on what sample size is too low for application of ML [61], several approaches to determine the minimally required sample size have been proposed [62]. Moreover, several collaborative initiatives, for instance the COINSTAC [63] platform or federated learning (FL) [64], have been developed with the aim of combatting the paucity of large and annotated datasets. FL is an emerging collaborative approach that encompasses multiple centers that train a machine learning model on their institutional private data, and subsequently integrates all model updates into a consensus model. Thereby, FL has been shown to perform comparably to conventional open interinstitutional data sharing models, while bypassing data privacy and confidentiality issues [65].

Another limitation of these studies is the deployment of imbalanced datasets. Classification accuracy was the most frequently provided performance metric in our systematic review. Four studies exclusively provided accuracy as the evaluation metric for their model [28,29,33,66]. However, classification accuracy fails to reliably estimate discriminatory power in the presence of class imbalance, as the impact of a class on model prediction depends on its representation within the data used in the training process [67]. A bias towards the majority class is introduced, as there are fewer cases in the minority class to contribute to overall classification accuracy. Thus, specific characteristics of the minority class are likely to be misidentified as noise and ignored during modeling [68]. Class imbalance can and should be mitigated using various approaches, such as data augmentation, resampling, or employment of penalized models.

All studies used internal validation techniques to evaluate model robustness. Cross-validation techniques, that were most frequently devised for internal validation, iteratively partition the dataset into training and validation sets and are particularly useful when data is scarce, by enabling the exploitation of the entire dataset for modeling. Holdout validation, dividing the samples into designated train and test sets, can, if not stratified, introduce bias when observations are unevenly distributed among training and testing sets, and should typically be reserved for larger datasets. Two studies [31,38] presented a three-way split of their dataset into training, validation, and testing sets. In this setting, the validation set takes on the task of hyperparameter tuning and is typically necessary in complex models with many hyperparameters. Nested cross-validation, described by three studies [35,37,38], addresses the same question of mathematical optimization of the model. To warrant that the models generalize well on unseen data, the choice of internal validation measures should be adapted to the underlying data and individual model and must be reported accordingly.

Several studies included MR scans from brain metastasis patients with multiple foci or failed to explicitly report on the inclusion of solitary BM patients. Recognizing this question is vital when aiming to provide tools for reliable discrimination of brain metastases from GBM, as clinically relevant diagnostic challenges occur in the differentiation of solitary metastases from higher-grade gliomas.

In this systematic review, we provided pooled summary results for the best performing classifiers from each study. Due to a broad heterogeneity among the studies and failure to report necessary effect estimates (AUC) and their variances, not all of them could be subjected to a quantitative meta-analysis. Only small subsets of three and two studies, respectively, were eligible for the conduct thereof. Consistency of effects across the studies is a prerequisite for generalizability in meta-analyses [44]. Further subsampling, as advised for exploration of the cause and type of heterogeneity in our meta-analyses, could not be performed due to the limited number of studies included in the analysis. Overall, our meta-analyses indicate a tendency of the classification performance of eligible studies, all published between 2019 and 2021. However, the limited number of eligible studies indicates the prematurity of this analysis and the need for further validation of these findings. Once the literature fulfils the necessary requirements for the conduct of a more comprehensive meta-analysis, future research could potentially include subgroup analyses, such as comparison of prediction models exploiting information from the intratumoral vs. the peritumoral region, to provide more robust evidence on the effect estimates.

To assess the quality of reporting, we performed a systematic analysis according to the TRIPOD statement. Overall, quality of reporting in the studies was poor. An overall median TRIPOD score of 0.48 (range: 0.17–0.79) signifies that, on average, more than half of the critical information for study development was not reported. Strikingly, our findings are to a large extent in keeping with a recently published systematic review investigating TRIPOD adherence in clinical prediction models in oncology using ML [69]: Dhiman et al. also found low adherence for Title, Abstract. and Predictor blinding; reporting on Background/Objectives and overall interpretation of study results was similarly high in our study cohort; major differences could be seen in the reporting of missing data, where our reviewed articles showed significantly lower adherence; however, given that Dhiman et al. investigated clinical prediction models, it is likely that their results do not translate immediately to ML models based on imaging. This persistent concern of insufficient reporting in the literature necessitates initiatives for data sharing and improvement of transparency.

