


Review article

Can we rely on machine learning algorithms as a trustworthy predictor for recurrence in high-grade glioma? A systematic review and meta-analysis

Ibrahim Mohammadzadeh^{a,*}, Behnaz Niroomand^{a,2}, Bardia Hajikarimloo^b,
 Mohammad Amin Habibi^{c,3}, Ali Mortezaei^{d,4}, Jina Behjati^{e,5}, Abdulrahman Albakr^f,
 Hamid Borghei-Razavi^{f,**} 

^a Skull Base Research Center, Loghman-Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran

^b Department of Neurological Surgery, University of Virginia, Charlottesville, VA, USA

^c Department of Neurosurgery, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran

^d Student Research Committee, Gonabad University of Medical Sciences, Gonabad, Iran

^e Functional Neurosurgery Research Center, Shohada Tajrish Comprehensive Neurosurgical Center of Excellence, Shahid Beheshti University of Medical Sciences, Tehran, Iran

^f Department of Neurological Surgery, Pauline Braathen Neurological Center, Cleveland Clinic Florida, Weston, FL, USA

ARTICLE INFO

Keywords:

High grade glioma
 Recurrence
 Brain tumor
 Prediction
 Machine learning
 Artificial intelligence

ABSTRACT

Early prediction of recurrence in high-grade glioma (HGG) is critical due to its aggressive nature and poor prognosis. Distinguishing true recurrence from treatment-related changes, such as radionecrosis, is a major diagnostic challenge. Machine learning (ML) offers a novel approach, leveraging advanced algorithms to analyze complex imaging data with high precision. A comprehensive search of PubMed, Embase, Scopus, Web of Science, and Google Scholar identified eligible studies. The sensitivity, specificity, accuracy, precision, F1 score, and the (area under the curve) AUC items were extracted from the included studies. After screening 1077 records, seven studies met the eligibility criteria for the systematic review, of which five were included in the meta-analysis. ML algorithm, particularly Support Vector Machines (SVM), demonstrated promising performance. A meta-analysis of five studies revealed a pooled sensitivity of 0.95 (95% CI: 0.84–0.99) and specificity of 0.80 (95% CI: 0.69–0.88). Additionally, the positive diagnostic likelihood ratio (DLR) was 4.75 (95% CI: 2.91–7.76), the negative DLR was 0.06 (95% CI: 0.02–0.21), and the diagnostic odds ratio was 80.97 (95% CI: 17.5–374.61). The diagnostic score was calculated as 4.39 (95% CI: 2.86–5.93), and the AUC was 0.86 (95% CI: 0.83–0.89). Subgroup analyses showed SVM-based models with higher sensitivity (0.98 vs. 0.87) and specificity (0.82 vs. 0.77) than non-SVM ($p = 0.13$). Comparing glioblastoma and Grade 3 tumors, sensitivities were 94% vs. 97%, and specificities were 79% vs. 83%, with no significant heterogeneity. These findings suggest that ML models, particularly SVM, offer promising diagnostic performance in distinguishing true tumor recurrence from treatment-related changes.

* Correspondence to: Medical doctor, skull-base Neurosurgery from Shahid Beheshti medical university, Tehran, Iran.

** Correspondence to: Associate Professor of Neurological Surgery at CCLCM of CWRU Director of Minimally Invasive Cranial and Pituitary Surgery Program Research Director, Neuroscience Institute, Cleveland Clinic Florida Region, USA

E-mail addresses: ibrahim.mohammadzadeh@sbmu.ac.ir, ibrahim.mdz7777@gmail.com (I. Mohammadzadeh), behnaz.nrm@gmail.com (B. Niroomand), kjh7vp@uvahealth.org (B. Hajikarimloo), Mohammad.habibi1392@yahoo.com (M.A. Habibi), alimortezaei97@yahoo.com (A. Mortezaei), Jina.behjati@gmail.com (J. Behjati), Dr.aalbakr@gmail.com (A. Albakr), borgheh2@ccf.org (H. Borghei-Razavi).

¹ ORCID ID: 0000-0002-8862-0778

² ORCID ID: 0000-0001-5184-7445

³ ORCID ID: 0000-0001-7600-6925

⁴ ORCID ID: 0000-0002-7217-3264

⁵ ORCID ID: 0009-0007-8408-0077

1. Introduction

Glioma, the most common type of brain and spinal cord neoplasm, is the name for all tumors that originate from glial cells [1]. Intracranial gliomas are classified based on histological features, such as astrocytic or oligodendrocytic origin, and according to the World Health Organization (WHO) grading system (grades 1–4). Of these, grades 3 and 4 are designated as high-grade gliomas (HGGs) [2]. These subtypes differ significantly in histology, cytological characteristics, and genetic profiles. For instance, glioblastoma (GBM) is highly aggressive and classified as IDH-wildtype, while astrocytomas are IDH-mutant [2]. About 57% of all gliomas are GBM [1], the grade IV glioma, which is characterized by its rapid growth, with an average survival time of 12–15 months [3,4]. The poor prognosis of HGGs is due to high recurrence rate, tumors' biological complexity and therapeutic evasion which emphasizes the critical need for innovative and more effective treatment strategies to address the devastating impact of HGG [4]. The treatment for HGG may cause treatment-related changes (e.g., radionecrosis, pseudoprogression) on standard MRI sequences, which harden the identification of tumor progression [5].

Artificial intelligence (AI) has emerged as a transformative tool in medical diagnostics, offering significant advancements in the detection and management of complex and critical conditions [6,7]. In the context of HGGs, AI plays a vital role in addressing diagnostic challenges, particularly in distinguishing tumor progression from treatment-related changes and detecting recurrence [8,9]. AI algorithms with the use of machine learning (ML) and deep learning (DL) models have demonstrated remarkable accuracy in analyzing medical imaging data, such as MRI and CT scans, facilitating early and precise diagnosis [10]. This is especially critical in the management of recurrent HGGs, where traditional diagnostic methods often struggle to differentiate between true tumor regrowth and post-treatment effects. Moreover, AI models can predict recurrence risk by integrating imaging landmarks, biomarkers, genomic profiles, and clinical data, enabling timely and personalized interventions [8,9,11].

In this study, our main goal was to demonstrate whether AI can be reliably used as a tool to predict the recurrence of HGG. In this regard, we conducted a systematic review and meta-analysis to assess the effectiveness and accuracy of AI algorithms in predicting the likelihood of recurrence in patients with HGG. The primary goal was to highlight how AI-based models can assist in clinical decision-making by providing more reliable, efficient, and accurate predictions of recurrence outcomes. Additionally, this study provides valuable insights for neurosurgeons and radiologists by helping distinguish true tumor recurrence from treatment-related changes. It can guide surgical planning, enhance imaging interpretation, and improve overall patient management.

2. Methods

This systematic review and meta-analysis focused on assessing the diagnostic accuracy of ML algorithms in predicting the recurrence of HGG. It has been internationally registered with PROSPERO Code CRD42024618667 and follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a standardized and transparent methodology [12].

2.1. Search strategy

Four main databases of medical literature, including PubMed, Web of Science, Scopus, and Embase, were systematically searched from their inception until December 2, 2024. The main keywords of this study were "High-grade glioma", "glioblastoma", "recurrent", "Artificial intelligence", "Machine learning", and "Deep learning". The Medical Subject Headings (MeSH) in PubMed and Emtree terms in Embase were utilized, along with other keywords, to tailor unique search strategies for each database. No restrictions were applied regarding publication date,

language, or publication type. Additionally, the first 100 results from Google Scholar were reviewed as part of a supplementary search. The complete search strategy syntax is provided in [Supplementary Table S1](#).

3. Inclusion criteria

Studies were considered for inclusion in this systematic review and meta-analysis if they fulfill the following criteria: 1) Original articles such as cohort studies, randomized clinical trials and case-series 2) Conducted on patients diagnosed with HGG, 4) Reporting all type of ML algorithms; i.e., Random Forest (RF), convolutional Neural Network (DNN), Decision Tree (DT), K-nearest neighbors (KNN), Support Vector Machine (SVM) in predicting recurrent HGG.

3.1. Exclusion criteria

Studies were excluded if they had the following criteria for exit: 1) non-original articles, such as reviews, editorials, or case reports; 2) Involving cases of glioma classified as low-grade (Grade I, II); 3) The exact grade of the glioma is not specified; 4) Lacking sufficient data or outcomes related for prediction of recurrence; and 5) Research not using ML algorithms for prediction of recurrence.

3.2. Study selection

All database records were imported into EndNote 21 software. Two reviewers (BN and BH), according to the eligibility criteria, independently screened the articles. First, Duplicate records were removed, and then a screening process was conducted using titles, abstracts and the full text of the selected articles. Studies that met the eligibility criteria were delivered to the data extraction. Any disagreements between the two reviewers were resolved by involving a third reviewer (IM).

