Neuro-Oncology

23(2), 214–225, 2021 | doi:10.1093/neuonc/noaa232 | Advance Access date 19 October 2020

Brain metastasis detection using machine learning: a systematic review and meta-analysis

Se Jin Cho, Leonard Sunwoo, Sung Hyun Baik, Yun Jung Bae, Byung Se Choi, and Jae Hyoung Kim

Department of Radiology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, Gyeonggi, Republic of Korea (S.J.C., L.S., S.H.B., Y.J.B., B.S.C., J.H.K.)

Corresponding Author: Leonard Sunwoo, MD, PhD, Department of Radiology, Seoul National University Bundang Hospital, 82, Gumi-ro 173beon-gil, Bundang-gu, Seongnam, Gyeonggi, 13620, Republic of Korea (leonard.sunwoo@gmail.com).

Abstract

Background. Accurate detection of brain metastasis (BM) is important for cancer patients. We aimed to systematically review the performance and quality of machine-learning-based BM detection on MRI in the relevant literature.

Methods. A systematic literature search was performed for relevant studies reported before April 27, 2020. We assessed the quality of the studies using modified tailored questionnaires of the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) criteria and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM). Pooled detectability was calculated using an inverse-variance weighting model.

Results. A total of 12 studies were included, which showed a clear transition from classical machine learning (cML) to deep learning (DL) after 2018. The studies on DL used a larger sample size than those on cML. The cML and DL groups also differed in the composition of the dataset, and technical details such as data augmentation. The pooled proportions of detectability of BM were 88.7% (95% CI, 84–93%) and 90.1% (95% CI, 84–95%) in the cML and DL groups, respectively. The false-positive rate per person was lower in the DL group than the cML group (10 vs 135, $P < 0.001$). In the patient selection domain of QUADAS-2, three studies (25%) were designated as high risk due to non-consecutive enrollment and arbitrary exclusion of nodules.

Conclusion. A comparable detectability of BM with a low false-positive rate per person was found in the DL group compared with the cML group. Improvements are required in terms of quality and study design.

Key Points

1. Larger datasets are used for training algorithms in the DL group than the cML group.
2. The DL group showed a lower false-positive rate per person than the cML group.
3. The quality and study design of the published literature should be improved.

Brain metastases (BM) are the most common malignant brain tumors in adults. Detection of BM is important because of its high incidence (about 20% of patients with systemic cancer), its contribution to mortality in patients with advanced-stage cancer, and the comparable local control rate of stereotactic radiosurgery to radiotherapy for limited BM.[1–3] Recent technological advances in MRI have led to more accurate BM detection.[4–6] However, human readers also confront several challenges such as excessive workload (due to the increased burden of the initial screening and follow-up MRIs), fatigue and fluctuations in concentration, mimickers of BM,[5] and risk of medico-legal problems.[7]

A computer-aided detection (CAD) system could potentially solve these problems. CAD can help radiologists enhance their reading efficacy by increasing vigilance.[8] On account of recent progress in artificial intelligence technology, the volume of research in CAD for BM on MRI has greatly increased, particularly with the advent of deep learning (DL).[9–20] The studies consistently report that CAD can automatically detect varying size of enhancing BM nodules on MRI using

Importance of the Study

Although there is no fully established quality assessment tool for reporting machine learning research, many guidelines such as the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) are currently being revised. At the point where a clear transition from cML to DL occurred for the automated detection of BM, this systematic review and meta-analysis of the algorithm performance and quality of published machine learning research highlights the technical details of cML versus DL and offers valuable information to guide studies in the future.

different machine learning algorithms, along with some false-positive lesions. To compare the results of these studies and to choose the optimal CAD algorithm for BM detection, a comparative study is needed. However, to date, the performance and technical details of CAD for BM on MRI in the literature have never been systematically reviewed.

