frontiers | Frontiers in Oncology

SYSTEMATIC REVIEW
published: 22 April 2022
doi: 10.3389/fonc.2022.856231

Machine Learning Models for Classifying High- and Low-Grade Gliomas: A Systematic Review and Quality of Reporting Analysis

Ryan C. Bahar[1†], Sara Merkaj[1,2†], Gabriel I. Cassinelli Petersen[1], Niklas Tillmanns[1], Harry Subramanian[1], Waverly Rose Brim[1], Tal Zeevi[1], Lawrence Staib[1], Eve Kazarian[1], MingDe Lin[1,3], Khaled Bousabarah[4], Anita J. Huttner[5], Andrej Pala[2], Seyedmehdi Payabvash[1], Jana Ivanidze[6], Jin Cui[1], Ajay Malhotra[1] and Mariam S. Aboian[1]*

[1] Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, United States, [2] Department of Neurosurgery, University of Ulm, Ulm, Germany, [3] Visage Imaging, Inc., San Diego, CA, United States, [4] Visage Imaging, GmbH., Berlin, Germany, [5] Department of Pathology, Yale-New Haven Hospital, Yale School of Medicine, New Haven, CT, United States, [6] Department of Radiology, Weill Cornell Medicine, New York, NY, United States

Objectives: To systematically review, assess the reporting quality of, and discuss improvement opportunities for studies describing machine learning (ML) models for glioma grade prediction.

Methods: This study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA) statement. A systematic search was performed in September 2020, and repeated in January 2021, on four databases: Embase, Medline, CENTRAL, and Web of Science Core Collection. Publications were screened in Covidence, and reporting quality was measured against the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement. Descriptive statistics were calculated using GraphPad Prism 9.

Results: The search identified 11,727 candidate articles with 1,135 articles undergoing full text review and 85 included in analysis. 67 (79%) articles were published between 2018-2021. The mean prediction accuracy of the best performing model in each study was 0.89 ± 0.09. The most common algorithm for conventional machine learning studies was Support Vector Machine (mean accuracy: 0.90 ± 0.07) and for deep learning studies was Convolutional Neural Network (mean accuracy: 0.91 ± 0.10). Only one study used both a large training dataset (n>200) and external validation (accuracy: 0.72) for their model. The mean adherence rate to TRIPOD was 44.5% ± 11.1%, with poor reporting adherence for model performance (0%), abstracts (0%), and titles (0%).

Conclusions: The application of ML to glioma grade prediction has grown substantially, with ML model studies reporting high predictive accuracies but lacking essential metrics and characteristics for assessing model performance. Several domains, including

OPEN ACCESS

Edited by:
Dario de Biase, University of Bologna, Italy

Reviewed by:
Vincenzo Di Nunno, AUSL Bologna, Italy
Chirag Kamal Ahuja, Post Graduate Institute of Medical Education and Research (PGIMER), India

***Correspondence:**
Mariam S. Aboian
mariam.aboian@yale.edu

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Neuro-Oncology and Neurosurgical Oncology, a section of the journal Frontiers in Oncology

Received: *16 January 2022*
Accepted: *25 March 2022*
Published: *22 April 2022*

Citation:
Bahar RC, Merkaj S, Cassinelli Petersen GI, Tillmanns N, Subramanian H, Brim WR, Zeevi T, Staib L, Kazarian E, Lin M, Bousabarah K, Huttner AJ, Pala A, Payabvash S, Ivanidze J, Cui J, Malhotra A and Aboian MS (2022) Machine Learning Models for Classifying High- and Low-Grade Gliomas: A Systematic Review and Quality of Reporting Analysis. Front. Oncol. 12:856231. doi: 10.3389/fonc.2022.856231

generalizability and reproducibility, warrant further attention to enable translation into clinical practice.

Systematic Review Registration: PROSPERO, identifier CRD42020209938.

Keywords: machine learning, deep learning, artificial intelligence, glioma, systematic review

1 INTRODUCTION

Gliomas are the most common primary brain malignancy (1). They are classified according to histopathologic and molecular World Health Organization (WHO) criteria: grades 1/2 (low-grade gliomas (LGG) and grades 3/4 [high-grade gliomas (HGG)] (2). Glioblastomas, WHO grade 4 tumors, are the most aggressive with a 15-month median overall survival (3).

Because prognosis (3, 4) and treatment (5) vary with glioma grade, accurate classification is essential for guiding clinical decision-making and mitigating risks posed by unnecessary or delayed surgery due to misdiagnosis (6). The gold standard for diagnosis, histopathology, requires surgical resection or stereotactic biopsy for analysis. These invasive procedures, however, carry significant risks and complications (7). Gliomas also exhibit intratumoral heterogeneity with associated sampling error (8). Therefore, a need exists for timely pre-operative whole-glioma grading. As a non-invasive tool for analyzing entire lesions, imaging overcomes the limitations of diagnostic surgical procedures. Although conventional MRI has had modest success in glioma grading (sensitivity 55-83%) (9), the diagnostic potential of imaging has expanded with the use of advanced imaging, radiomics, and artificial intelligence.

Radiomics quantitatively characterizes medical images using image-derived features that serve as biomarkers for tumor phenotypes (10). Artificial intelligence technologies, such as machine learning (ML), have augmented radiomics. By leveraging robust high-dimensional data, ML enhances predictive performance (11). Deep learning (DL) is a subtype of ML that has sparked recent interest given its superior performance in image analysis and suitability for high volumes of data (12). For imaging applications, DL generates useful outputs from input images using multilayer neural networks. Convolutional Neural Networks are the primary DL architecture for image classification (13).

In clinical practice, ML models may increase the value of diagnostic imaging and enhance patient management, for example, by motivating earlier grade-appropriate interventions (14, 15). Despite these opportunities, ML has not been implemented clinically because of numerous technical (data requirements, need for training, low standardization and interpretability) and non-technical (ethical, financial, legal, educational) barriers (16).

High-quality scientific reporting is necessary for readers to critically interpret or replicate studies and encourage translation into practice. Prior work indicates that reporting quality in prediction studies is poor (17). To address this, the Transparent Reporting of a multivariable model for Individual Prognosis or Diagnosis (TRIPOD) Statement was published in 2015 (18). Most of TRIPOD is applicable to ML-based prediction model studies; however, ML-specific guidelines are lacking. The need for such guidelines has initiated development of a TRIPOD extension for ML-based prediction models (TRIPOD-AI) (19, 20).

While ML demonstrates promise for accurate glioma grading, few works have characterized the state of ML in glioma grade prediction (21–23). A systematic review of the literature can identify potential ML methods for clinical use and generate insights for implementation. This study aims to (1) systematically review and synthesize the body of literature using ML for classification of glioma grade, (2) evaluate study reporting quality using TRIPOD, and (3) discuss opportunities for bridging the ML bench-to-clinic implementation gap.