Based on the deficiencies in current reporting that we identified in our study, we formulated the following recommendations. We propose that authors clearly specify the proportion of solitary brain metastases versus multiple brain metastases in their datasets. The differentiation of solitary brain metastasis from high grade gliomas represents a clinically relevant problem that can be assisted by ML algorithms. On the contrary, differentiation of multiple metastases and solitary metastases from gliomas can have important implications for algorithms that screen studies from normal to abnormal. We advise the use of multi-center hospital datasets for algorithm training and validation. For reporting of model performance, we suggest including multiple performance metrics and statistical testing. Validation of studies in clinically applicable, representative, and independent datasets is crucial for the accurate estimation of generalizability. We understand that such databases may not be readily available; therefore, clear indication of the methods used for validation is critical for future research. Based on our TRIPOD adherence assessment, we recommend providing more descriptive titles that describe the model tested, improving the discussion of results and methods within the abstract, and including balanced datasets with equal representation of tumor types for initial model development. When algorithms are developed for specifically imbalanced dataset applications, then the clear description of the different entities within the dataset is important. We advise authors to sufficiently characterize the predictors used for modeling and to explicitly mention the absence and presence of missing data, respectively. Backed by recent reviews and editorials, we stipulate that strict adherence to standardized reporting guidelines leads to more transparency and can ultimately facilitate model translation and to clinical practice [10,11].

There are several limitations to our study. It is possible that we did not identify every relevant article in the field. To address this limitation, we used four bibliographical databases, as recommended by the Cochrane Handbook for the conduct of systematic reviews. The search was conducted by two institutional librarians and was repeated two times, most recently in September 2021. Another potential limitation is the exclusion of "abstracts only" studies from our systematic review. We acknowledge that this decision, aimed to warrant the inclusion of peer-reviewed results only, could come at the expense of missing pertinent or even contesting evidence to our findings. We prioritized the highest quality of studies over preliminary reports published at scientific meetings. Furthermore, we emphasize that TRIPOD is a quality assessment tool that is tailored for regression-based multivariate prediction models [42], instead of ML techniques that can pursue a different approach for classification. While TRIPOD still provides a rigorous evaluation of ML methods similar to checklists with a focus on AI, such as CLAIM [70], the language in TRIPOD is focused on multivariate regression models. Hence, we acknowledge that translation of TRIPOD to ML studies can be effortful and imprecise. We endorse the development of the TRIPOD–AI extension for explicit use in ML studies, which is currently under development [71].

5. Conclusions

We show that the literature demonstrates early evidence for the efficacy of ML algorithms for glioma versus BM classification and paves the way for clinical implementation of potential algorithms. Significant limitations include small datasets, imbalanced representation of pathologies, and lack of external validation of algorithms. This necessitates initiatives for data or algorithm sharing and development of representative multi-center datasets that allow individualization of algorithms to patient populations and imaging protocols from different institutions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers14061369/s1, Figure S1: Search strategy used on bibliographic databases; Figure S2: Meta-analysis results from random effect models displayed as forest plots; Figure S3: Types of imaging features across all studies; Figure S4: (a) Number of times that algorithm types were reported in a study vs. algorithm representation among the best classifiers extracted from each study; (b) AUC of the best reported classifiers grouped by type of algorithm leveraged for classification; Figure S5: AUC of the best reported classifiers grouped by type of imaging modality leveraged for classification.

Author Contributions: Conceptualization, L.J., H.S., A.M. (Ajay Malhotra), M.S.A.; methodology, L.J., W.R.B., L.S., M.S.A.; software, L.J., A.B.; formal analysis, L.J., W.R.B.; investigation, L.J., W.R.B., M.v.R.; data curation, L.J., W.R.B., M.v.R.; writing—original draft preparation, L.J.; writing—review and editing, L.J., S.M., H.S., T.Z., S.P., K.B., M.L., J.C., A.M. (Amit Mahajan), A.O., M.H.J., V.L.C., A.M. (Ajay Malhotra), B.S., M.S.A.; visualization, L.J., G.C.P.; supervision, B.S., M.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: M.A. received funding from American Society of Neuroradiology Fellow Award 2018. This publication was made possible by KL2 TR001862 (M.A.) from the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH), and NIH roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH. S.P. has grant support from NIH/NINDS K23NS118056, foundation of American Society of Neuroradiology (ASNR) #1861150721, Doris Duke Charitable Foundation (DDCF) #2020097, and NVIDIA. M.L. receives funding from NIH/NCI R01 CA206180 (unrelated to this work).