On the other hand, classical machine learning (cML) and DL algorithms are both inherently prone to overfitting and spectrum bias. Thus, a robust study design is required to avoid such biases and to enhance the clinical impact and generalizability. However, currently there are no established quality assessment criteria specific to the systematic review of machine learning studies.[21–23] In addition, it would be meaningful to compare technical details of both groups, such as the sample size of the dataset, use of external validation, or use of BM nodule size criteria. Therefore, this systematic review and meta-analysis aimed to evaluate the impact of DL over cML on the performance of CAD for BM detection, to assess the quality and methodological appropriateness of included studies, and to provide guidance for future research.

Materials and Methods

This systematic review and meta-analysis was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[24]

Literature Search

A search of MEDLINE and EMBASE databases was performed to find original literature that reported the detectability of machine learning using MRI data for patients with BM. The following search terms were used: ((brain metastasis) OR (brain metastases) OR (metastatic brain tumor) OR (intra-axial metastatic tumor) OR (cerebral metastasis) OR (cerebral metastases)) AND ((automated) OR (computer aided) OR (computer-aided) OR (CAD) OR (radiomic) OR (texture analysis) OR (deep learning) OR (machine learning) OR (neural network) OR (artificial intelligence)). No beginning search date was set, with the literature search being updated until April 27, 2020. The search was limited to publications written in the English language. The bibliographies of relevant articles were searched to identify any other appropriate articles.

Inclusion Criteria

Studies satisfying the following criteria were included: (1) involved patients with BM; (2) machine learning using MRI data was the index test; and (3) contained sufficient data for the detectability (proportion) analysis of the index test.

Exclusion Criteria

Studies or subsets of studies were excluded if any of the following criteria were met: (1) case reports or case series including fewer than 10 patients regardless of the topic; (2) letters, editorials, conference abstracts, systematic reviews or meta-analyses, consensus statements, guidelines, and review articles; (3) articles not focusing on the current topic; (4) articles with, or with suspicion of, overlapping populations; and (5) contained insufficient data for the detectability analysis of machine learning using MRI data for the patients with BM.

Two radiologists, S.J.C. and L.S., with 7 and 10 years, respectively, of experience in neuroimaging, independently performed the literature search and selection.

Data Extraction

We extracted data using standardized forms according to the PRISMA guidelines.[24] Herein, the DL group was defined as the studies that utilized deep neural networks (eg, convolutional neural networks or its derivatives) as their main algorithm. Otherwise, the studies were classified into the cML group.[21] The following data were extracted:

1. Characteristics of the included studies: authors, year of publication, institution, country of origin, duration of data recruitment, study design (prospective vs retrospective), category of validation (internal vs external, random split vs temporal split if internal validation), number of patients in each dataset (total, developmental, and test set, respectively. Due to the unclear word across the studies, we defined the developmental set as all datasets except for test set (ie, the validation set for the fine-tuning step in the DL is considered as the developmental set[23]), male-to-female ratio (total, developmental, and test set, respectively), and proportion of lung cancer (which is the most common primary cancer for BM) among the enrolled patients with cancer (total, developmental, and test set, respectively), number of metastatic nodules in each dataset

(total, developmental, and test set, respectively), mean size of metastatic nodules in each dataset (total, developmental, and test set, respectively), proportion of nodules equal to or larger than 10 mm (total, developmental, and test set, respectively), patient inclusion criteria, and performer who determined the ground truth.

2. MRI characteristics: MRI machine and vendor, magnetic field strength (T), in-plane resolution, and slice thickness (mm).
3. Machine learning characteristics: specific type or name of algorithm of machine learning, presence of skull stripping, signal intensity normalization, segmentation, data augmentation.
4. Detectability and false-positive rate per person (the number of false-positive lesions per patient) of machine learning using MRI data for the patients with BM.