2 MATERIALS AND METHODS

This study followed the guidelines in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA) statement (24) and was registered with the International Prospective Register of Systematic Reviews (PROSPERO, CRD42020209938). An institutional librarian searched the literature published through September 18, 2020, using four databases: Cochrane Central Register of Controlled Trials, EMBASE, Medline, and Web of Science Core Collection. A multi-database approach was pursued because prior work has demonstrated that a single database search may omit pertinent studies (25). Keywords and controlled vocabulary included the following terms and combinations thereof: "artificial intelligence," "machine learning," "deep learning," "radiomics," "magnetic resonance imaging," "glioma," and related terms. The search was repeated on January 29, 2021, to gather additional articles published through this date. A full search strategy is provided in **Appendix A1 (Supplementary)**. A second institutional librarian reviewed the search prior to execution.

Only peer-reviewed studies were imported into Covidence (Veritas Health Innovation Ltd) for screening. Covidence is an online tool designed to streamline the systematic review process. Duplicate studies were identified and removed. Study abstracts were then screened for relevance to neuro-oncology by two of three independent reviewers: an experienced board-certified neuroradiologist, radiology resident, and graduate student in artificial intelligence. The board-certified neuroradiologist resolved discrepancies in screening recommendations. Relevant articles were subsequently assessed for eligibility. To ensure completeness, appropriateness, and understandability of eligible studies, the following exclusion criteria were established: (1) abstract-only; (2) not primary literature; (3) non-English; (4) unrelated to artificial intelligence; (5) unrelated to gliomas;

(6) unrelated to imaging; (7) non-human research subjects; and (8) duplicates. Studies were not excluded based on publication year in order to have comprehensive analysis of historical and contemporary literature. Eligible studies underwent full text review to identify those using ML to classify gliomas by grade. Studies exclusively developing predictive models with distinct focuses (e.g., predicting glioma IDH status, glioma segmentation) were not included in analysis. A PRISMA flow diagram describing our study selection process is depicted in **Figure 1**.

Whole data was independently extracted by two trained medical student researchers using a standardized Microsoft Excel (Microsoft Corporation) form. Conflicts were resolved through team discussion and consensus. When necessary, articles were carefully re-reviewed to obtain missing information after data extraction. The following data points were extracted: article characteristics (title, lead author, country of lead author, publication year), data characteristics (data source, country (or countries) of data acquisition, dataset size, types and number of tumors for training/testing/validation, model validation technique), grading characteristics (study definition of HGG and LGG, gold standard for glioma grading), model characteristics (best performing ML classifier, classification task, supervised/semi-supervised/unsupervised learning, types of features in classifier, imaging sequences used by classifier, measures of classifier performance) and reporting characteristics (TRIPOD items, explained below).

Reporting quality was assessed against the TRIPOD statement in agreement with the TRIPOD adherence assessment form (26) and author explanations (18, 27). TRIPOD contains 20 main items (e.g., main item 5) that apply to studies developing prediction models, 10 of which contain subitems (e.g., 5a, 5b, 5c). Among the 30 total items that can be evaluated and scored, three (item 5c, 11, 14b) were excluded because they were not applicable to our studies. The remaining 27 items were scored for every study. Each item includes one or more elements, all of which must score a "yes" for the item to score "1." To calculate a study's adherence rate to TRIPOD, the number of items scoring "1" was divided by the total number of scored items for the study. Adherence rate for a given TRIPOD item across all studies was calculated by dividing the number of studies scoring "1" for that item by the total number of studies scored.

TRIPOD adherence rates and descriptive statistics (e.g., frequencies, mean ± standard deviation) were calculated and displayed with GraphPad Prism 9 (GraphPad Software). GraphPad Prism 9 is a scientific graphing and statistical software supporting data analysis. Descriptive statistics were

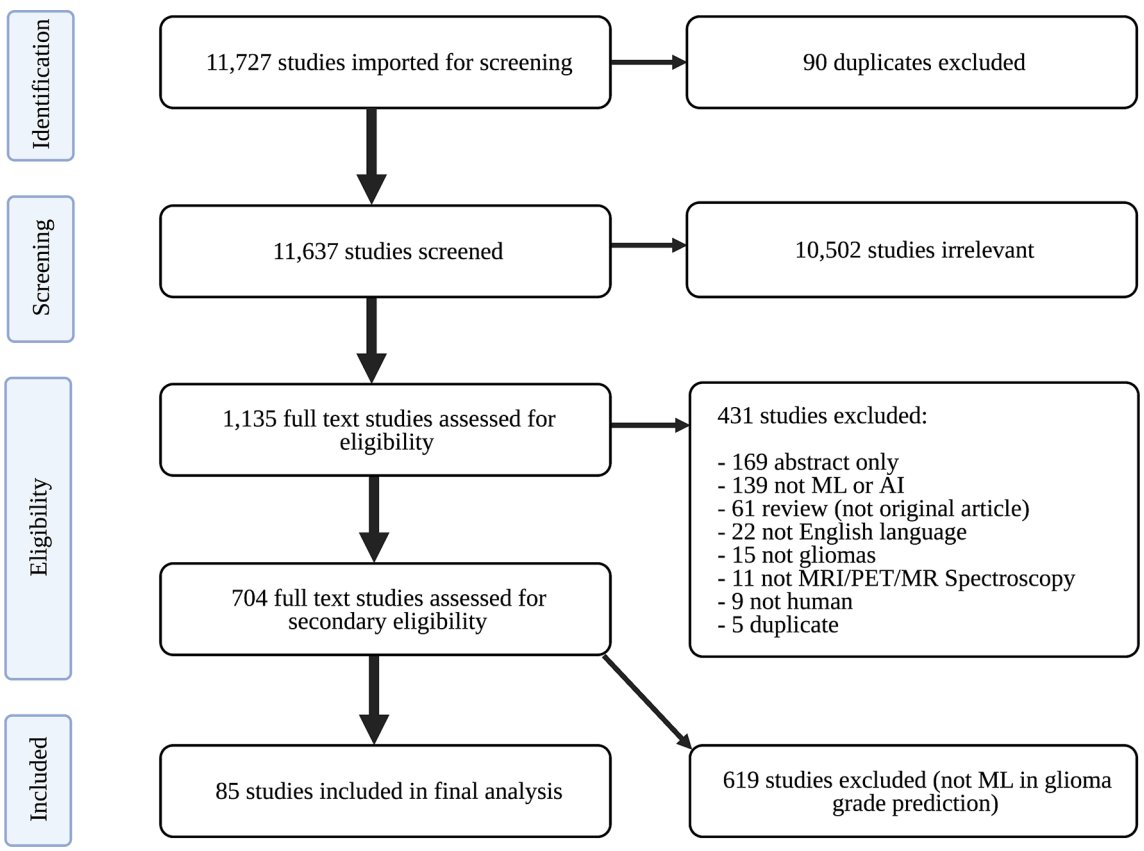

FIGURE 1 | PRISMA flow diagram of study search strategy. ML, machine learning; PRISMA, Preferred Reporting Guidelines for Systematic Reviews and Meta-Analyses.

obtained to summarize study characteristics, dataset and model characteristics, features, imaging modalities, and model prediction performance, among other domains. Only the best performing classifiers' performance metrics (accuracy, AUC, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score) defined in each study are presented. Best performing classifiers were determined based on accuracy results. In the few instances when accuracy was not reported, AUC determined the best performing classifier. All studies meeting our inclusion criteria ((1) identified by search strategy; (2) relevant to neuro-oncology; (3) not excluded during eligibility assessment; and (4) clearly evaluate ML-based classification of glioma grade) contributed to the sample size of our study. References for included studies are listed in **Appendix A4 (Supplementary)**.