Institutional Review Board Statement: Ethical review and approval were waived because this study is a systematic review.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article or supplementary material.

Conflicts of Interest: A.O. is listed as associate editor for Cancers. M.L. is an employee and stockholder of Visage Imaging, Inc. and is a board member of Tau Beta Pi engineering honor society. K.B. is an employee of Visage Imaging, GmbH.

References

1. Ostrom, Q.T.; Gittleman, H.; Farah, P.; Ondracek, A.; Chen, Y.; Wolinsky, Y.; Stroup, N.E.; Kruchko, C.; Barnholtz-Sloan, J.S. CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006–2010. *Neuro Oncol.* **2013**, *15* (Suppl. 2), ii1–ii56. [CrossRef] [PubMed]
2. Cagney, D.N.; Martin, A.M.; Catalano, P.J.; Redig, A.J.; Lin, N.U.; Lee, E.Q.; Wen, P.Y.; Dunn, I.F.; Bi, W.L.; Weiss, S.E.; et al. Incidence and prognosis of patients with brain metastases at diagnosis of systemic malignancy: A population-based study. *Neuro Oncol.* **2017**, *19*, 1511–1521. [CrossRef] [PubMed]
3. Rassy, E.; Zanaty, M.; Azoury, F.; Pavlidis, N. Advances in the management of brain metastases from cancer of unknown primary. *Future Oncol.* **2019**, *15*, 2759–2768. [CrossRef] [PubMed]
4. Ostrom, Q.T.; Gittleman, H.; Stetson, L.; Virk, S.M.; Barnholtz-Sloan, J.S. Epidemiology of gliomas. *Cancer Treat Res.* **2015**, *163*, 1–14.
5. Barajas, R.F., Jr.; Cha, S. Metastasis in Adult Brain Tumors. *Neuroimaging Clin. N. Am.* **2016**, *26*, 601–620. [CrossRef]
6. Patchell, R.A.; Tibbs, P.A.; Walsh, J.W.; Dempsey, R.J.; Maruyama, Y.; Kryscio, R.J.; Markesbery, W.R.; Macdonald, J.S.; Young, B. A randomized trial of surgery in the treatment of single metastases to the brain. *N. Engl. J. Med.* **1990**, *322*, 494–500. [CrossRef]
7. Schwartz, K.M.; Erickson, B.J.; Lucchinetti, C. Pattern of T2 hypointensity associated with ring-enhancing brain lesions can help to differentiate pathology. *Neuroradiology* **2006**, *48*, 143–149. [CrossRef]
8. Choy, G.; Khalilzadeh, O.; Michalski, M.; Do, S.; Samir, A.E.; Pianykh, O.S.; Geis, J.R.; Pandharipande, P.V.; Brink, J.A.; Dreyer, K.J. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology* **2018**, *288*, 318–328. [CrossRef]
9. ACR Data Science Institute AI Central. Available online: https://aicentral.acrdsi.org (accessed on 17 January 2022).
10. Ebrahimian, S.; Kalra, M.K.; Agarwal, S.; Bizzo, B.C.; Elkholy, M.; Wald, C.; Allen, B.; Dreyer, K.J. FDA-regulated AI algorithms: Trends, strengths, and gaps of validation studies. *Acad. Radiol.* **2021**. *online ahead of print*. [CrossRef]