Quality Assessment

Since there is no established quality assessment tool that focuses on machine learning methodology, we selected several items from Checklist for Artificial Intelligence in Medical Imaging (CLAIM), a recently published guideline,[25] and applied them to tailored questionnaires of the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) criteria.[26] Specifically, the following CLAIM items were synthetically considered when assessing each domain of the QUADAS-2: (1) in the risk of bias of patient selection domain: data sources, selection of data subsets, and how missing data were handled; (2) in the risk of bias of index test domain: statistical measures of significance and uncertainty, and robustness or sensitivity analysis; (3) in the risk of bias of reference standard domain of the QUADAS-2: sufficient detail to allow replication about definition of ground truth, rationale for choosing the reference standard, qualifications and preparation of annotators for source of ground truth annotations; and (4) in concerns regarding the applicability of index test domain: validation or testing on external data.

Two reviewers (S.J.C. and L.S.) independently performed the data extraction and quality assessment. Disagreements between the 2 reviewers were discussed at a research meeting until a consensus was reached.

Data Synthesis and Analyses

The current work aimed to systematically review the relevant topic, including a detailed quality assessment, and to perform a pooled proportion analysis of detectability of machine learning using MRI data for patients with BM. The pooled proportions were calculated using an inverse-variance weighting model.[27–29] A random-effects meta-analysis of proportions was utilized to calculate the overall proportions. The study heterogeneity was evaluated using Higgins inconsistency index (I^2), with substantial heterogeneity being indicated by an I^2 value greater than 50%.[30] The comparative statistical significance of the false-positive rate per person between 2 groups was obtained using multilevel mixed-effects Poisson regression. All statistical analyses were conducted by one author (S.J.C., with 3 years of experience in conducting systematic reviews and meta-analysis), using the “meta” package in R, version 3.6.3 (http://www.r-project.org/).

Results

Literature Search

A systematic literature search (Figure 1) initially identified 1044 articles from OVID-MEDLINE and EMBASE. After removing 202 duplicates, the screening of the remaining 842 titles and abstracts yielded 21 potentially eligible articles. No additional article was identified after searching the bibliographies of these articles. After full-text reviews of the 21 provisionally eligible articles, nine articles were excluded because they were not in the field of interest,[31–37] contained potentially overlapping data,[38] or contained insufficient information in terms of detectability of machine learning using MRI data for patients with BM.[39] Finally, 12 articles were included in the present systematic review and meta-analysis.[9–20]

Characteristics of the Included Studies

Table 1 shows the detailed characteristics of the 12 studies, including 5 on cML[9,13,15,16,18] and 7 on DL.[10–12,14,17,19,20] One study was a multicenter study,[17] but the others were all single-center studies.[9–16,18–20] One study was a prospective design,[13] while the others were all retrospective.[9–12,14–20] One study performed external validation,[17] and another study performed internal validation with a temporal split.[16] All other studies performed internal validation with a random split.[9–15,18–20] The number of enrolled patients across all studies was 3620, with individual studies ranging from 26 to 1632 patients. The number of metastatic nodules across all studies was 10 258, with studies ranging from 62 to 3264. There was a clear transition from cML to DL after 2018 (Figure 2). The number of enrolled patients and the metastatic nodules were relatively smaller in the cML group, ranging from 26 to 140 and 62 to 584, respectively,[9,13,15,16,18] than the DL group, which ranged from 158 to 1632 and 932 to 3264, respectively.[10–12,14,17,19,20] The ratio of patient numbers in the developmental set to that in the test set in 4 of the included studies was higher than 4 (all of which were in the DL group).[10–12,19] The ratio was equal to or lower than 1 in 4[9,13,15,18] (all of which were in the cML group) of the remaining 8 studies.[9,13–18,20] In terms of the number of metastatic nodules, the ratio of the developmental/test set was higher than 4 in 3 of the included studies,[10,16,19] equal to or lower than 4 in 6,[9,13,15,17,18,20] and available in 3 studies.[11,12,14] The proportion of lung cancer among patients with underlying cancer and the proportion of the nodules equal to or larger than 10 mm are presented in Table 1. Overall, studies included highly probable metastatic nodules. In detail, 3 of the enrolled studies used a size threshold as the inclusion criteria.[12,13,18] The ground truths were determined by radiologists and/or neuro-oncologists across the studies. Among them, 4 studies took the follow-up MRI into consideration to determine the ground truth.[9,15,16,19]