3 RESULTS

3.1 Study Characteristics

The search identified 11,727 candidate articles, with 11,637 studies screened for relevance to neuro-oncology. Agreement between screeners was substantial [(Cohen's kappa: 0.77 ± 0.04, see **Table A1 (Supplementary)**]. 1,135 articles underwent full text review, and 85 articles were included in analysis (**Figure 1**).

67 articles (79%) were published between 2018 and 2021, with 26 articles (31%) published in 2019 alone (**Figure 2**). Based on lead author affiliations, most articles were from China, the US, or India (n=45, 51%) (**Figure 3**).

36 articles (42%) defined HGG as grade 3 and 4 and LGG as grade 1 and 2. 17 articles (20%) defined HGG as grade 4 and LGG as grades 2 and 3. 32 articles (38%) didn't define grades for HGG and LGG (**Figure 4**).

3.2 Study Findings

3.2.1 Dataset and Model Characteristics

Among the 84 articles with identifiable patient data sources, data was most commonly acquired multi-nationally (n=38, 45%), entirely in China (n=15, 18%), or entirely in the US (n=10, 12%) (**Figure 3**). BraTS (28) and TCIA (29) datasets, which are publicly available multi-institutional datasets containing multi-parametric MRI scans, were used in 45% (n=38) of studies. Conventional ML was the primary ML model for glioma grade prediction in 59 (69%) studies and DL in 26 (31%) studies. Of all

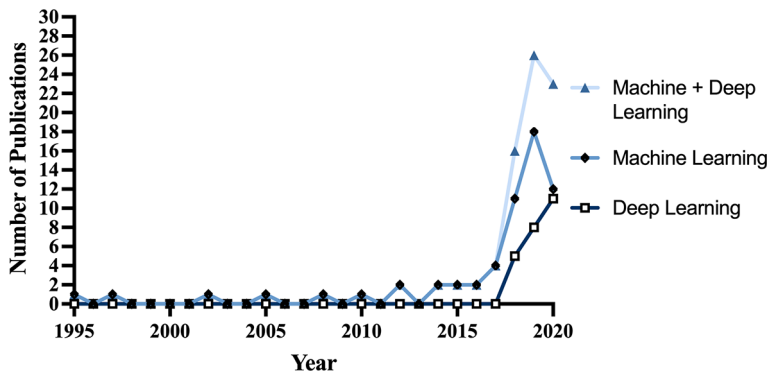

FIGURE 2 | Number of studies published per year from 1995-2020.

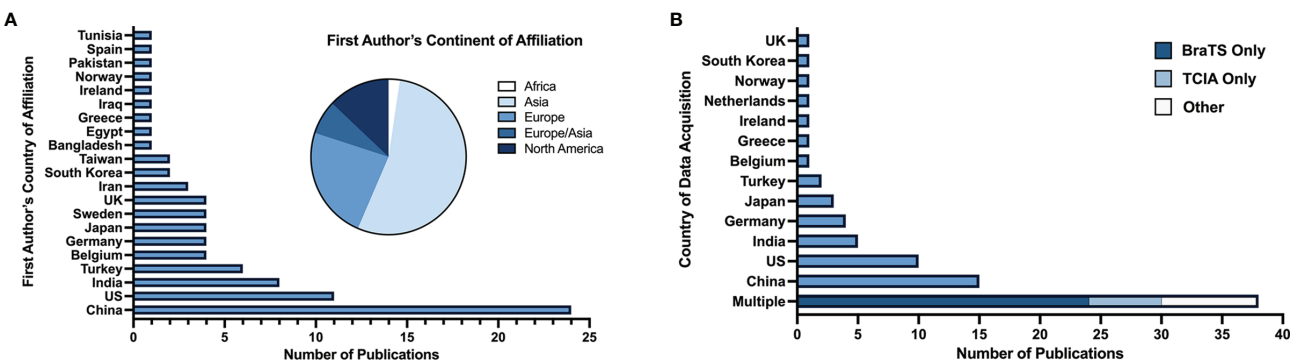

FIGURE 3 | **(A)** Number of studies by first author's country of affiliation and respective continent. **(B)** Number of studies by country (or countries) of data acquisition.

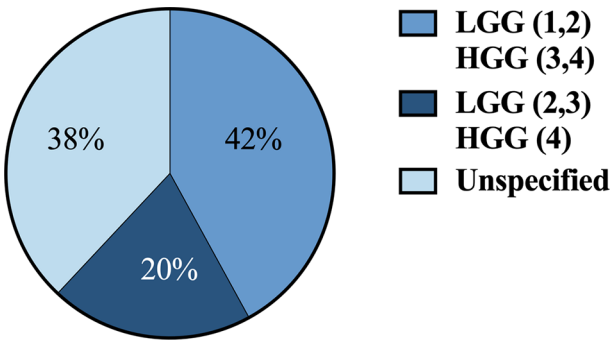

FIGURE 4 | Classification systems used across studies for defining HGG vs. LGG by grade 1-4. HGG, high-grade gliomas; LGG, low-grade gliomas.

85 studies included, 80 (94%) reported the number of patients in their datasets (mean: 177 ± 140). Studies developing conventional ML models reported mean dataset sizes of 168 ± 150 patients. DL studies reported mean dataset sizes of 199 ± 109 patients. Among the 67 studies whose best performing models were binary classifiers of HGG and LGG and reported the number of HGG and LGG used in model development, 58 (87%) had imbalanced datasets characterized by an unequal number of HGG and LGG patients. Most studies (n=44, 66%) used datasets containing more HGG than LGG patients (i.e., HGG : LGG ratio >1). 14 studies (21%) had fewer HGG than LGG patients (HGG : LGG ratio <1).

Only 5 (6%) studies reported external validation. Of the 80 other studies, 68 (85%) reported internal validation and 12 (15%) did not clearly report validation methods. 82 (96%) of studies had supervised learning algorithms and 3 (4%) used semi-supervised learning. No studies reported unsupervised learning algorithms. The gold standard for glioma grading was histopathology in all studies.