11. Lin, M. What's Needed to Bridge the Gap Between US FDA Clearance and Real-world Use of AI Algorithms. *Acad. Radiol.* **2021**. *online ahead of print*. [CrossRef]
12. 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. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
13. Swinburne, N.C.; Schefflein, J.; Sakai, Y.; Oermann, E.K.; Titano, J.J.; Chen, I.; Tadayon, S.; Aggarwal, A.; Doshi, A.; Nael, K. Machine learning for semi-automated classification of glioblastoma, brain metastasis and central nervous system lymphoma using magnetic resonance advanced imaging. *Ann. Transl. Med.* **2019**, *7*, 232. [CrossRef] [PubMed]
14. Park, J.E.; Kim, H.S.; Lee, J.; Cheong, E.N.; Shin, I.; Ahn, S.S.; Shim, W.H. Deep-learned time-signal intensity pattern analysis using an autoencoder captures magnetic resonance perfusion heterogeneity for brain tumor differentiation. *Sci. Rep.* **2020**, *10*, 21485. [CrossRef] [PubMed]
15. Shrot, S.; Salhov, M.; Dvorski, N.; Konen, E.; Averbuch, A.; Hoffmann, C. Application of MR morphologic, diffusion tensor, and perfusion imaging in the classification of brain tumors using machine learning scheme. *Neuroradiology* **2019**, *61*, 757–765. [CrossRef]
16. Yamashita, K.; Yoshiura, T.; Arimura, H.; Mihara, F.; Noguchi, T.; Hiwatashi, A.; Togao, O.; Yamashita, Y.; Shono, T.; Kumazawa, S.; et al. Performance evaluation of radiologists with artificial neural network for differential diagnosis of intra-axial cerebral tumors on MR images. *AJNR Am. J. Neuroradiol.* **2008**, *29*, 1153–1158. [CrossRef]
17. Tateishi, M.; Nakaura, T.; Kitajima, M.; Uetani, H.; Nakagawa, M.; Inoue, T.; Kuroda, J.I.; Mukasa, A.; Yamashita, Y. An initial experience of machine learning based on multi-sequence texture parameters in magnetic resonance imaging to differentiate glioblastoma from brain metastases. *J. Neurol. Sci.* **2020**, *410*, 116514. [CrossRef]
18. Tsolaki, E.; Svolos, P.; Kousi, E.; Kapsalaki, E.; Fountas, K.; Theodorou, K.; Tsougos, I. Automated differentiation of glioblastomas from intracranial metastases using 3T MR spectroscopic and perfusion data. *Int. J. Comput. Assist. Radiol. Surg.* **2013**, *8*, 751–761. [CrossRef]
19. Yang, G.; Jones, T.L.; Barrick, T.R.; Howe, F.A. Discrimination between glioblastoma multiforme and solitary metastasis using morphological features derived from the p:q tensor decomposition of diffusion tensor imaging. *NMR Biomed.* **2014**, *27*, 1103–1111. [CrossRef]
20. Abidin, A.Z.; Dar, I.; D'Souza, A.M.; Lin, E.P.; Wismüller, A. Investigating a quantitative radiomics approach for brain tumor classification. In *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*; International Society for Optics and Photonics: Bellingham, DC, USA, 2019.
21. Georgiadis, P.; Cavouras, D.; Kalatzis, I.; Daskalakis, A.; Kagadis, G.C.; Sifaki, K.; Malamas, M.; Nikiforidis, G.; Solomou, E. Improving brain tumor characterization on MRI by probabilistic neural networks and non-linear transformation of textural features. *Comput. Methods Programs Biomed.* **2008**, *89*, 24–32. [CrossRef]
22. Bae, S.; An, C.; Ahn, S.S.; Kim, H.; Han, K.; Kim, S.W.; Park, J.E.; Kim, H.S.; Lee, S.K. Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: Model development and validation. *Sci. Rep.* **2020**, *10*, 12110. [CrossRef]
23. Artzi, M.; Bressler, I.; Bashat, D.B. Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis. *J. Magn. Reson. Imaging* **2019**, *50*, 519–528. [CrossRef] [PubMed]
24. Yang, G.; Jones, T.L.; Howe, F.A.; Barrick, T.R. Morphometric model for discrimination between glioblastoma multiforme and solitary metastasis using three-dimensional shape analysis. *Magn. Reson. Med.* **2016**, *75*, 2505–2516. [CrossRef] [PubMed]
25. Dong, F.; Li, Q.; Jiang, B.; Zhu, X.; Zeng, Q.; Huang, P.; Chen, S.; Zhang, M. Differentiation of supratentorial single brain metastasis and glioblastoma by using peri-enhancing oedema region-derived radiomic features and multiple classifiers. *Eur. Radiol.* **2020**, *30*, 3015–3022. [CrossRef]
26. Meier, R.; de Mortanges, A.P.; Wiest, R.; Knecht, U. Exploratory Analysis of Qualitative MR Imaging Features for the Differentiation of Glioblastoma and Brain Metastases. *Front. Oncol.* **2020**, *10*, 581037. [CrossRef]
27. Zacharaki, E.I.; Wang, S.; Chawla, S.; Yoo, D.S.; Wolf, R.; Melhem, E.R.; Davatzikos, C. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magn. Reson. Med.* **2009**, *62*, 1609–1618. [CrossRef] [PubMed]
28. Tsolaki, E.; Svolos, P.; Kousi, E.; Kapsalaki, E.; Fezoulidis, I.; Fountas, K.; Theodorou, K.; Kappas, C.; Tsougos, I. Fast spectroscopic multiple analysis (FASMA) for brain tumor classification: A clinical decision support system utilizing multi-parametric 3T MR data. *Int. J. Comput. Assist. Radiol. Surg.* **2015**, *10*, 1149–1166. [CrossRef] [PubMed]
29. Zacharaki, E.I.; Kanas, V.G.; Davatzikos, C. Investigating machine learning techniques for MRI-based classification of brain neoplasms. *Int. J. Comput. Assist. Radiol. Surg.* **2011**, *6*, 821–828. [CrossRef] [PubMed]
30. Svolos, P.; Tsolaki, E.; Kapsalaki, E.; Theodorou, K.; Fountas, K.; Fezoulidis, I.; Tsougos, I. Investigating brain tumor differentiation with diffusion and perfusion metrics at 3T MRI using pattern recognition techniques. *Magn. Reson. Imaging* **2013**, *31*, 1567–1577. [CrossRef]
31. Sachdeva, J.; Kumar, V.; Gupta, I.; Khandelwal, N.; Ahuja, C.K. A package-SFERCB-"Segmentation, feature extraction, reduction and classification analysis by both SVM and ANN for brain tumors". *Appl. Soft Comput.* **2016**, *47*, 151–167. [CrossRef]
32. Payabvash, S.; Aboian, M.; Tihan, T.; Cha, S. Machine Learning Decision Tree Models for Differentiation of Posterior Fossa Tumors Using Diffusion Histogram Analysis and Structural MRI Findings. *Front. Oncol.* **2020**, *10*, 71. [CrossRef]