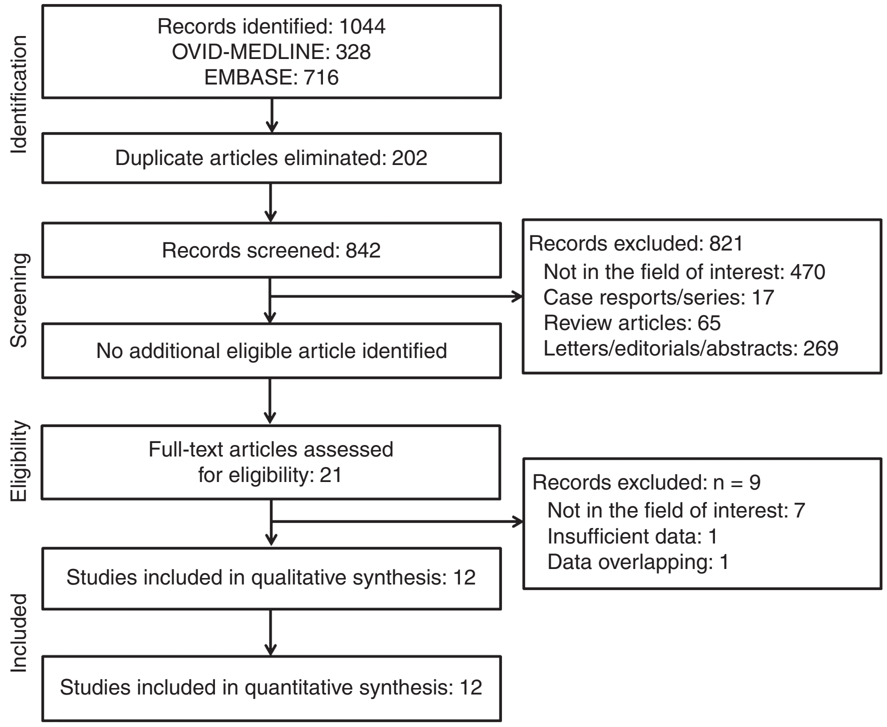
Identification
Records identified: 1044
OVID-MEDLINE: 328
EMBASE: 716
Duplicate articles eliminated: 202
Screening
Records screened: 842
Records excluded: 821
Not in the field of interest: 470
Case resports/series: 17
Review articles: 65
Letters/editorials/abstracts: 269
No additional eligible article identified
Eligibility
Full-text articles assessed for eligibility: 21
Records excluded: n = 9
Not in the field of interest: 7
Insufficient data: 1
Data overlapping: 1
Included
Studies included in qualitative synthesis: 12
Studies included in quantitative synthesis: 12

Fig. 1 Flow diagram of the study selection process.

MRI and Machine Learning Characteristics of the Included Studies

MRI examinations were performed using 1.5T scanners in 3 studies,[9,11,15] 3T scanners in 3 studies,[13,17,18] and either 1.5T or 3T scanners in the other 6 studies (Table 2).[10,12,14,15,19,20] One study acquired the images with a slice thickness of 2.5 mm,[9] another study with 1 to 2 mm,[17] and the others with equal to or lower than 1 mm,[11–16,18–20] except for one study in which the slice thickness data was not available. All studies in the cML group used a template-matching algorithm,[9,13,15,16,18] including one study that additionally used a cross-correlation method,[13] and another, which additionally used a K-means clustering method.[16] In the DL group, 5 of the 7 studies used a 3D convolutional neural network,[10–12,17,19] one study used a 2.5D convolutional neural network,[14] and the other study used a single-shot detector.[20] In terms of the detailed techniques, 6 of the studies performed skull stripping (3 in cML and 3 in DL). All 12 studies performed intensity normalization, 7 of the studies performed tumor segmentation (2 in cML and 5 in DL), 7 of the studies performed data augmentation (none in cML and all 7 in DL), and 3 of the studies thresholded the enhancement degree (1 in cML and 2 in DL).