3.2.2 Features

Texture (second-order) features and first-order features were the most common feature subsets, extracted in 45 (53%) and 42 (49%) studies, respectively. Shape and/or size features (n=28, 33%) and DL extracted features (n=20, 24%) were also common. Hemodynamic (n=5), qualitative (n=6), higher-order (n=4) and spectroscopic features (n=8) were observed in less than 10% of studies. Definitions for feature types are provided in **Table A7 (Supplementary)**.

3.2.3 Imaging Modalities

T1-weighted contrast-enhanced (T1CE) imaging was the most common sequence used in best performing models (n=54, 64%), followed by T2 (n=46, 54%) and FLAIR (n=40, 47%). T1 pre-contrast was less common (n=35, 41%). Perfusion-weighted imaging (n=15), MR Spectroscopy (n=9) and diffusion-weighted imaging (n=12) were used in 11-18% of models. PET and fMRI were only used in one model each.

3.2.4 Prediction Performance

A summary of model performance measures across studies is shown in **Table 1**. The mean glioma grade prediction accuracy of the best performing algorithm per study was 0.89 ± 0.09. This parameter was determined by taking the prediction accuracy of the best performing algorithm in each study for all studies and calculating a mean value and standard deviation. Lower accuracies were reported for models undergoing external validation (mean: 0.82 ± 0.09, n=5). DL models had a mean prediction accuracy of 0.92 ± 0.08 and conventional ML models 0.88 ± 0.09.

The most common best performing conventional ML model was Support Vector Machine (mean accuracy: 0.90 ± 0.07) and DL model was Convolutional Neural Network (mean accuracy: 0.91 ± 0.10) (**Figure 5**).

We grouped all studies by data source into 4 categories: BraTS, TCIA, single center, and multicenter (excluding BraTS and TCIA) data. Studies which used BraTS as a data source had a mean accuracy of 0.93 ± 0.04 (n=27) and studies using TCIA had a mean accuracy of 0.91 ± 0.08 (n=12). Single center datasets were the most common (n=43) with a mean accuracy of 0.88 ± 0.07, and multicenter hospital datasets the least common (n=6, mean accuracy: 0.80 ± 0.18).

We additionally identified studies whose models were built on relatively large (n≥200) datasets and externally validated, two characteristics indicating potential generalizability. Only one study (1%) had both characteristics (accuracy: 0.72) (30). Further analysis of model performance by dataset source, dataset size, validation technique, and glioma grade classification task can be found in **Appendix A2** and **Tables A2-A5 (Supplementary)**. Characteristics of the 10 studies reporting the highest accuracy results for their best performing algorithms are summarized in **Table 2**. Characteristics of all included studies may be seen in **Table A6 (Supplementary)**.

3.3 Quality Assessment

The mean adherence rate to TRIPOD was 44.5% ± 11.1%, with poor reporting adherence in categories including model performance (0%), abstract (0%), title (0%), justification of sample size (2.4%), full model specification (2.4%), and participant demographics and missing data (7.1%). High reporting adherence was observed for results interpretation (100%), background (98.8%), study design/source of data (96.5%), and objectives (95.3%) (**Figure 6**).

TABLE 1 | Mean (± standard deviation) aggregate performance metrics across studies.

Accuracy (n=82)	AUC (n=48)	Sensitivity (n=55)	Specificity (n=51)	Positive Predictive Value (n=12)	Negative Predictive Value (n=6)	F1 Score (n=7)
0.89 ± 0.09 (0.53-1.00)	0.92 ± 0.07 (0.73-1.00)	0.89 ± 0.09 (0.63-1.00)	0.88 ± 0.11 (0.55-1.00)	0.90 ± 0.09 (0.68-1.00)	0.82 ± 0.08 (0.73-0.94)	0.89 ± 0.11 (0.67-0.98)

n, number of studies reporting metric.

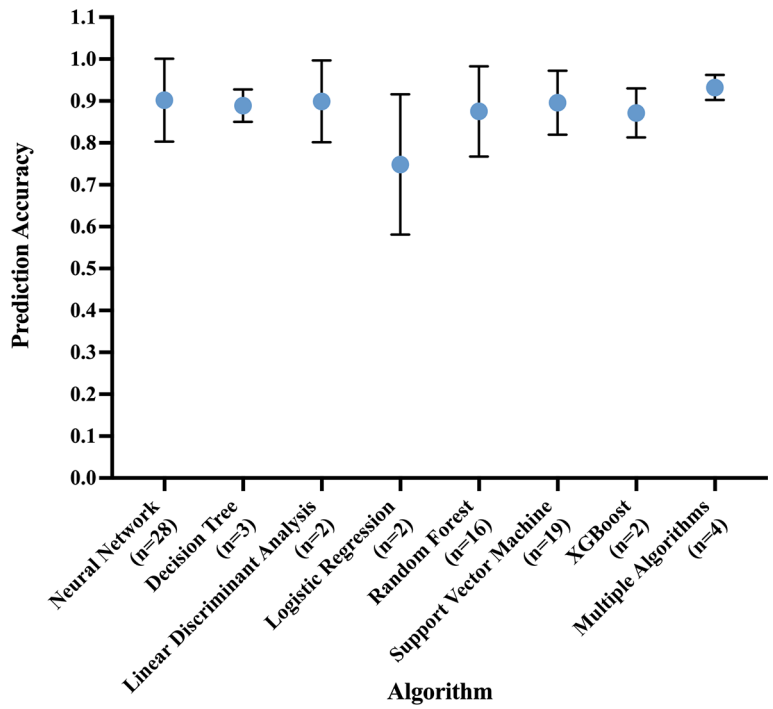

FIGURE 5 | Prediction accuracy of most common algorithm types, measured in the best performing algorithm of each study. Circle at mean. Error bars indicate standard deviation.

Titles (0%) did not identify the development of prediction models. Abstracts (0%) frequently lacked source of data, overall sample size, and calibration methods. Regarding model performance (0%), very few studies reported measures for model calibration or confidence intervals. Most studies failed to specify model regression coefficients (2.4%) or provide justifications of sample size (2.4%), e.g., how sample size was arrived at according to statistical or practical grounds. More detailed explanations of TRIPOD items and their adherence rates can be found in **Table A8 (Supplementary)**.

4 DISCUSSION

Our systematic review analyzed 85 articles describing ML applications for glioma grade prediction and revealed several trends. First, the number of studies published per year grew steadily between 2016 and 2019. Second, imaging sequences and ML models became less conventional, with the emergence of advanced MRI sequences (MR Spectroscopy, Perfusion) in the early 2000s (41) and DL models in 2018 (42). Third, datasets recently expanded to encompass multiple institutions, with BraTS and TCIA datasets appearing in 2017. While ML model studies report high predictive accuracies, they underreport critical model performance measures, lack a common validation dataset, and vary remarkably in glioma classification systems, ML algorithms, feature types and imaging sequences used for prediction, etc., limiting model comparison. Here, we identify several opportunities for improvement to prepare models for multicenter clinical adoption.