33. Qin, J.; Li, Y.; Liang, D.; Zhang, Y.; Yao, W. Histogram analysis of absolute cerebral blood volume map can distinguish glioblastoma from solitary brain metastasis. *Medicine* **2019**, *98*, e17515. [CrossRef] [PubMed]
34. Chen, C.; Ou, X.; Wang, J.; Guo, W.; Ma, X. Radiomics-Based Machine Learning in Differentiation Between Glioblastoma and Metastatic Brain Tumors. *Front. Oncol.* **2019**, *9*, 806. [CrossRef] [PubMed]
35. Ortiz-Ramon, R.; Ruiz-Espana, S.; Molla-Olmos, E.; Moratal, D. Glioblastomas and brain metastases differentiation following an MRI texture analysis-based radiomics approach. *Phys. Med.* **2020**, *76*, 44–54. [CrossRef] [PubMed]
36. Shin, I.; Kim, H.; Ahn, S.; Sohn, B.; Bae, S.; Park, J.; Kim, H.; Lee, S.-K. Development and Validation of a Deep Learning–Based Model to Distinguish Glioblastoma from Solitary Brain Metastasis Using Conventional MR Images. *Am. J. Neuroradiol.* **2021**, *42*, 838–844. [CrossRef] [PubMed]
37. Priya, S.; Liu, Y.; Ward, C.; Le, N.H.; Soni, N.; Maheshwarappa, R.P.; Monga, V.; Zhang, H.; Sonka, M.; Bathla, G. Machine learning based differentiation of glioblastoma from brain metastasis using MRI derived radiomics. *Sci. Rep.* **2021**, *11*, 10478. [CrossRef]
38. De Causans, A.; Carré, A.; Roux, A.; Tauziède-Espariat, A.; Ammari, S.; Dezamis, E.; Dhermain, F.; Reuzé, S.; Deutsch, E.; Oppenheim, C. Development of a machine learning classifier based on radiomic features extracted from post-contrast 3D T1-weighted MR images to distinguish glioblastoma from solitary brain metastasis. *Front. Oncol.* **2021**, *11*, 638262. [CrossRef]
39. Liu, Z.; Jiang, Z.; Meng, L.; Yang, J.; Liu, Y.; Zhang, Y.; Peng, H.; Li, J.; Xiao, G.; Zhang, Z. Handcrafted and Deep Learning-Based Radiomic Models Can Distinguish GBM from Brain Metastasis. *J. Oncol.* **2021**, *2021*, 5518717. [CrossRef]
40. Samani, Z.R.; Parker, D.; Wolf, R.; Hodges, W.; Brem, S.; Verma, R. Distinct Tumor Signatures using Deep Learning-based Characterization of the Peritumoral Microenvironment in Glioblastomas and Brain Metastases. *Sci. Rep.* **2021**, *11*, 14469. [CrossRef]
41. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ* **2015**, *350*, g7594. [CrossRef]
42. Moons, K.G.; Altman, D.G.; Reitsma, J.B.; Ioannidis, J.P.; Macaskill, P.; Steyerberg, E.W.; Vickers, A.J.; Ransohoff, D.F.; Collins, G.S. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **2015**, *162*, W1–W73. [CrossRef]
43. Lewis, S.; Clarke, M. Forest plots: Trying to see the wood and the trees. *BMJ* **2001**, *322*, 1479–1480. [CrossRef] [PubMed]
44. Higgins, J.P.; Thompson, S.G.; Deeks, J.J.; Altman, D.G. Measuring inconsistency in meta-analyses. *BMJ* **2003**, *327*, 557–560. [CrossRef] [PubMed]