Quality Assessment of the Studies via QUADAS-2

The quality of the included studies was assessed according to the QUADAS-2 criteria,[26] under the consideration of the CLAIM guideline.[25] The results are presented with a diagram in Supplementary Figure 1, and the detailed assessment is presented in Table 3 according to the domain of the risk of bias and concern of applicability, respectively. In the patient selection domain of risk of bias, two enrolled studies were considered to have a low risk of bias,[14,16] 3 studies were considered to have a high risk of bias due to non-consecutive patient selection and excluding nodules by size criteria,[12,13,18] whereas the other studies with non-consecutive patient selection were considered to have an unclear high risk of bias.[9–11,15,17,19,20] In the index test domain, all studies were considered to have a low bias risk because the ground truth was blinded to the machine, and a prespecified threshold (determined in the algorithm development phase) was used in the test phase. In the reference standard domain, 5 studies were considered to have a low risk of bias since they defined the ground truth of metastasis under the consideration of the follow-up MRI,[9,15–17,19] whereas the others were considered unclear.[10–14,20] In the flow and timing domain, two studies were considered to have an unclear risk of bias because some patients in these studies were excluded during the size thresholding of the BM nodules,[13,18] while the others were considered to have a low risk of bias.[9–12,14–17,19,20] In the index test domain of concern of applicability, 2 studies which performed internal validation with a temporal split or external validation were considered to have a low concern of applicability, whereas the others that used internal validation with a random split were considered to have an unclear concern of applicability.[9–15,18–20] Besides, all studies were considered to have low applicability in the patient selection and reference standard domains.

Table 1 Characteristics of the included studies

Group	Source	Affiliation	Recruitment Duration of Data	Study design	Validation	Total Pt. No. (Dev/ Test)	Total Male/ Female Ratio (Dev/Test)	Total Proportion of Lung cancer, % (Dev/Test)	Total Nodule No. (Dev/Test)	Total Mean Nodule Size (Dev/Test)	Total Proportion of Nodule Size ≥ 10 mm, % (Dev/Test)	Enroll Criteria	Ground Truth
cML	Ambrosini RD et al 2010[9]	University of Rochester, USA	NA	Retro.	NA	31 (9/22)	NA (NA/ NA)	NA (NA/ NA)	203 (124/79)	7mm (7mm/7mm)	NA (NA/ NA)	NA	By radiologist, with FU MRI
cML	Farjam R et al 2012[13]	University of Michigan, USA	NA	Pros.	Internal random split (5-fold CV)	29 (9/20)	NA (NA/ NA)	NA (NA/ NA)	292 (106/186)	NA (NA/ NA)	NA (NA/ 0)	Nodule size > 5mm	By a radiologist
cML	Pérez-Ramírez Ú et al 2016[15]	Universitat Politècnica de València', Spain	January 2014-February 2014	Retro.	Internal random split	30 (8/22)	NA (NA/ NA)	47.4 (50/ 45)	62 (20/42)	7.7mm (7.6mm/7.8mm)	40.3 (35/ 42.9)	No history of radiosurgery	By a radiologist and a radiation neuro-oncologist, with FU MRI
cML	Sunwoo L et al 2017[16]	Seoul National University Bundang Hospital, Republic of Korea	January 2015-March 2016	Retro.	Internal temporal split	140 (80/60)	1.1 (1.1/1)	80 (77.5/ 83.3)	584 (450/134)	4.8 mm (5mm/4.5mm)	22.9 (23.6/20.9)	No extra-axial metastasis or other coexistent diseases, sufficient MRI quality, nodule number < 50	By two radiologists, with FU MRI
cML	Yang S et al 2013[18]	Severance Hospital, Republic of Korea	NA	Retro.	Internal random split	26 (7/19)	NA (NA/ NA)	NA (NA/ NA)	86 (33/53)	NA (NA/ NA)	0 (0/0)	Nodule size ≥ 1mm, <10mm	By a radiologist
DL	Bousabarah K et al 2020[10]	Faculty of Medicine and University Hospital Cologne, Germany	April 2012-June 2019	Retro.	Internal random split	509 (469/40)	1.1 (1.1/1.9)	54 (54.4/ 50)	1223 (1149/83)	1.34 cm^3 (1.3cm^3/1.92 cm^3	NA (NA/ NA)	Candidate for stereotactic surgery, Sufficient MRI sequence	By radiation oncologists or neurosurgeons
DL	Charron O et al 2018[11]	Paul Strauss Center, France	2010–2016	Retro.	Internal random split	182 (164/18)	NA (NA/ NA)	NA (NA/ NA)	412 (NA/ NA†)	8.7 mm (NA/ NA)	33.7 (NA/ NA)	No history of surgery or radiation, sufficient MRI sequence	By a radiation oncologist