4.1 Study Datasets, Validation Techniques, Classification Systems, and Reporting Quality

Prior to broad clinical use, ML models must be trained and validated on large, multi-institutional datasets to ensure generalizability (43). Dataset sizes, however, were low in our study, and most publications lacked external validation. These findings are consistent with those from a similar systematic review by Tabatabaei et al. (23). Moreover, while studies based on highly curated datasets, including BraTS or TCIA, showed consistently high accuracy results, algorithms trained on these datasets without external validation may not have reproducible results in clinical practice, where imaging protocols are less standardized, image quality is variable, and tumor presentations are heterogeneous. To show that models perform well across distinct populations and are fit for broad clinical implementation, future works should use sizable, less-curated, multicenter datasets and externally validate their models.

ML models should also be trained according to standardized definitions of glioma grade. Interestingly, definitions were variable for HGG and LGG, with some studies considering grade 3 gliomas to be high-grade and others low-grade. Lack of a unified classification system may hinder predictive model

TABLE 2 | Characteristics of the 10 studies reporting the highest accuracy results for their best performing models, including: glioma grade classification task, dataset source and size, ratio of high- to low-grade gliomas, validation technique, imaging sequences used in prediction, feature types used in prediction, best performing algorithm (based on accuracy results), and performance metrics.

Paper	Glioma Grade Classification Task	Dataset	HGG : LGG Ratio	Validation Technique	Imaging Sequences	Features	Best Algorithm	Performance Metrics
Hedyehzadeh et al. (2020) (31)	2/3 vs. 4	TCIA (n=461 patients)	1.3:1 (262 HGG, 199 LGG in total set)	Internal (4-fold cross-validation)	T1, T1CE, T2, FLAIR	Texture	Support Vector Machine	Accuracy = 1.00 Sensitivity = 1.00 Specificity = 1.00
BashirGonbadi and Khotanlou (2019) (32)	1/2 vs. 3/4	BraTS (n=285 patients)	2.8:1 (210 HGG, 75 LGG in total set)	Internal (Holdout, 15% of dataset)	T1, T1CE, T2, FLAIR	Deep learning extracted	Convolutional Neural Network	Accuracy = 0.9918
Polly et al. (2018) (33)	HGG vs. LGG (unclear)	BraTS (n=160 images)	1:1 (50 HGG, 50 LGG in testing set)	Unspecified	T2	First-order, Shape, Texture	Support Vector Machine	Accuracy = 0.99 Sensitivity = 1.00 Specificity = 0.9803
De Looze et al. (2018) (34)	HGG vs. LGG (unclear)	Single center hospital (n=381 patients)	Unclear	Internal (5-fold cross-validation)	T1, T1CE, T2, FLAIR, Diffusion	Qualitative	Random Forest	Accuracy = 0.99 AUC = 0.99 Sensitivity = 1.00 Specificity = 0.92
Sharif et al. (2020) (35)	HGG vs. LGG (unclear)	BraTS (n=30 patients)	2.3:1 (7 HGG, 3 LGG in testing set)	Internal (Holdout, 10-fold cross-validation)	T1, T1CE, T2, FLAIR	Deep learning extracted	Convolutional Neural Network	Accuracy = 0.987
Muneer et al. (2019) (36)	1 vs. 2 vs. 3 vs. 4	Single center hospital (n=20 patients)	1.3:1.6:1:1.5 (39 grade 1, 51 grade 2, 31 grade 3, 47 grade 4 images in testing set)	Internal (Holdout, 30% of dataset)	T2	Deep learning extracted	VGG19 (Deep Convolutional Neural Network)	Accuracy = 0.9825 Sensitivity = 0.9272 Specificity = 0.9813 Positive Predictive Value = 0.9471 F1 Score = 0.9371
Dandil and Bicer (2020) (37)	1/2 vs. 3 vs. 4 vs. meningioma	INTERPRET (n=179 patients)	Unclear	Unspecified	MR Spectroscopy (Time of Echo 20ms and 136ms)	First-order, Shape and size, Texture	Long Short-Term Memory (Neural Network)	Accuracy = 0.982 AUC = 0.9936 Sensitivity = 1.00 Specificity = 0.9753
Tian et al. (2018) (38)	2 vs. 3/4	Single center hospital (n=153 patients)	2.6:1 (111 HGG, 42 LGG in total set)	Internal (10-fold cross-validation)	T1, T1CE, T2, Diffusion, Perfusion (3D Arterial Spin Labeling)	Texture	Support Vector Machine	Accuracy = 0.981 AUC = 0.992 Sensitivity = 0.987 Specificity = 0.974
Lo et al. (2019) (39)	2 vs. 3 vs. 4	TCIA (n=130 patients)	1:1.4:1.9(30 grade 2, 43 grade 3 and 57 grade 4 in total set)	Internal (10-fold cross-validation)	T1CE	Deep learning extracted	Deep Convolutional Neural Network	Accuracy = 0.979 AUC = 0.9991
Kumar et al. (2020) (40)	1/2 vs. 3/4	BraTS (n=285 patients)	2.8:1 (210 HGG, 75 LGG in total set)	Internal (5-fold cross-validation)	T1, T1CE, T2, (T2W)-FLAIR	First-order, Shape, Texture	Random Forest	Accuracy = 0.9754 AUC = 0.9748 Sensitivity = 0.9762 Specificity = 0.9733 F1 Score = 0.983

Testing or validation metrics are reported when available, otherwise training metrics are reported. HGG, high-grade gliomas; LGG, low-grade gliomas; ML, machine learning; PRISMA-DTA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy; T1CE, T1-weighted contrast-enhanced; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

performances on external datasets, given that the images used for segmentation, feature extraction, and model training/testing are labeled HGG or LGG based on non-uniform definitions. As glioma grade guides clinical management, it is essential that algorithm outputs of "HGG" and "LGG" reflect a universal definition consistent with current WHO criteria.