3.3. Data extraction

The data extracted from eligible articles included the first author's name, publication year, country, number of patients, mean age, gender composition, study design, and reference imaging modality (e.g., computed tomography (CT) or magnetic resonance imaging (MRI)). The data also included performance metrics such as sensitivity, specificity, accuracy, precision, F1 score, and AUC. The items were registered into a predesigned Excel sheet. Two authors (BN and IM) separately reviewed and extracted the data of each article.

3.4. Quality assessment

The risk of bias (ROB) for this study was assessed through the PROBAST tool, which evaluates articles on four aspects, including participants, predictors, results, and analysis [13]. Each domain was assessed for ROB, categorizing articles as low, high, or unclear ROB while also evaluating the alignment of prediction models with the research question.

3.5. Statistical analysis

The meta-analysis was performed for the best-performing prognostic model from each study. The true-and-false positive true-and-false negative values that were calculated from sensitivity and specificity were pooled in the analysis. All statistical analyses were conducted using the MIDAS package in STATA version 17.

4. Result

4.1. Study characteristics

The systematic search yielded 1077 records, of which 28 articles

have been evaluated by their full text for eligibility. Seven articles met the criteria for inclusion in the systematic review, and five were included in the meta-analysis. The PRISMA flow diagram is presented in Fig. 1. All included studies were retrospective cohorts. The sample sizes of the included literature ranged from 29 to 134 participants, comprising a total of 533 patients, with females comprising 39.8% of the population (female-to-male ratio: 0.66). The mean age of the participants was 54.3 years. Three studies were from the United States, two from China, and one each from Australia and India. Out of the eight identified algorithms, only five were included in the meta-analysis, as the highest-performing algorithm from each study was selected for inclusion.

Out of the eight algorithms, six algorithms were machine learning (RF, SVM, DT, RBF and KNN) and two of them were deep learning (CNN

and MFFE U-Net). The SVM was the most frequently used algorithm, featured in three studies (Fig. 2) (Tables 1 and 2).

SVM emerged as the best-performing algorithm among the included studies, achieving the highest average values for accuracy (0.9), sensitivity (1.0), specificity (0.933), and AUC (0.965). It was followed by CNN, which demonstrated an accuracy of 0.82, sensitivity of 1.0, specificity of 0.6, and AUC of 0.8 (Fig. 3).

MRI, as the basis of the diagnostic modalities, was the imaging reference in 100% of studies. Among the validation methods, three-fold cross-validation was the most frequent approach for performance evaluation in the presented encoders (three out of the seven studies). Additionally, 5- and 10-fold cross-validation were also employed in other studies. Radiomics-based features were the most common input

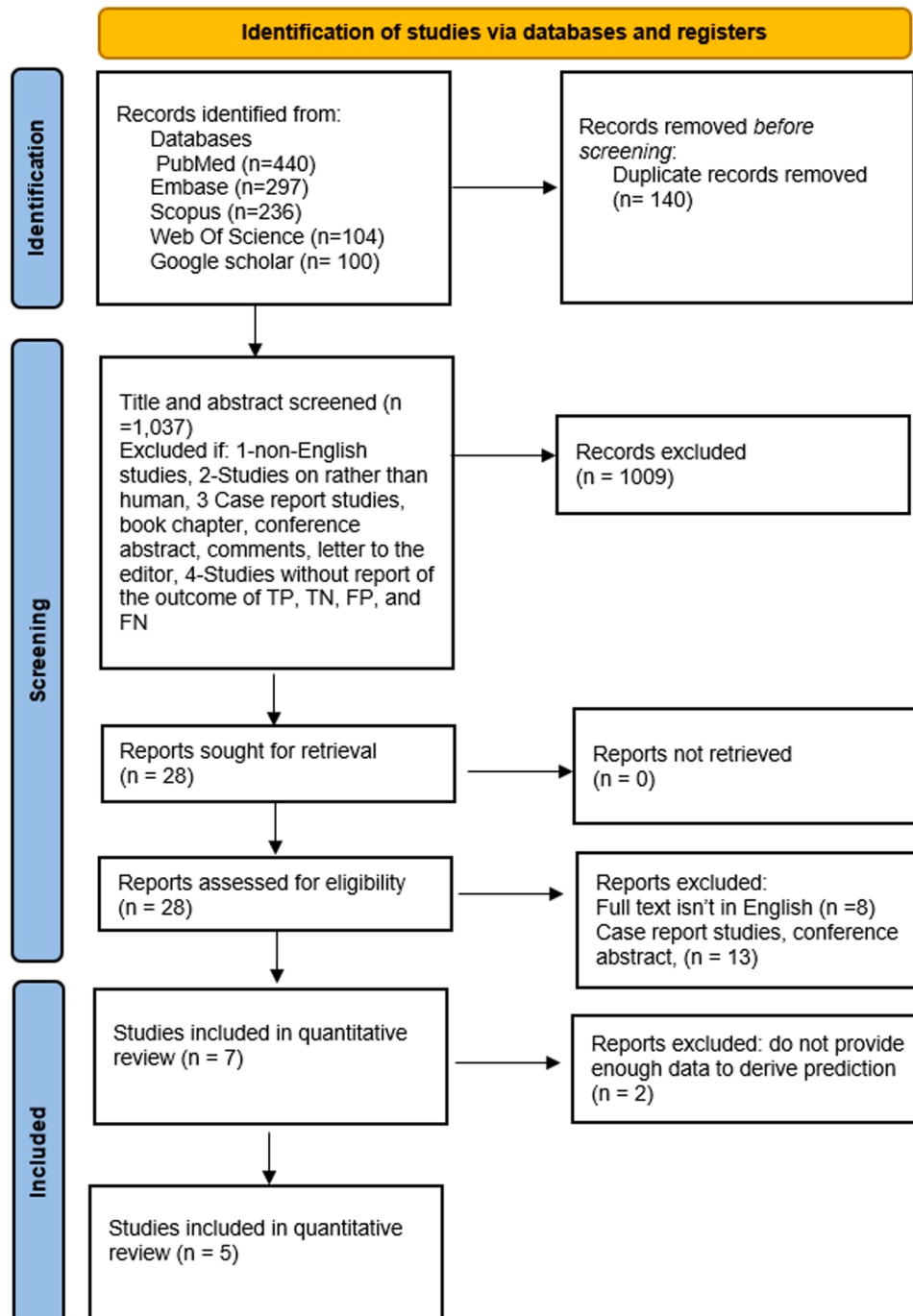


Fig. 1. PRISMA flowchart of the study selection process.

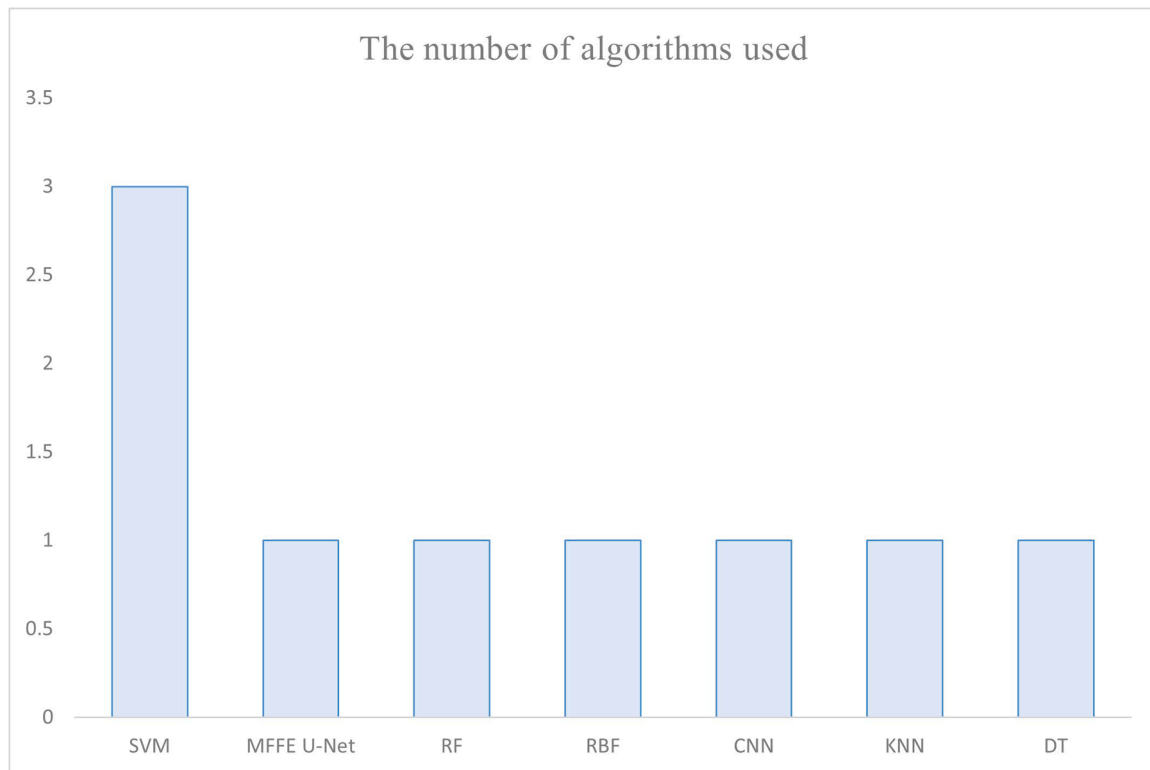


Fig. 2. Frequency of algorithms used in the analyzed studies.

characteristics employed in 5 out of the 7 studies, highlighting their key role in predictive modeling and analysis.