Table 1 Continued

Group	Source	Affiliation	Recruitment Duration of Data	Study design	Validation	Total Pt. No. (Dev/ Test)	Total Male/ Female Ratio (Dev/Test)	Total Proportion of Lung cancer, % (Dev/Test)	Total Nodule No. (Dev/Test)	Total Mean Nodule Size (Dev/Test)	Total Proportion of Nodule Size ≥ 10 mm, % (Dev/Test)	Enroll Criteria	Ground Truth
DL	Dikici E et al 2020[12]	Ohio State University College of Medicine, USA	NA	Retro.	Internal random split	158 (126/32)	NA (NA/ NA)	NA (NA/ NA)	932 (NA/ NA†)	5.5 mm (NA/ NA)	8.7 (NA/ NA)	Nodule size ≤ 15mm	By a radiologist
DL	Grøvik E et al 2020[14]	Stanford University, USA	June 2016-June 2018	Retro.	Internal random split	156 (105/51)	0.5 (NA/ NA)	NA (NA/ NA)	NA (NA/856)	NA (NA/ NA)	NA (NA/ NA)	No history of surgery or radiation, sufficient MRI sequence	By two radiologists
DL	Xue J et al 2019[17]	Multicenter *, China	October 2016-May 2019,	Retro.	Internal random split (4-fold CV) & External	1632 (1201/431)	1.3 (1.3/1.1)	NA (NA/ NA)	3264 (2402/862)	NA (4.01 cm^3/ NA)	NA (NA/ NA)	No extra-axial metastasis or other coexistent diseases, sufficient MRI protocol	By a radiologist and an oncologist
DL	Zhang M et al 2020[19]	Brigham and Women's Hospital, Harvard Medical School, USA	NA	Retro.	Internal random split	361 (316/45)	1.6 (NA/ NA)	NA (NA/ NA)	2053 (1777/276)	NA (NA/ NA)	35.8	NA	By two radiologists, with FU MRI
DL	Zhou Z et al 2020[20]	University of Texas MD Anderson Cancer Center, USA	January 2011-August 2018	Retro.	Internal random split	266 (212/54)	1.5 (0.8/0.9)	39.1 (37.3/ 46.3)	1147 (913/383)	10 mm (NA/ NA)	47.6 (34.2/61.1)	Candidate for stereotactic surgery	By 3 radiologists and a radiation neuro-olcolgist

Abbreviation: CV, cross-validation; Dev/Test; Developmental data set/ Test data set; DL, deep learning; cML, classical machine learning; NA, not available; No., number; Pt. patients; Pros, prospective; Retro, retrospective;

* Shandong Provincial Hospital; Affiliated Hospital of Qingdao University Medical College; Second Hospital of Shandong University; † We assumed number of the nodules of the test set by estimating the proportion between patients sample size or by a division of "n" of cross-validation (eg, n-fold cross-validation); ‡ Proportion of lung cancer among the total underlying cancers.