An alternative to binary high-or-low grading is to report numerical glioma grades (1, 2, 3, or 4) and tumor entities. Importantly, grade and entity classifications are evolving. In 2016, purely histopathological classification was succeeded by classification based on both molecular and histopathologic parameters (44). In 2021, cIMPACT-NOW established further changes to glioma grading, for example, by redefining GBM to be an IDH (Isocitrate Dehydrogenase)-wild-type lesion distinct from IDH-mutant grade 4 astrocytomas (45, 46). Classification changes may have led to inconsistencies in tumor entities and grades reported in glioma grade prediction studies across the years, limiting model comparison. As WHO criteria continue to

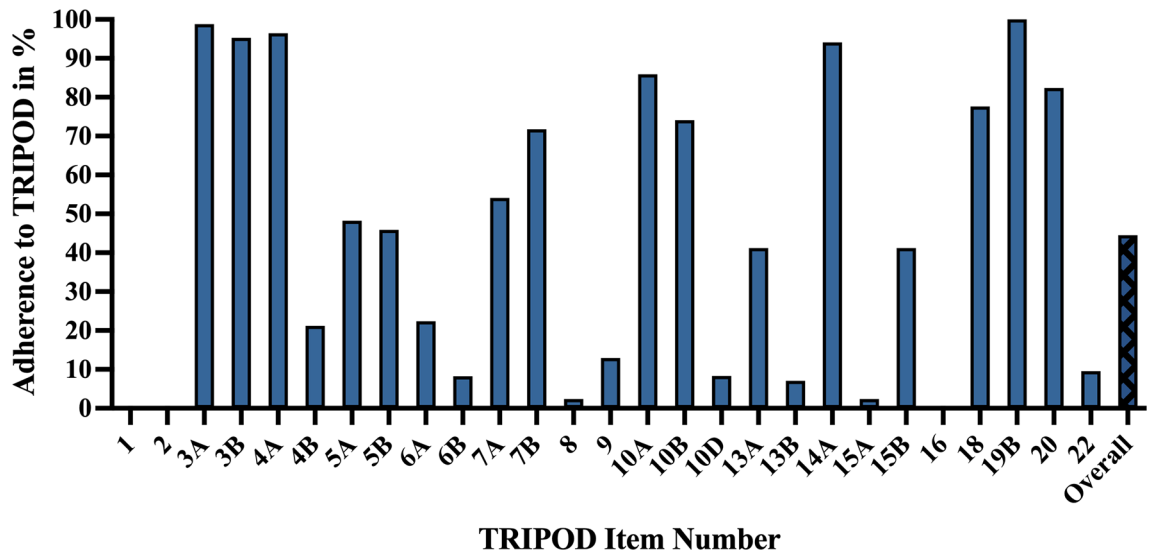

FIGURE 6 | TRIPOD adherence of machine learning glioma grade prediction studies. Adherence rate for individual items represents the percent of studies scoring a point for that item: 1 – title. 2 – abstract. 3a – background. 3b – objectives. 4a – study design. 4b – study dates. 5a – study setting. 5b – eligibility criteria. 6a – outcome assessment. 6b – blinding assessment of outcome. 7a – predictor assessment. 7b – blinding assessment of predictors. 8 – sample size justification. 9 – missing data. 10a – predictor handling. 10b – model type, model-building, and internal validation. 10d – model performance. 13a – participant flow and outcomes. 13b – participant demographics and missing data. 14a – model development (participants and outcomes). 15a – full model specification. 15b – using the model. 16 – model performance. 18 – study limitations. 19b – results interpretation. 20 – clinical use and research implications. 22 – funding. Overall – mean TRIPOD adherence rate of all studies. TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

evolve and affect generalizability of study results, we recommend that studies clearly reference the criteria used in glioma grading, report the glioma entities and corresponding grade used in model development, and describe the predictive performances by both entity and grade. This will promote comprehensible and traceable results over time. Moreover, integrated techniques that characterize disease according to both radiological and biological features are emerging in neuro-oncology (47). We advise future researchers to consider the implementation of these techniques into ML model development studies predicting glioma grade, molecular markers, response to treatment, prognosis, and other applications within neuro-oncology.

Finally, reporting of ML models should be transparent, thorough, and reproducible to facilitate proper assessment for use in clinical practice (48). Several comprehensive checklists are used to assess reporting quality of diagnostic models, including Checklist for Artificial Intelligence in Medical Imaging (49) and TRIPOD. In our study, mean adherence to TRIPOD was low, with key assessment elements such as model performance scoring poorly. These findings reflect inadequate study reporting. To address this, we recommend future studies use appropriate reporting frameworks to guide all phases of study execution, from initial design through manuscript writing. For ML studies, the relevance of TRIPOD as a benchmark for reporting quality may be questioned. Published explanations and elaborations of TRIPOD focus on regression-based models, a shortcoming that TRIPOD authors have recently acknowledged (19). We support their initiative to create a TRIPOD Statement specific to ML (TRIPOD-AI) (19, 20), and in the context of this work, to improve the reporting quality of literature concerning ML in glioma grade prediction.

4.2 Limitations

This study has several limitations. First, the timing and criteria of our search may have missed relevant studies (e.g., recent and unpublished works). Moreover, 191 of the 1,135 (16.8%) studies assessed for eligibility were excluded because they were abstracts (n=169, 14.9%) or not in English (n=22, 1.9%), creating a potential selection bias. However, full texts were required for complete data extraction and quality of reporting analysis, and we unfortunately did not have the resources to translate non-English articles. Second, we determined best performing algorithms based on accuracy, which excluded the three studies that did not report accuracy results for their models. Accuracy, furthermore, may be considered a flawed performance metric for ML models applied to imbalanced datasets (50), which constituted most datasets in our study. With imbalanced datasets, ML models intrinsically overfit toward the majority class, risking higher misclassification rates for minority classes (51, 52). Because accuracy may be high even if a minority class is poorly predicted, we recommend study authors consistently report a full slate of model performance metrics. Including metrics sensitive to performance differences within imbalanced datasets (e.g., AUC) (53) will enable a more thorough assessment of ML model performance. Third, the inconsistent definitions for HGG and LGG, evolving grading criteria, high heterogeneity of our included articles and low

number of articles reporting confidence intervals for their performance metrics limited the pooling of results across studies and subsequent generation of conclusions. As a result, we could not perform a meta-analysis (54, 55).

5 CONCLUSION

The application of ML to glioma grade prediction has grown substantially, with ML model studies reporting high predictive accuracies but lacking essential metrics and characteristics for assessing model performance. To increase the generalizability, standardization, reproducibility, and reporting quality necessary for clinical translation, future studies need to (1) train and test on large, multi-institutional datasets, (2) validate on external datasets, (3) clearly report glioma entities, corresponding glioma grades, and a full state of predictive performance metrics by both grade and entity, and (4) adhere to reporting guidelines.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

SM receives funding in part from the Biomedical Education Program (BMEP). RB receives funding in part from the National Institute of Diabetes and Digestive and Kidney Disease of the National Institutes of Health under Award Number T35DK104689. MA received funding from American Society of Neuroradiology Fellow Award 2018. This publication was made possible by KL2 TR001862 (MA) from the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH), and NIH roadmap for Medical Research. Seyedmehdi Payabvash has grant support from NIH/NINDS K23NS118056, foundation of American Society of Neuroradiology (ASNR) #1861150721, Doris Duke Charitable Foundation (DDCF) #2020097, and NVIDIA. Funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

We would like to thank our institutional librarians (Alexandria Brackett and Thomas Mead) for developing and executing our search strategy, and Mary Hughes and Vermetha Polite for their technical support. We also would like to thank Julia Shatalov for assisting with eligibility screening and Irena Tocino for her continuous support throughout the execution of this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.856231/full#supplementary-material