4.2. Sensitivity and specificity

The result of the meta-analysis demonstrated a pooled sensitivity of 0.95 [95 % CI: 0.84–0.99], with considerable heterogeneity noted with an I^2 of 59.06 [95 % CI: 18.68–99.44]. The χ^2 test of heterogeneity had a Q and degrees of freedom (df) of 9.77 and 4, respectively (p -value <0.001). The pooled specificity was 0.8 [95 % CI 0.69–0.88], with very low heterogeneity observed between the studies, with an I^2 value of 0 [95 % CI: 0–100]. The χ^2 test of heterogeneity had a Q and degrees of freedom (df) of 3.55 and 4, respectively (p value <0.001) (Fig. 4).

4.3. Positive and negative diagnostic likelihood ratio (DLR)

The positive likelihood ratio of the ML models was 4.75 [95 % CI: 2.91–7.76], with very low heterogeneity (I^2 = 0, 95 % CI: 0–100). The χ^2 test of heterogeneity had a Q and df of 4.49 and 4, respectively (p value <0.001). On the other hand, the pooled negative likelihood ratio was 0.06 [95 % CI: 0.02–0.21], with moderate heterogeneity (I^2 = 54.84, 95 % CI: 9.83–99.85). The χ^2 test of heterogeneity had a Q and df of 8.86 and 4, respectively (p value <0.001) (Fig. 5).

4.4. Diagnostic score and diagnostic odds ratio (DOR)

The diagnostic score obtained from the pooled data was 4.39 [95 % CI: 2.86–5.93], with significant heterogeneity (I^2 = 95.21 %, 95 % CI: 92.44–97.99). The χ^2 test of heterogeneity had a Q and df of 83.56 and 4, respectively (p value <0.001). The DOR of the was 80.97 [95 % CI: 17.5–374.61], with severe heterogeneity (I^2 = 100 %, 95 % CI: 100–100). The χ^2 test of heterogeneity had a Q and df of 3.5e + 14 and 4, respectively (p value <0.001) (Fig. 6).

4.5. Area under curve

The pooled area under the SROC curve, which reflects combined sensitivity and specificity, yielded an AUC of 0.86 (95 % CI: 0.83–0.89) (Fig. 7).

4.6. Subgroup analysis between studies utilizing SVM and Non-SVM algorithms

A subgroup analysis was performed to evaluate the sensitivity and specificity of machine learning models based on the presence or absence of SVM algorithms. Studies utilizing SVM (n = 2) demonstrated a pooled sensitivity of 0.98 (95 % CI: 0.95–1.00) and a specificity of 0.82 (95 % CI: 0.71–0.94). In contrast, studies without SVM (n = 3) showed a pooled sensitivity of 0.87 (95 % CI: 0.73–1.00) and a specificity of 0.77 (95 % CI: 0.62–0.92). The joint model revealed a likelihood ratio test (LRT Chi-squared) value of 4.11 (p value= 0.13), suggesting no statistically significant difference between the two groups. However, the heterogeneity (I^2) among studies utilizing SVM was 51 %, indicating moderate variability in the included studies.

4.7. Subgroup analysis: GBM vs. Grade 3 tumors

The subgroup analysis comparing GBM and Grade 3 tumors included two and three studies, respectively. The sensitivity for detecting GBM was 94 % (95 % CI: 86 %–100 %), with a specificity of 79 % (95 % CI: 67 %–90 %). In contrast, the sensitivity for Grade 3 tumors was 97 % (95 % CI: 91 %–100 %), and the specificity was 83 % (95 % CI: 67 %–98 %). The p -value for sensitivity in the GBM group was 0.98, indicating no significant heterogeneity, while the p -value for specificity was 0.29. The joint model analysis showed no substantial heterogeneity across the studies (I^2 = 0 %), highlighting consistency in the findings. These results suggest comparable diagnostic performance across the two subgroups, with high sensitivity and moderate specificity.

Table 1
Demographic and HGG recurrent characteristics.

Author/ Year	Type of study	Country	No. of patients in train/ test group	Recurrence criteria	Follow up (Mo)	Mean age /female %	Inclusion criteria	Exclusion criteria	Type of treatment	Grade of glioma (WHO Grade)
S. Rathore et al. [14]/2018	Retrospective	US	31/ 59	Histopathology-proven recurrence; follow-up MRI showing recurrence in pre-defined R-ROI	NA	57.05/ 47.2	Histopathological diagnosis of glioblastoma, no history of prior tumor or surgery, and availability of preoperative and postoperative MRI data	Residual tumor after surgery	Gross total resection followed by temozolomide-based chemoradiotherapy	IV
S. Bacchi et al. [5]/2019	Retrospective	Australia	44/ 11	NA	NA	56/ 54.6	NA	NA	NA	III or IV
Chougule et al. [15]/2022	Retrospective	India	23/ 6	a greater than 25 % increase in the sum of the products of the perpendicular diameters of the enhancing lesions with the smallest tumor measurement; (b) an appearance of any new lesion beyond the margin of the surgical resection in the new location; or (c) progressive increase in rCBV values upon repeat imaging	12	NA/ 17.24	diagnosed with GBM based on the 2016 WHO classifications, receiving the standard treatment, which includes radiotherapy with concomitant adjuvant temozolomide after gross total surgical resection of the tumor region defined by T1 perfusion MRI (rCBV maps) ^{22,23} ; and (c) having follow-up periods of more than 1-year postsurgery	Patients with no confirmed histology and poor Karnofsky Performance Scale with poor follow-up	Surgical resection, radiotherapy, temozolomide	III or IV
Y. Lao et al. [16]/2022	Retrospective	US	20/ 30	Clinically confirmed recurrence through MRI scans compared to follow-up proximity-based models	NA	50.5/ 20	Histologically confirmed GBM, Post-operative MRI and recurrence scans available, Follow-up scans acquired ≥ 3 months post-treatment	Poor MRI quality, Pseudo-progression within 3 months post-radiotherapy	Surgical resection followed by radiotherapy	III or IV
J. Ren et al. [8]/2023	Retrospective	China	90/ 40	RANO criteria: $\geq 25\%$ increase in focal enhancement or mass effect after 3 months	> 3 months	54.4/ 38	Patients underwent surgical treatment, pathologically confirmed grade II-IV gliomas according to the 2021 WHO classification, treated with radiotherapy or chemotherapy, performed MRI routine and contrast-enhanced scans after adjuvant therapy, and received more than 3 months of follow-up.	the lesion located under the curtain or in the brain stem, incomplete surgical resection, large artifacts or poor-quality images did not meet diagnostic requirements, had other types of central nervous system diseases and loss of follow-up	Surgical resection followed by chemoradiotherapy	II-IV
P. Du et al. [17]/2023	Retrospective	China	98/ 36	Recurrence determined based on follow-up MRI using RANO criteria within one year after surgery	12	NA/ 62	Age ≥ 18 years, Confirmed GBM diagnosis, Complete tumor resection, pre- and post-operative MRI data available, Standardized post-operative therapy	History of brain tumors, predominantly hemorrhagic lesions, Incomplete data	Surgical resection, followed by radiotherapy and TMZ chemotherapy	IV
Ch. Jiao et al. [18]/2024	Retrospective	US	35/ 10	Manual annotation of recurrence on T1ce, proximity-based estimation	NA	NA	Confirmed GBM recurrence, pre- and post-surgery MRI available, High-risk recurrence annotation	Poor MRI quality, Missing key MR sequences	Surgical resection and radiotherapy	IV

Abbreviation: GBM: Glioblastoma, MRI: Magnetic Resonance Imaging, rCBV: Relative Cerebral Blood Volume, T1ce: T1-weighted Contrast-Enhanced, TMZ: Temozolomide, WHO: World Health Organization, RANO: Response Assessment in Neuro-Oncology, Mo: Month, ROI: Region of Interest, US: United States.

4.8. Publication bias assessment

Deeks' test for funnel plot asymmetry showed no significant publication bias ($P = 0.25$). While the bias coefficient was not statistically

significant (8.72, 95% CI: -8.53 to 25.97, $P = 0.315$), the intercept was significant (1.94, 95% CI: 0.27 to 3.60, $P = 0.023$), suggesting potential bias despite no strong evidence of small-study effects. (Fig. 8).

Table 2
AI algorithms characteristics and performance metrics.