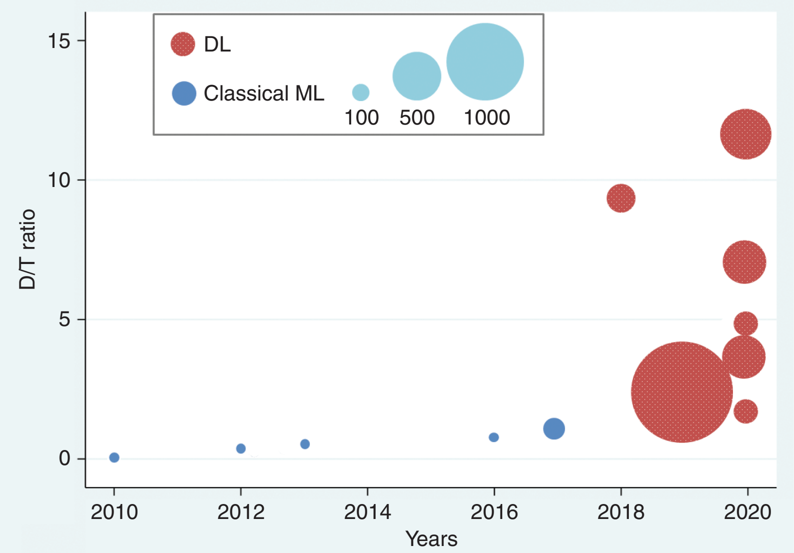

Fig. 2 Distribution of studies by the year of publication (x-axis) and the ratio of patient numbers in the developmental set compared with the test set (D/T ratio) (y-axis). The size of the circles represents the sample size of each study. ML, machine learning; DL, deep learning.

Pooled Detectability Performance of the MRI

For all 12 included studies, the pooled proportion of detectability of machine learning using MRI data for the patients with BM was 90.0% (95% confidence interval [CI], 85–93%), ranging from 81.1% to 98% (Table 2 and Figure 3). Heterogeneity was present ($I^2 = 90\%$). The subgroup pooled proportion of detectability of the cML group, and DL group were 88.7% (95% CI, 84–93%) and 90.1% (95% CI, 84–95%), respectively, with no statistical difference. The DL group showed a significantly lower false-positive rate per person than the cML group (10 vs 135, $P < 0.001$). Otherwise, there was no significant factor affecting the heterogeneity among the type of machine learning (cML vs DL), sample size ratio (developmental set/test set), nodule number ratio (developmental set/ test set), and nodular number per person on the meta-regression analysis.

Discussion

In this study, the machine learning studies of BM detection on brain MRI were systematically evaluated concerning the demographics, MRI and machine learning methodology, and quality of reporting. In terms of methodology, machine learning research can be categorized into cML and DL. A clear transition from cML to DL was noted after 2018, and the number of papers on automated detection of BM has gradually increased since then. Therefore, at this point, it is important to systematically review the published machine learning research and provide guidance for future studies. We found that cML studies were based on a smaller sample size than DL studies. Although the pooled proportion of detectability of BM between the cML and DL groups was not significantly different (88.7% and 90.1%, respectively), the false-positive rate per person was significantly lower in the DL group than the cML group ($P < 0.001$). Based on the modified version of CLAIM and QUADAS-2 criteria, we found that at least some studies were considered to have a high or unclear risk of bias in the domains of patient selection, reference standard, and flow and timing. These results shed light on the current state of the technology, as well as the need for quality improvement.