REFERENCES

1. Ostrom QT, Cioffi G, Gittleman H, Patil N, Waite K, Kruchko C, et al. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012-2016. *Neuro Oncol* (2019) 21 (Suppl 5):v1–v100. doi: 10.1093/neuonc/noz150
2. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary. *Neuro Oncol* (2021) 23(8):1231–51. doi: 10.1093/neuonc/noab106
3. Tran B, Rosenthal MA. Survival Comparison Between Glioblastoma Multiforme and Other Incurable Cancers. *J Clin Neurosci* (2010) 17 (4):417–21. doi: 10.1016/j.jocn.2009.09.004
4. Ohgaki H, Kleihues P. Population-Based Studies on Incidence, Survival Rates, and Genetic Alterations in Astrocytic and Oligodendroglial Gliomas. *J Neuropathol Exp Neurol* (2005) 64(6):479–89. doi: 10.1093/jnen/64.6.479
5. Gallego Perez-Larraya J, Delattre JY. Management of Elderly Patients With Gliomas. *Oncologist* (2014) 19(12):1258–67. doi: 10.1634/theoncologist.2014-0170
6. Zonari P, Baraldi P, Crisi G. Multimodal MRI in the Characterization of Glial Neoplasms: The Combined Role of Single-Voxel MR Spectroscopy, Diffusion Imaging and Echo-Planar Perfusion Imaging. *Neuroradiology* (2007) 49 (10):795–803. doi: 10.1007/s00234-007-0253-x
7. Thon N, Tonn JC, Kreth FW. The Surgical Perspective in Precision Treatment of Diffuse Gliomas. *Onco Targets Ther* (2019) 12:1497–508. doi: 10.2147/OTT.S174316
8. Hu LS, Hawkins-Daarud A, Wang L, Li J, Swanson KR. Imaging of Intratumoral Heterogeneity in High-Grade Glioma. *Cancer Lett* (2020) 477:97–106. doi: 10.1016/j.canlet.2020.02.025
9. Law M, Yang S, Wang H, Babb JS, Johnson G, Cha S, et al. Glioma Grading: Sensitivity, Specificity, and Predictive Values of Perfusion MR Imaging and Proton MR Spectroscopic Imaging Compared With Conventional MR Imaging. *AJNR Am J Neuroradiol* (2003) 24(10):1989–98.
10. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More Than Pictures, They Are Data. *Radiology* (2016) 278(2):563–77. doi: 10.1148/radiol.2015151169
11. Giger ML. Machine Learning in Medical Imaging. *J Am Coll Radiol* (2018) 15 (3 Pt B):512–20. doi: 10.1016/j.jacr.2017.12.028
12. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep Learning: A Primer for Radiologists. *Radiographics* (2017) 37(7):2113–31. doi: 10.1148/rg.2017170077
13. Cheng PM, Montagnon E, Yamashita R, Pan I, Cadrin-Chênevert A, Perdigón Romero F, et al. Deep Learning: An Update for Radiologists. *Radiographics* (2021) 41(5):1427–45. doi: 10.1148/rg.2021200210
14. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The Practical Implementation of Artificial Intelligence Technologies in Medicine. *Nat Med* (2019) 25(1):30–6. doi: 10.1038/s41591-018-0307-0
15. Lasocki A, Tsui A, Tacey MA, Drummond KJ, Field KM, Gaillard F. MRI Grading Versus Histology: Predicting Survival of World Health Organization Grade II-IV Astrocytomas. *AJNR Am J Neuroradiol* (2015) 36(1):77–83. doi: 10.3174/ajnr.A4077