Author/Year	Validation	Type of reference	Input characteristics	Selected features	Method of radiomics	No. of extracted/ final (radiomics features)	Sequences analyzed	AI algorithm	Best AI predictor	Accuracy	Sensitivity (Recall)	Specificity	Precision	F1score	AUROC
S. Rathore et al. [14]/ 2018	Leave-One-Out Cross Validation (LOOCV)	MRI	Radiomics	Distance from tumor, intensity measures, texture features, and diffusion measures	Automated	More than 20 voxel-based radiomic features from multiparametric MRI	T1, T1CE, T2, T2-FLAIR, DTI, DSC-MRI	SVM with Gaussian kernel	SVM	0.8954	0.9706	0.7673	NA	NA	0.91
S. Bacchi et al. [5]/2019	5-fold cross-validation	MRI	Radiomics-based features derived from MRI sequences	DWI + FLAIR	Automated	NA	DWI, FLAIR, Post-contrast T1, ADC, DWI+FLAIR, DWI+Post-contrast T1, DWI+ADC	CNN	CNN	0.82	1	0.6	NA	0.86	0.8
Chougule et al. [15]/2022	3-fold crossvalidation	MRI	Radiomics	First-order and GLCM features for local recurrence, GLDM features for distant recurrence	Automated	133	T1CE, FLAIR, ADC	RF	RF	0.714	0.66	0.74	NA	NA	NA
Y. Lao et al. [16]/2022	10-fold cross-validation	MRI	Voxel-level features including intensity, proximity to stem cell niches (SCN), and tumor cavity	High-risk recurrence regions (HRRs) based on SCN and tumor cavity proximity	Automated	NA	T1, T1CE, T2, T2-FLAIR, ADC	RBF, SVM	SVM	NA	0.8	NA	0.69	0.73	NA
J. Ren et al. [8]/2023	5-fold cross-validation	MRI	Radiomics	72 key radiomics features (38 from PoE, 34 from ED)	Automated	1316/ 72	T1WI, CE-T1WI, T2WI, T2-FLAIR, Multimodality	SVM, KNN	SVM	0.9	100	0.933	NA	NA	0.965
P. Du et al. [17]/2023	3-fold cross-validation	MRI	Radiomics and clinicopathological features	12 optimal radiomics features and 5 clinical predictors (age, Rad-score, MGMT promoter methylation, KPS, TERT mutation)	Semi-automated	4306/12	CE-T1WI, T2-Flair, and DWI	DT	DT	0.833	0.867	0.81	NA	NA	0.719
Ch. Jiao et al. [18]/2024	3-fold cross-validation	MRI	Radiomics features derived from multi-modal MRI and stem cell niche proximity estimation	High-risk recurrence (HRR) features from T1ce, proximity maps	Automated	NA	T1, T1ce, T2, FLAIR, ADC	MFFE U-Net	MFFE U-Net	NA	0.85	NA	0.79	0.82	NA

Abbreviation: ADC: Apparent Diffusion Coefficient, AI: Artificial Intelligence, AUROC: Area Under Receiver Operating Characteristic, CE-T1WI: Contrast-Enhanced T1-Weighted Imaging, CNN: Convolutional Neural Network, DTI: Diffusion Tensor Imaging, DSC-MRI: Dynamic Susceptibility Contrast MRI, DT: Decision Tree, FLAIR: Fluid-Attenuated Inversion Recovery, GLCM: Gray-Level Co-Occurrence Matrix, GLDM: Gray-Level Dependence Matrix, HRR: High-Risk Recurrence, KNN: K-Nearest Neighbors, LOOCV: Leave-One-Out Cross Validation, MGMT: O6-Methylguanine-DNA Methyltransferase, MFFE U-Net: Multi-Feature Fusion and Enhancement U-Net, MRI: Magnetic Resonance Imaging, PoE: Predictive Outcome Estimation, RANO: Response Assessment in Neuro-Oncology, RF: Random Forest, Rad-score: Radiomics Score, SCN: Stem Cell Niches, SVM: Support Vector Machine, TERT: Telomerase Reverse Transcriptase, TMZ: Temozolomide, T1CE: T1-weighted Contrast-Enhanced, T2-FLAIR: T2 Fluid-Attenuated Inversion Recovery, WHO: World Health Organization.

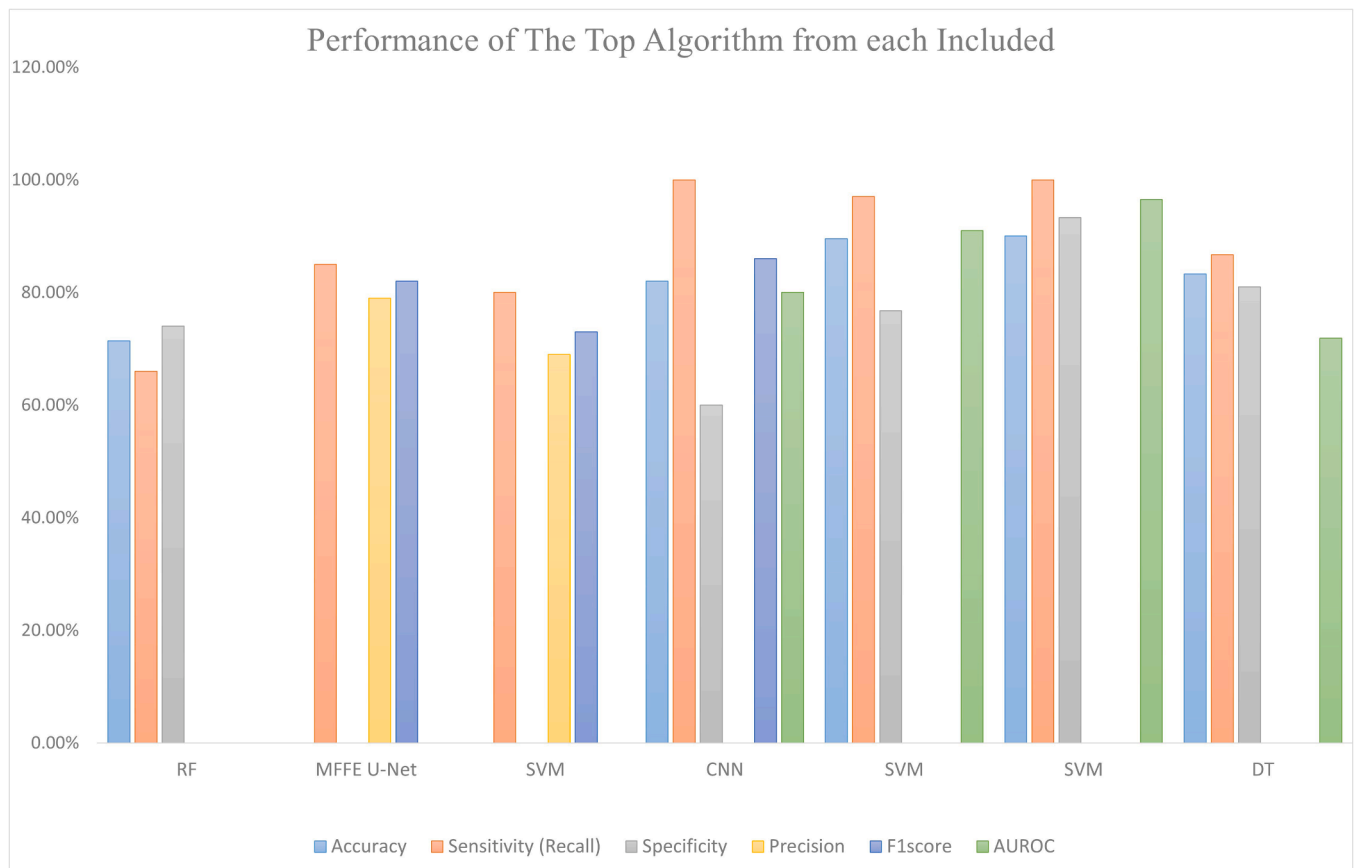


Fig. 3. Performance comparison of machine learning algorithms, including Accuracy, Sensitivity, Specificity and AUROC.

4.8.1. Quality assessment (PROBAST)

As described in the methods section, the majority of studies demonstrated low risk of bias across the four domains (participants, predictors, outcomes, and analysis) and application, with 91.53 %, 93.64 %, 96.90 %, 92.80 %, and 92.74 % respectively. A small percentage of studies showed high risk in these domains, particularly in the predictors and analysis categories, with 6.36 % and 7.20 % respectively. No studies were categorized as having an unclear risk of bias (Fig. 9).

4.8.2. Applicability of the ML tools (PROBAST)

Most studies showed low concern across the domains. Specifically, 98.84 % of studies demonstrated low concern in the participants domain, followed by 95.55 % in predictors, 97.57 % in outcomes, 95.43 % in analysis, and 95.78 % in overall applicability. A smaller percentage showed high concern, ranging from 1.16 % to 4.57 %. No domain was categorized as having unclear concern (Fig. 10).