45. Pisner, D.A.; Schnyer, D.M. Support vector machine. In *Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 101–121.
46. Ugga, L.; Perillo, T.; Cuocolo, R.; Stanzione, A.; Romeo, V.; Green, R.; Cantoni, V.; Brunetti, A. Meningioma MRI radiomics and machine learning: Systematic review, quality score assessment, and meta-analysis. *Neuroradiology* **2021**, *63*, 1293–1304. [CrossRef] [PubMed]
47. Cho, S.J.; Sunwoo, L.; Baik, S.H.; Bae, Y.J.; Choi, B.S.; Kim, J.H. Brain metastasis detection using machine learning: A systematic review and meta-analysis. *Neuro. Oncol.* **2021**, *23*, 214–225. [CrossRef] [PubMed]
48. Zaharchuk, G.; Gong, E.; Wintermark, M.; Rubin, D.; Langlotz, C. Deep learning in neuroradiology. *Am. J. Neuroradiol.* **2018**, *39*, 1776–1784. [CrossRef]
49. Abrol, A.; Fu, Z.; Salman, M.; Silva, R.; Du, Y.; Plis, S.; Calhoun, V. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. Commun.* **2021**, *12*, 353. [CrossRef]
50. Peng, H.; Gong, W.; Beckmann, C.F.; Vedaldi, A.; Smith, S.M. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Anal.* **2021**, *68*, 101871. [CrossRef]
51. Chan, H.-P.; Samala, R.K.; Hadjiiski, L.M.; Zhou, C. Deep learning in medical image analysis. *Adv. Exp. Med. Biol.* **2020**, *1213*, 3–21.
52. Valverde, J.M.; Imani, V.; Abdollahzadeh, A.; de Feo, R.; Prakash, M.; Ciszek, R.; Tohka, J. Transfer learning in magnetic resonance brain imaging: A systematic review. *J. Imaging* **2021**, *7*, 66. [CrossRef]
53. McBee, M.P.; Awan, O.A.; Colucci, A.T.; Ghobadi, C.W.; Kadom, N.; Kansagra, A.P.; Tridandapani, S.; Auffermann, W.F. Deep Learning in Radiology. *Acad. Radiol.* **2018**, *25*, 1472–1480. [CrossRef]
54. Ren, Y.; Zhang, L.; Suganthan, P.N. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Comput. Intell. Mag.* **2016**, *11*, 41–53. [CrossRef]
55. Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
56. Askaner, K.; Rydelius, A.; Engelholm, S.; Knutsson, L.; Lätt, J.; Abul-Kasim, K.; Sundgren, P. Differentiation between glioblastomas and brain metastases and regarding their primary site of malignancy using dynamic susceptibility contrast MRI at 3T. *J. Neuroradiol.* **2019**, *46*, 367–372. [CrossRef] [PubMed]
57. Suh, C.; Kim, H.; Jung, S.; Kim, S. Diffusion-weighted imaging and diffusion tensor imaging for differentiating high-grade glioma from solitary brain metastasis: A systematic review and meta-analysis. *Am. J. Neuroradiol.* **2018**, *39*, 1208–1214. [CrossRef] [PubMed]
58. Lu, S.; Ahn, D.; Johnson, G.; Cha, S. Peritumoral diffusion tensor imaging of high-grade gliomas and metastatic brain tumors. *Am. J. Neuroradiol.* **2003**, *24*, 937–941. [PubMed]
59. Xu, J.; Glicksberg, B.S.; Su, C.; Walker, P.; Bian, J.; Wang, F. Federated learning for healthcare informatics. *J. Healthc. Inform. Res.* **2021**, *5*, 1–19. [CrossRef]