16. Jin W, Fatehi M, Abhishek K, Mallya M, Toyota B, Hamarneh G. Artificial Intelligence in Glioma Imaging: Challenges and Advances. *J Neural Eng* (2020) 17(2):021002. doi: 10.1088/1741-2552/ab8131
17. Park JE, Kim HS, Kim D, Park SY, Kim JY, Cho SJ, et al. A Systematic Review Reporting Quality of Radiomics Research in Neuro-Oncology: Toward Clinical Utility and Quality Improvement Using High-Dimensional Imaging Features. *BMC Cancer* (2020) 20(1):29. doi: 10.1186/s12885-019-6504-5
18. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* (2015) 162(1):55–63. doi: 10.7326/M14-0697
19. Collins GS, Moons KGM. Reporting of Artificial Intelligence Prediction Models. *Lancet* (2019) 393(10181):1577–9. doi: 10.1016/S0140-6736(19)30037-6
20. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, 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(7):e048008. doi: 10.1136/bmjopen-2020-048008
21. Buchlak QD, Esmaili N, Leveque JC, Bennett C, Farrokhi F, Piccardi M. Machine Learning Applications to Neuroimaging for Glioma Detection and Classification: An Artificial Intelligence Augmented Systematic Review. *J Clin Neurosci* (2021) 89:177–98. doi: 10.1016/j.jocn.2021.04.043
22. Sohn CK, Bisdas S. Diagnostic Accuracy of Machine Learning-Based Radiomics in Grading Gliomas: Systematic Review and Meta-Analysis. *Contrast Media Mol Imaging* (2020) 2020:2127062. doi: 10.1155/2020/2127062
23. Tabatabaei M, Razaei A, Sarrami AH, Saadatpour Z, Singhal A, Sotoudeh H. Current Status and Quality of Machine Learning-Based Radiomics Studies for Glioma Grading: A Systematic Review. *Oncology* (2021) 99(7):433. doi: 10.1159/000515597
24. Frank RA, Bossuyt PM, McInnes MDF. Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy: The PRISMA-DTA Statement. *Radiology* (2018) 289(2):313–4. doi: 10.1148/radiol.2018180850
25. Whiting P, Westwood M, Burke M, Sterne J, Glanville J. Systematic Reviews of Test Accuracy Should Search a Range of Databases to Identify Primary Studies. *J Clin Epidemiol* (2008) 61(4):357–64. doi: 10.1016/j.jclinepi.2007.05.013
26. TRIPOD Statement Web Site. *Adherence to TRIPOD* (2020). Available at: https://www.tripod-statement.org/adherence/ (Accessed 10 Jul 2021).
27. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* (2015) 162(1):W1–73. doi: 10.7326/M14-0698
28. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE T Med Imaging* (2015) 34(10):1993–2024. doi: 10.1109/Tmi.2014.2377694
29. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digital Imaging* (2013) 26(6):1045–57. doi: 10.1007/s10278-013-9622-7
30. Park YW, Choi YS, Ahn SS, Chang JH, Kim SH, Lee SK. Radiomics MRI Phenotyping With Machine Learning to Predict the Grade of Lower-Grade Gliomas: A Study Focused on Nonenhancing Tumors. *Korean J Radiol* (2019) 20(9):1381–9. doi: 10.3348/kjr.2018.0814
31. Hedyehzadeh M, Nezhad SYD, Safdarian N. Evaluation of Conventional Machine Learning Methods for Brain Tumour Type Classification. *Cr Acad Bulg Sci* (2020) 73(6):856–65. doi: 10.7546/Crabs.2020.06.14
32. Bashir Gonbadi F, Khotanlou H. Glioma Brain Tumors Diagnosis and Classification in MR Images Based on Convolutional Neural Networks. *9th International Conference on Computer and Knowledge Engineering* (Iccke 2019) (2019) 1–5. doi: 10.1109/ICCKE48569.2019.8965143
33. Polly FP, Shil SK, Hossain MA, Ayman A, Jang YM. Detection and Classification of HGG and LGG Brain Tumor Using Machine Learning. *32nd International Conference on Information Networking (Icoin)*, (2018) 813–7. doi: 10.1109/ICOIN.2018.8343231
34. De Looze C, Beausang A, Cryan J, Loftus T, Buckley PG, Farrell M, et al. Machine Learning: A Useful Radiological Adjunct in Determination of a Newly Diagnosed Glioma's Grade and IDH Status. *J Neuro-Oncol* (2018) 139 (2):491–9. doi: 10.1007/s11060-018-2895-4
35. Sharif MI, Li JP, Khan MA, Saleem MA. Active Deep Neural Network Features Selection for Segmentation and Recognition of Brain Tumors Using MRI Images. *Pattern Recogn Lett* (2020) 129:181–9. doi: 10.1016/j.patrec.2019.11.019
36. Muneer KVA, Rajendran VR, Joseph KP. Glioma Tumor Grade Identification Using Artificial Intelligent Techniques. *J Med Syst* (2019) 43(5). doi: ARTN 113 10.1007/s10916-019-1228-2
37. Dandil E, Bicer A. Automatic Grading of Brain Tumours Using LSTM Neural Networks on Magnetic Resonance Spectroscopy Signals. *Iet Image Process* (2020) 14(10):1967–79. doi: 10.1049/iet-ipr.2019.1416
38. Tian Q, Yan LF, Zhang X, Zhang X, Hu YC, Han Y, et al. Radiomics Strategy for Glioma Grading Using Texture Features From Multiparametric MRI. *J Magnetic Resonance Imaging* (2018) 48(6):1518–28. doi: 10.1002/jmri.26010
39. Lo CM, Chen YC, Weng RC, Hsieh KLC. Intelligent Glioma Grading Based on Deep Transfer Learning of MRI Radiomic Features. *Appl Sci-Basel* (2019) 9 (22). doi: ARTN 4926 10.3390/app9224926
40. Kumar R, Gupta A, Arora HS, Pandian GN, Raman B. CGHF: A Computational Decision Support System for Glioma Classification Using Hybrid Radiomics- and Stationary Wavelet-Based Features. IEEE Access (2020) 8:79440–58. doi: ARTN 4926 10.1109/Access.2020.2989193
41. Devos A, Simonetti AW, van der Graaf M, Lukas L, Suykens JAK, Vanhamme L, et al. The Use of Multivariate MR Imaging Intensities Versus Metabolic Data From MR Spectroscopic Imaging for Brain Tumour Classification. *J Magn Reson* (2005) 173(2):218–28. doi: 10.1016/j.jmr.2004.12.007
42. Ge C, Gu IY, Jakola AS, Yang J. Deep Learning and Multi-Sensor Fusion for Glioma Classification Using Multistream 2d Convolutional Networks. *Annu Int Conf IEEE Eng Med Biol Soc* (2018) 2018:5894–7. doi: 10.1109/EMBC.2018.8513556
43. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology* (2018) 288(2):318–28. doi: 10.1148/radiol.2018171820
44. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary. *Acta Neuropathol* (2016) 131(6):803–20. doi: 10.1007/s00401-016-1545-1
45. Brat DJ, Aldape K, Colman H, Figrarella-Branger D, Fuller GN, Giannini C, et al. cIMPACT-NOW Update 5: Recommended Grading Criteria and Terminologies for IDH-Mutant Astrocytomas. *Acta Neuropathol* (2020) 139 (3):603–8. doi: 10.1007/s00401-020-02127-9
46. Weller M, van den Bent M, Preusser M, Le Rhun E, Tonn JC, Minniti G, et al. EANO Guidelines on the Diagnosis and Treatment of Diffuse Gliomas of Adulthood. *Nat Rev Clin Oncol* (2021) 18(3):170–86. doi: 10.1038/s41571-020-00447-z
47. Maggio I, Franceschi E, Gatto L, Tosoni A, Di Nunno V, Tonon C, et al. Radiomics, Mirnomics, and Radiomirrnomics in Glioblastoma: Defining Tumor Biology From Shadow to Light. *Expert Rev Anticancer Ther* (2021) 21:1265–72. doi: 10.1080/14737140.2021.1971518
48. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* (2011) 155(8):529–36. doi: 10.7326/0003-4819-155-8-201110180-00009
49. Mongan J, Moy L, Kahn CEJr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell* (2020) 2(2):e200029. doi: 10.1148/ryai.2020200029
50. Saito T, Rehmsmeier M. The Precision-Recall Plot is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PloS One* (2015) 10(3):e0118432. doi: 10.1371/journal.pone.0118432
51. Knowler WC, Pettitt DJ, Savage PJ, Bennett PH. Diabetes Incidence in Pima-Indians - Contributions of Obesity and Parental Diabetes. *Am J Epidemiol* (1981) 113(2):144–56. doi: 10.1093/oxfordjournals.aje.a113079
52. Li DC, Liu CW, Hu SC. A Learning Method for the Class Imbalance Problem With Medical Data Sets. *Comput Biol Med* (2010) 5):509–18. doi: 10.1016/j.compbiomed.2010.03.005
53. Ling CX, Huang J, Zhang H. AUC: A Better Measure Than Accuracy in Comparing Learning Algorithms. *Lect Notes Artif Int* (2003) 2671:329–41. doi: 10.1007/3-540-44886-1_25

54. Cronin P, Kelly AM, Altaee D, Foerster B, Petrou M, Dwamena BA. How to Perform a Systematic Review and Meta-Analysis of Diagnostic Imaging Studies. *Acad Radiol* (2018) 25(5):573–93. doi: 10.1016/j.acra.2017.12.007
55. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi YThe Cochrane Collaboration. Chapter 10: Analysing and Presenting Results. In: JJ Deeks, PM Bossuyt, C Gatsonis, editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0* Birmingham, UK: The Cochrane Collaboration (2010). Available at: http://srdta.cochrane.org/

Author Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: Author ML is an employee and stockholder of Visage Imaging, Inc., and unrelated to this work, receives funding from NIH/NCI R01 CA206180 and is a board member of Tau Beta Pi engineering honor society. KB is an employee of Visage Imaging, GmbH. JI has funding support for an investigator-initiated clinical trial from Novartis Pharmaceuticals (unrelated to this work).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.