5. Discussion

This systematic review and meta-analysis determined the performance of ML algorithms in detecting recurrent HGG, addressing the issue of accurately distinguishing between actual tumor progression and treatment-related changes (TRC) on MRI. Based on the results of our meta-analysis, a pooled sensitivity of 0.95 and a specificity of 0.80 indicated that ML algorithms excel at identifying true recurrences while maintaining acceptable accuracy in ruling out false positives. The positive likelihood ratio of 4.75 suggests that a positive test result substantially increases the probability of true recurrence. In contrast, the low negative likelihood ratio of 0.06 indicates that negative results effectively rule out recurrence. The DOR of 80.97 and a diagnostic score 4.39 further support the algorithms' strong discriminative ability. The area under the ROC curve of 0.86 represents good diagnostic accuracy in

clinical terms, demonstrating that ML algorithms show considerable promise as diagnostic tools for HGG recurrence detection.

The Association for Neuro-Oncology has recommended PET-CT imaging to be used for treatment response assessment in gliomas, as it has higher diagnostic accuracy than MRI in identifying actual tumor progression and TRC [8]. In the early 2010s, thallium-201 SPECT and dual-tracer PET imaging approaches achieved moderate diagnostic performance, with sensitivities and specificities ranging from 80 % to 83 % [19,20]. The field progressed by introducing more sophisticated tracers, particularly 11C-choline PET/CT, which improved sensitivity to 100 % but maintained specificity at 70 % [21]. Recent advances in FET-PET imaging have further refined these results, achieving a sensitivity of 91.6 % with a specificity of 76.9 % [22]. Our meta-analysis of ML approaches represents a further evolution in diagnostic capability based on conventional MRI or multimodality MRI achieving a balanced performance with high sensitivity (95 %) and maintaining good specificity (80 %). Compared with PET-CT, conventional MRI is more available in various hospitals and does not expose patients to ionizing radiation [8].

The evolution of MRI-based diagnosis for recurrent HGG has demonstrated significant advancement over the past decade. Early dynamic contrast-enhanced (DCE)-MRI approaches reported seemingly perfect diagnostic metrics with 100 % sensitivity and specificity, though from a limited patient cohort who had surgery and radiation therapy for glioma [23]. The field then progressed toward combining multiple MRI sequences with susceptibility-weighted MRI (SWMRI) and dynamic susceptibility contrast (DSC) perfusion-weighted imaging (PWI), showing a high specificity (100 %) and moderate sensitivity (71.9 %) in patients underwent radiation therapy or gamma knife surgery followed by resection and developed new measurable enhancement more than six months after complete response [24].

DL is a subset of ML approaches that may include artificial neural

networks (ANN), convolutional neural networks (CNN) and recurrent neural networks (RNN). Radiomics features of patients with a histopathologically-confirmed diagnosis of HGG over seven years were retrospectively analyzed by the 3D CNN model and validated by 5-fold cross-validation, achieving an AUC of 0.80 with the combination of DWI+FLAIR in distinguishing progression/recurrence from TRC [5]. Radiomics features indicate the high-performing quantitative features extracted from medical images, which cannot be recognized by the naked eye and may be related to genetic features [17]. Based on a study by P. Du et al., the predictive importance of radiomics(rad)-score was ranked second, inferior to age, and preceding MGMT promoter methylation, preoperative Karnofsky Performance Status (KPS), and TERT promoter mutation as important predictive factors for recurrence within 1 year after total resection in GBM patients [17]. The rad-score was obtained from T1 weighted imaging (T1WI), T2 weighted imaging (T2WI), T2-fluid attenuated inversion recovery (T2-Flair), diffusion weighted imaging (DWI) and contrast enhanced (CE)-T1WI. Using a DT model combining the above predictive factors, the model achieved an AUC: 0.850 in the training set and 0.719 in the test set [17]. Glioma recurrence involves malignant cells invading healthy brain tissue via the vascular network or myelinated white matter (WM) fibers, though the factors influencing their pathway choice remain unclear. Predicting recurrence location is challenging, but most cases arise in the pre-existing peritumoral edema [15]. Another radiomics-based model was established from the postoperative enhancement and edema regions from four routine MRI sequences including, T1WI, CE-T1WI, T2-FLAIR,

and T2WI. The results showed that multimodality based on the whole region best distinguished recurrence from TRC, compared to each modality alone, with AUC of 0.965 for SVM and 0.955 for KNN [8]. Most of the features in this study were derived from CE-T1WI and T2WI. CE-T1WI, due to information enhancement after employing contrast agents, allowed for assessing of blood-brain barrier impairment. T2WI depicted the cellular proliferation state of neighboring tissues through free water reaction [8]. However, multimodal radiomics and diffusion multicompartments models usually only focus on the pre-existing peritumoral edema as the area of probable recurrence, while relapse in distant/multifocal locations is not considered [15]. Based on a spatio-temporal radiomics-based trajectory for GBM, a longitudinal radiomics analysis based on three to 13 multimodality MRI time points was performed using the multimodal voxel wise radiomic features from the recurrence areas on FLAIR, diffusion weighted imaging (DWI)-derived ADC maps, and T1CE compared with normative WM for each pre-recurrence time point and the temporally discriminative features [15].

Multiple statistical tools including regression analysis, Deeks' test, and PROBAST quality assessment revealed no significant reporting bias and consistently high methodological standards across studies. The minimal identified concerns in predictor selection and analysis are typical for medical imaging ML studies. These findings validate that the superior performance of ML approaches in glioma recurrence detection reflects genuine capabilities rather than methodological bias or poor study quality.

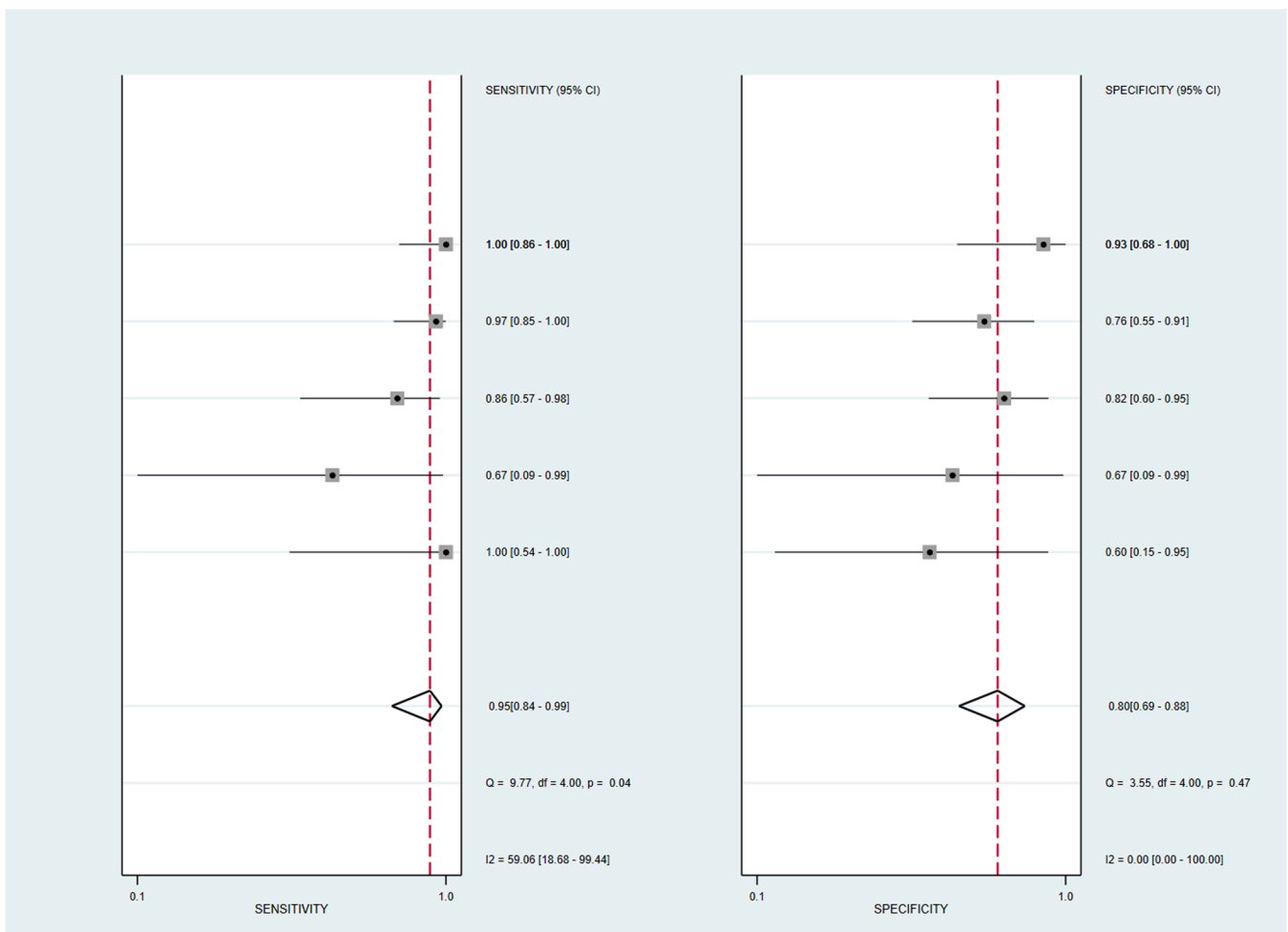


Fig. 4. Sensitivity and specificity of ML algorithms.