60. Wichmann, J.L.; Willemink, M.J.; de Cecco, C.N. Artificial Intelligence and Machine Learning in Radiology: Current State and Considerations for Routine Clinical Implementation. *Invest. Radiol.* **2020**, *55*, 619–627. [CrossRef]
61. Althnian, A.; AlSaeed, D.; Al-Baity, H.; Samha, A.; Dris, A.B.; Alzakari, N.; Elwafa, A.A.; Kurdi, H. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Appl. Sci.* **2021**, *11*, 796. [CrossRef]
62. Balki, I.; Amirabadi, A.; Levman, J.; Martel, A.L.; Emersic, Z.; Meden, B.; Garcia-Pedrero, A.; Ramirez, S.C.; Kong, D.; Moody, A.R. Sample-size determination methodologies for machine learning in medical imaging research: A systematic review. *Can. Assoc. Radiol. J.* **2019**, *70*, 344–353. [CrossRef]
63. Plis, S.M.; Sarwate, A.D.; Wood, D.; Dieringer, C.; Landis, D.; Reed, C.; Panta, S.R.; Turner, J.A.; Shoemaker, J.M.; Carter, K.W. COINSTAC: A privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. *Front. Neurosci.* **2016**, *10*, 365. [CrossRef]
64. Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K. The future of digital health with federated learning. *NPJ Digit. Med.* **2020**, *3*, 119. [CrossRef] [PubMed]
65. Sheller, M.J.; Edwards, B.; Reina, G.A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R.R. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **2020**, *10*, 12598. [CrossRef] [PubMed]
66. Blanchet, L.; Krooshof, P.; Postma, G.; Idema, A.; Goraj, B.; Heerschap, A.; Buydens, L. Discrimination between metastasis and glioblastoma multiforme based on morphometric analysis of MR images. *Am. J. Neuroradiol.* **2011**, *32*, 67–73. [CrossRef] [PubMed]
67. Branco, P.; Torgo, L.; Ribeiro, R. A survey of predictive modelling under imbalanced distributions. *arXiv* **2015**, arXiv:1505.01658.
68. Noorhalim, N.; Ali, A.; Shamsuddin, S.M. Handling imbalanced ratio for class imbalance problem using smote. In Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017), Langkawi, Malaysia, 7–8 November 2017.
69. Dhiman, P.; Ma, J.; Navarro, C.A.; Speich, B.; Bullock, G.; Damen, J.A.; Kirtley, S.; Hooft, L.; Riley, R.D.; van Calster, B.; et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J. Clin. Epidemiol.* **2021**, *138*, 60–72. [CrossRef]
70. Mongan, J.; Moy, L.; Kahn, C.E., Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [CrossRef]
71. Collins, G.S.; Dhiman, P.; Navarro, C.L.A.; Ma, J.; Hooft, L.; Reitsma, J.B.; Logullo, P.; Beam, A.L.; Peng, L.; van Calster, B.; et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **2021**, *11*, e048008. [CrossRef]