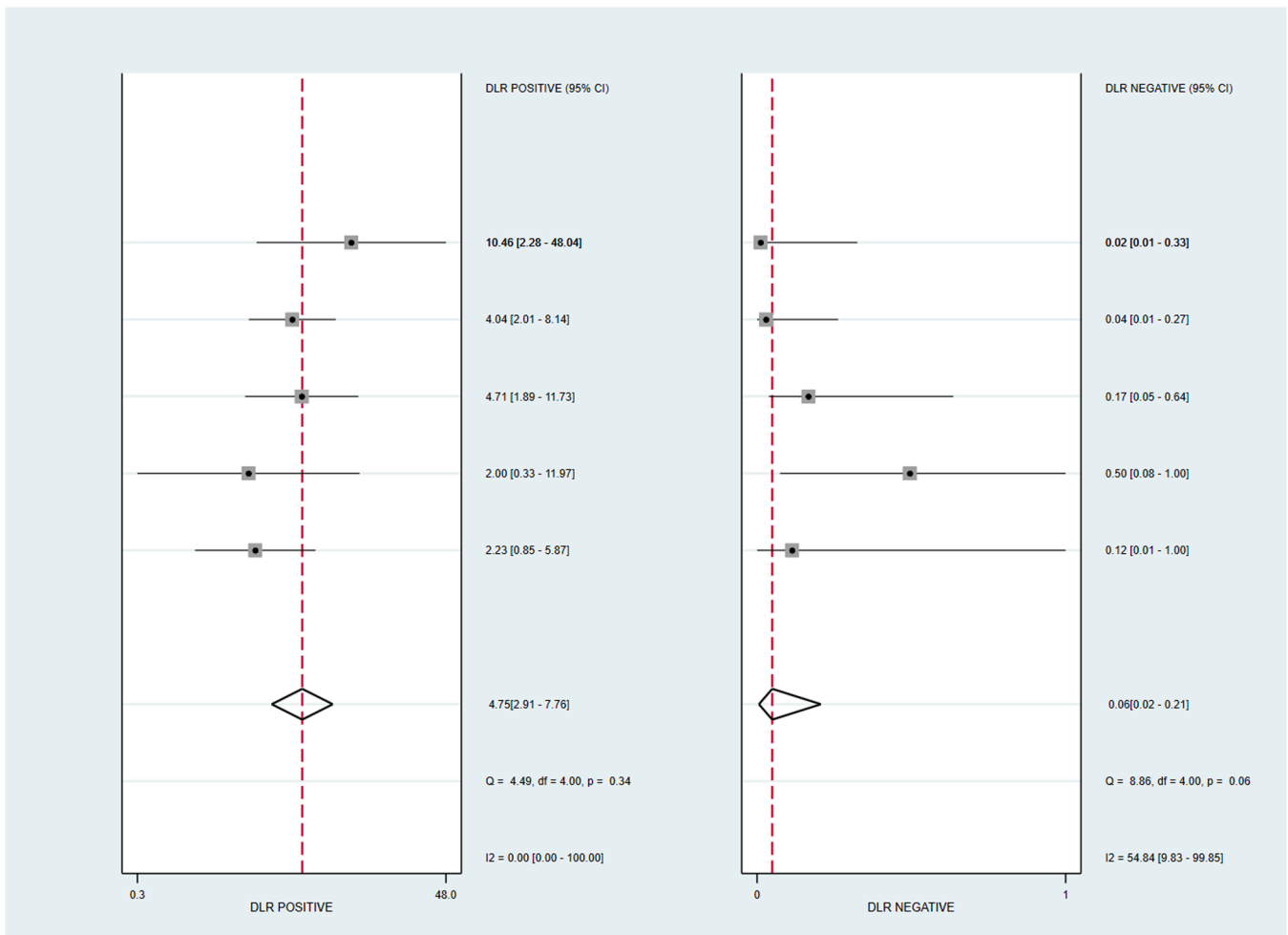


Fig. 5. Positive and negative DLR of ML algorithms.

this study demonstrates the potential of AI-based models, particularly SVM, in accurately predicting recurrence in HGG. With high sensitivity and specificity, these models show major diagnostic challenges: distinguishing true tumor recurrence from treatment-related changes. The findings indicate that machine learning tools must be integrated into clinical workflows to improve diagnostic precision and decision-making.

5.1. Limitations

Our study has several limitations. First, the retrospective nature of the included studies and our meta-analysis may introduce selection bias and limit statistical power. The relatively small sample sizes and geographical concentration, primarily in the United States and China, restrict generalizability. Second, considerable heterogeneity was observed across studies in terms of sensitivity and diagnostic odds ratios, driven by variations in study design, methodology, and MRI sequence selection. Differences in feature extraction methods and the use of diverse machine learning algorithms, ranging from SVM to CNN, further complicate direct performance comparisons and standardization of metrics. Lastly, the limited number of studies meeting inclusion criteria

may affect the completeness of our findings. Future research should focus on large-scale, multicenter prospective studies with standardized methodologies, consistent feature selection, and external validation datasets to improve the reliability and clinical applicability of ML models for glioma recurrence detection.

6. Conclusion

In conclusion, this systematic review and meta-analysis demonstrates the promising performance of ML models, particularly SVM, in detecting glioma recurrence. The pooled diagnostic metrics suggest strong potential for clinical application, though prospective multicenter validation with standardized protocols remains essential. As research in this field continues to evolve, close collaboration among clinicians, radiologists, and data scientists is crucial to develop and validate effective, clinically applicable models that can improve survival outcomes.

Ethics declaration

The study is deemed exempt from receiving ethical approval.

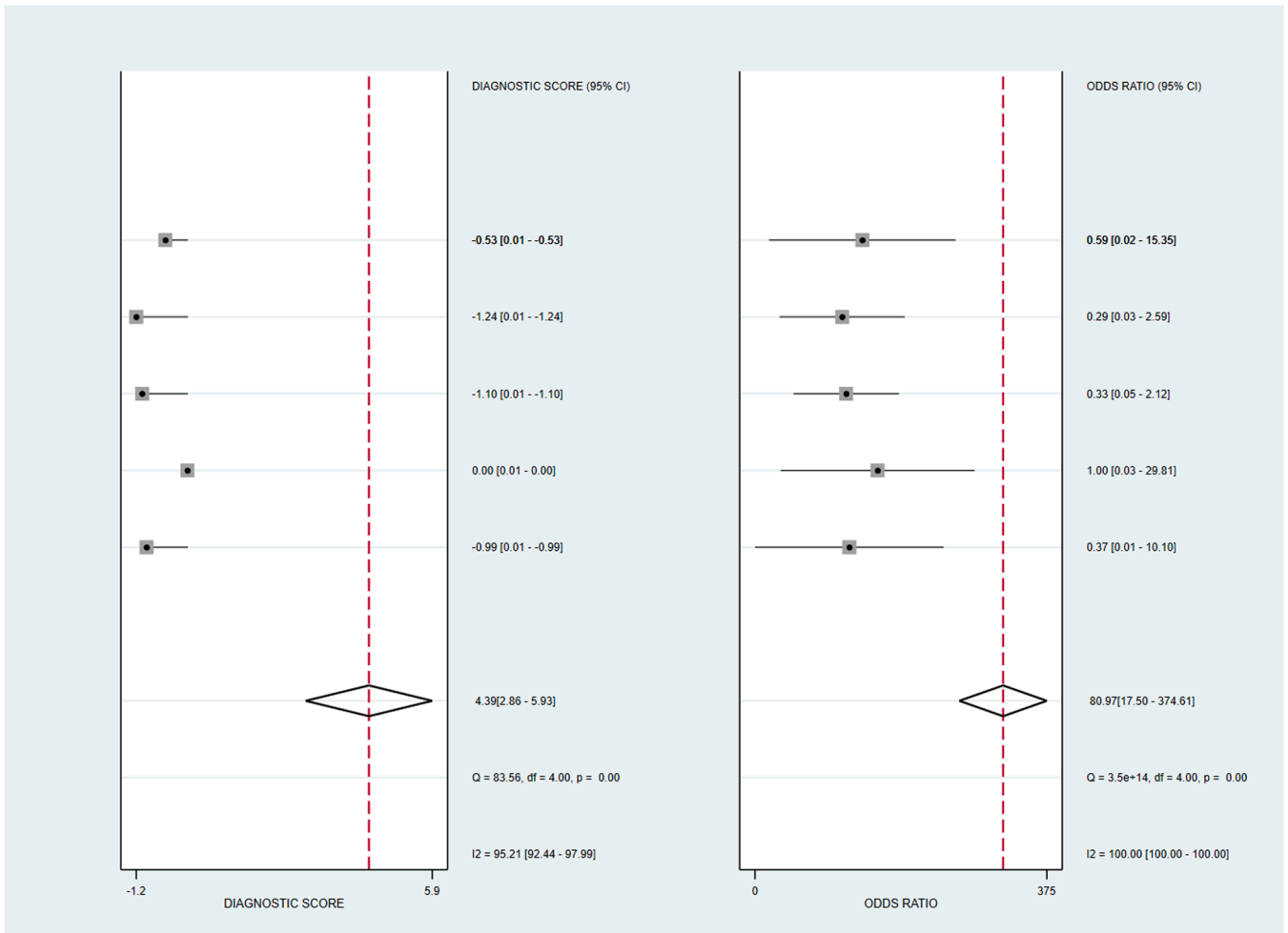


Fig. 6. Diagnostic score and diagnostic odds ratio of ML algorithms.

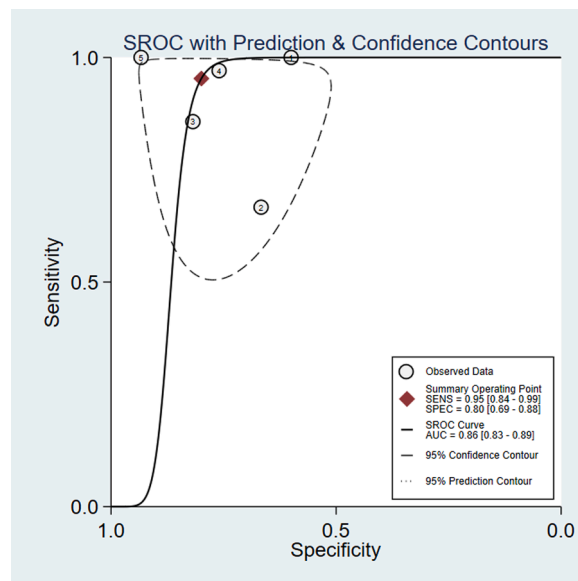


Fig. 7. Summary receiver operator characteristic curve (SROC) of ML algorithms.

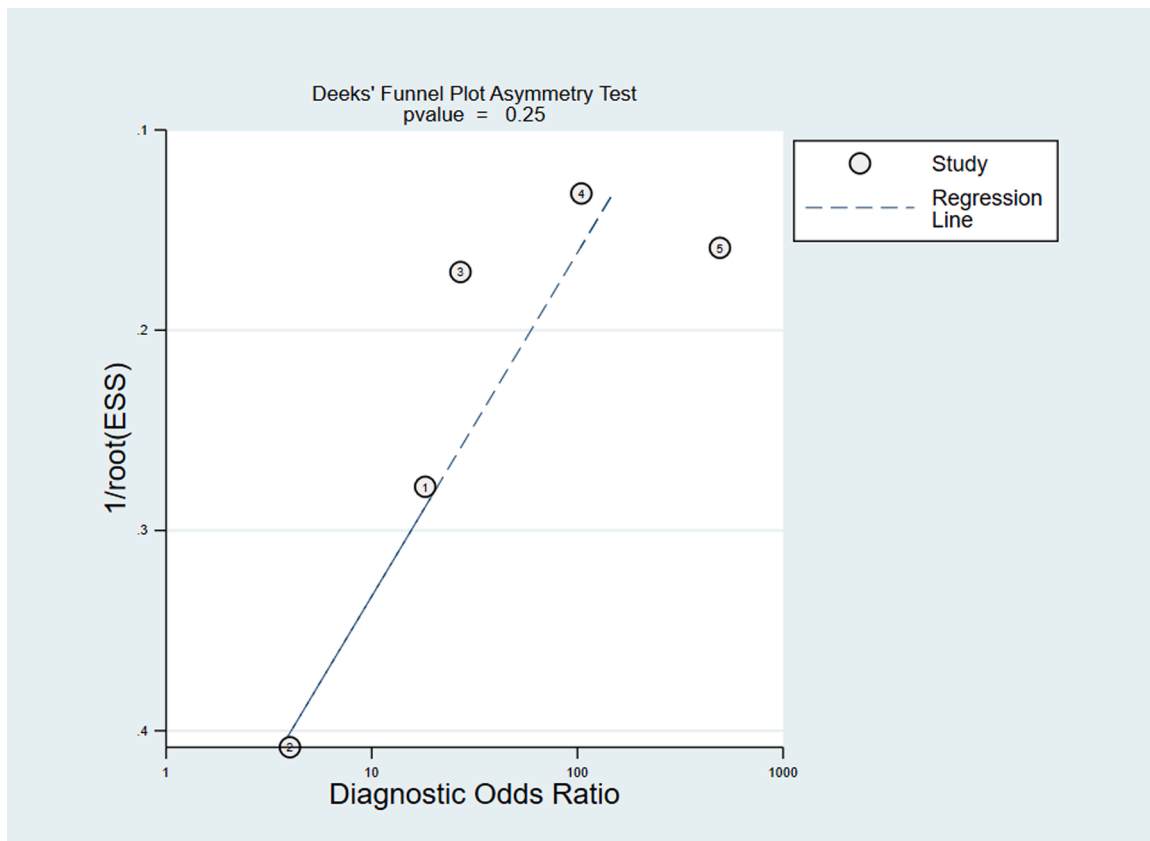


Fig. 8. PROBAST chart showing the distribution of risk of bias across the domains of Participants, Predictors, Outcome, and Analysis.

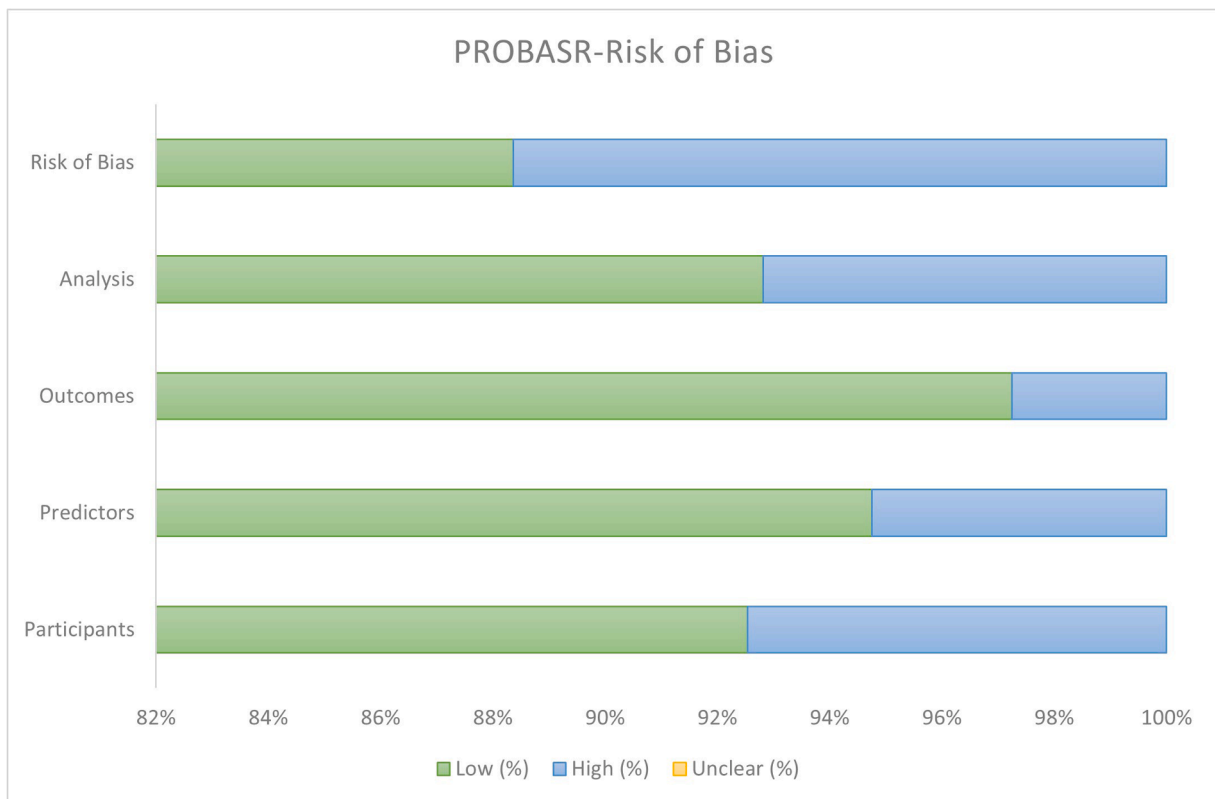


Fig. 9. Deeks' Funnel Plot asymmetry test for publication bias, displaying the diagnostic odds ratio against 1/root (ESS). The plot includes individual studies (circles) and a regression line (dashed).

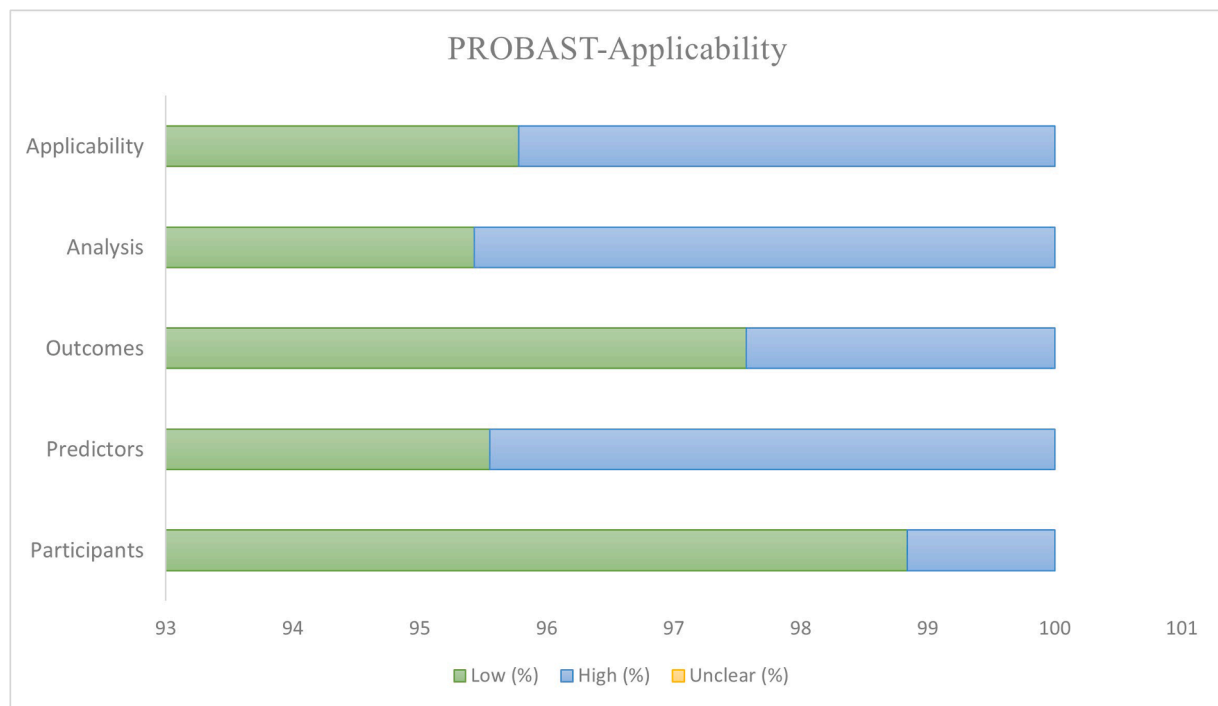


Fig. 10. PROBAST chart illustrating the applicability assessment across the domains of Participants, Predictors, and Outcome.

Funding declaration

The authors did not receive support from any organization for the submitted work.

Authors' contribution

The conception and design of the study: Ibrahim Mohammadzadeh (IM) and Hamid Borghei-Razavi (HBR). Acquisition of data: Behnaz Niroomand (BN) and Bardia Hajikarimloo. Analysis and interpretation of data: IM, BN. Drafting the article: IM, BN and HBR. Revising it critically for important intellectual content: Mohammad Amin Habibi, Jina Behjati, Abdulrahman Albakr and Ali Mortezaei. Final approval of the version to be submitted: All authors.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

CRedit authorship contribution statement

Borghei-Razavi Hamid: Project administration, Writing – original draft, Writing – review & editing. **Mortezaei Ali:** Writing – review & editing. **Behjati Jina:** Writing – review & editing. **Albakr Abdulrahman:** Methodology, Writing – review & editing. **Mohammadzadeh Ibrahim:** Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Niroomand Behnaz:** Formal analysis, Writing – original draft, Writing – review & editing. **Hajikarimloo Bardia:** Methodology, Writing – review & editing. **Habibi Mohammad Amin:** Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

Nothing to acknowledge.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.clineuro.2025.108762](https://doi.org/10.1016/j.clineuro.2025.108762).

References

- [1] M. Alizadeh, et al., Radiomics: the new promise for differentiating progression, recurrence, pseudoprogression, and radionecrosis in glioma and glioblastoma multiforme, *Cancers* 15 (18) (2023) 4429.
- [2] R. Wei, et al., Application of intraoperative ultrasound in the resection of high-grade gliomas, *Front. Neurol.* 14 (2023).
- [3] K. Shim, et al., Radiomics-based neural network predicts recurrence patterns in glioblastoma using dynamic susceptibility contrast-enhanced MRI, *Sci. Rep.* 11 (2021) 9974.
- [4] S. Rathore, et al., A radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma. *Medical Imaging 2018 Image-Guided Procedures, Robotic Interventions, and Modeling*, SPIE, 2018.
- [5] S. Bacchi, et al., Deep learning in the detection of high-grade glioma recurrence using multiple MRI sequences: a pilot study, *J. Clin. Neurosci.* 70 (2019) 11–13.
- [6] P. Esmailzadeh, Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations, *Artif. Intell. Med.* 151 (2024) 102861.
- [7] I. Mohammadzadeh, et al., Leveraging machine learning algorithms to forecast delayed cerebral ischemia following subarachnoid hemorrhage: a systematic review and meta-analysis of 5,115 participants, *Neurosurg. Rev.* 48 (1) (2025) 26.
- [8] J. Ren, et al., Multimodality MRI radiomics based on machine learning for identifying true tumor recurrence and treatment-related effects in patients with postoperative glioma, *Neurol. Ther.* 12 (5) (2023) 1729–1743.
- [9] L. Pei, et al., Deep neural network analysis of pathology images with integrated molecular data for enhanced glioma classification and grading, *Front. Oncol.* 11 (2021) 668694.
- [10] L. Jin, et al., Artificial intelligence neuropathologist for glioma classification using deep learning on hematoxylin and eosin stained slide images and molecular markers, *Neuro-Oncology* 23 (1) (2021) 44–52.

- [11] X. Hu, et al., Support vector machine multiparametric MRI identification of pseudoprogression from tumor recurrence in patients with resected glioblastoma, *J. Magn. Reson. Imaging* 33 (2) (2011) 296–305.
- [12] M.J. Page, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *bmj* (2021) 372.
- [13] R.F. Wolff, et al., PROBAST: a tool to assess the risk of bias and applicability of prediction model studies, *Ann. Intern. Med.* 170 (1) (2019) 51–58.
- [14] S. Rathore, H. Akbari, J. Doshi, et al., Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning, *J. Med. Imaging* 5 (2) (2018), 021219-021219.
- [15] T. Chougule, et al., Radiomics signature for temporal evolution and recurrence patterns of glioblastoma using multimodal magnetic resonance imaging, *NMR Biomed.* 35 (3) (2022) e4647.
- [16] Y. Lao, D. Ruan, A. Vassantachart, et al., Voxelwise prediction of recurrent high-grade glioma via proximity estimation-coupled multidimensional support vector machine, *Int. J. Radiat. Oncol. *Biol. *Phys.* 112 (5) (2022) 1279–1287.
- [17] P. Du, et al., The application of decision tree model based on clinicopathological risk factors and pre-operative MRI radiomics for predicting short-term recurrence of glioblastoma after total resection: a retrospective cohort study, *Am. J. Cancer Res.* 13 (8) (2023) 3449–3462.
- [18] C. Jiao, Y. Lao, W. Zhang, et al., Multi-modal fusion and feature enhancement U-Net coupling with stem cell niches proximity estimation for voxel-wise GBM recurrence prediction, *Phys. Med. Biol.* 69 (15) (2024) 155021.
- [19] S. Matsunaga, et al., Semiquantitative analysis using thallium-201 SPECT for differential diagnosis between tumor recurrence and radiation necrosis after gamma knife surgery for malignant brain tumors, *Int. J. Radiat. Oncol. *Biol. *Phys.* 85 (1) (2013) 47–52.
- [20] H. Shishido, et al., Diagnostic value of 11C-methionine (MET) and 18F-fluorothymidine (FLT) positron emission tomography in recurrent high-grade gliomas; differentiation from treatment-induced tissue necrosis, *Cancers* 4 (1) (2012) 244–256.
- [21] W. Li, et al., 11C-choline PET/CT tumor recurrence detection and survival prediction in post-treatment patients with high-grade gliomas, *Tumor Biol.* 35 (2014) 12353–12360.
- [22] A.D. Puranik, et al., FET PET to differentiate between post-treatment changes and recurrence in high-grade gliomas: a single center multidisciplinary clinic controlled study, *Neuroradiology* (2024) 1–7.
- [23] V.A. Larsen, et al., Evaluation of dynamic contrast-enhanced T1-weighted perfusion MRI in the differentiation of tumor recurrence from radiation necrosis, *Neuroradiology* 55 (2013) 361–369.
- [24] T.-H. Kim, et al., Combined use of susceptibility weighted magnetic resonance imaging sequences and dynamic susceptibility contrast perfusion weighted imaging to improve the accuracy of the differential diagnosis of recurrence and radionecrosis in high-grade glioma patients, *Oncotarget* 8 (12) (2016) 20340.