As the overall incidence of BM is estimated to be 20% in systemic oncology patients, and it is a major contributor to cancer mortality in patients with advanced-stage cancer, detection of BM is essential so that treatment can be initiated.[1,2] Advances in neuroimaging, particularly the use of 3D MR imaging, has enhanced the sensitivity of BM detection at initial cancer detection.[4–6] Among the common underlying cancer origin including lung, breast, skin, colon, pancreas, testes, ovary, cervix, renal cell carcinoma, and melanoma,[1] the current National Comprehensive Cancer Network guidelines recommend screening brain MRI for patients with lung cancer (stage II to IV non-small-cell lung cancer, small cell lung cancer of any stage) and melanoma (stage IIIC to IV) regardless of the patient's neurologic suspicion.[40] In addition, due to recent technological advances in stereotactic radiosurgery, accurate diagnosis and localization of BM have become even more important. However, despite such demand, there are several challenges in accurate BM detection for radiologists. Thinner slice thickness images led to an increased detection rate of tiny BM nodules but increase the workload, detect mimickers such as small vessels more frequently,[5] and increase the risk of medico-legal problems in case of detection failure.[7]

CAD systems for BM using machine learning methodology have been proposed to overcome these issues.[9–20] Generally, the 2 groups (cML and DL) share

Table 2 Characteristics and performance of the MRI machines and machine learning techniques

Group	Source	MRI				Machine learning						Performance	
		Machine	Tesla	In-plane resolution, mm	Slice thickness, mm	Algorithm (name)	Skull stripping	Intensity Normalization	Segmentation	Data Augmentation	Enhancement degree threshold	Detectability	False-positive rate
cML	Ambrosini RD et al 2010[9]	Signa, GE	1.5	0.43	2.5	Template-matching algorithm	Yes	Yes	No	No	NA	90.5	36
cML	Farjam R et al 2012[13]	NA, Philips	3	0.94	1	template-matching algorithm + cross-correlation method	No	Yes	No	NA	Yes	93.5	3.3
cML	Pérez-Ramírez Ú et al 2016[15]	Magnetom Symphony; Siemens	1.5	0.5	1	template-matching algorithm	NA	Yes	Yes	NA	NA	88.1	6.91
cML	Sunwoo L et al 2017[16]	Intera; Achieva; Ingenia, Philips	1.5; 3	1	1	template-matching and K-means clustering algorithm	Yes	Yes	Yes	No	NA	87.3	302.4
cML	Yang S et al 2013[18]	Magnetom Trio, Siemens	3	1	1	template-matching algorithm	Yes	Yes	No	No	NA	81.1	0.42
DL	Bousabarah K et al 2020[10]	Ingenia; Archieva; Intera, Philips	3; 3; 1.5	NA	NA	3D CNN (U-Net, mo U-Net)	NA	Yes	Yes	Yes	NA	82	0.35
DL	Charron O et al 2018[11]	NA	1.5	0.82	1	3D CNN (DeepMedic)	Yes	Yes	Yes	Yes	Yes	98	14.2
DL	Dikici E et al 2020[12]	Aera; Avanto; Espree; Skyra;Triotrim, Siemens; Optima MR450w; Signa HDxt, GE	1.5 or 3	0.49–1	0.8–1	3D CNN (CropNet)	NA	Yes	Yes	Yes	NA	90	9.12
DL	Grøvik E et al 2020[14]	SIGNA Explorer; TwinSpeed; Discovery 750; SIGNA Architect, GE	1.5; 1.5; 3; 3; 3	NA	1	2.5D CNN (modified GoogLeNet)	Yes	Yes	Yes	Yes	NA	83	3.4
DL	Xue J et al 2019[17]	MAGNETOM; Discovery MR750, Skyra; GE	3	1	1.5; 1.2; 1	3D CNN (BMDS net)	No	Yes	Yes	Yes	NA	96†	NA
DL	Zhang M et al 2020[19]	NA, Siemens and GE	1.5; 3	0.43–1, 0.47–1	0.6, 1	3D CNN (Faster R-CNN)	Yes	Yes	NA	Yes	Yes	95.6	19.9
DL	Zhou Z et al 2020[20]	Signa HDxt; Discovery MR750w, GE	1.5; 3	0.94	1	single-shot detector	No	Yes	No	Yes	NA	81	7

Abbreviations: CI, confidence interval; CNN, convolutional neural network; DL, deep learning; cML, classical machine learning; 3D, 3-dimensional.

* Charron O et al 2018 performed additional analysis for the necrotic tumor; Farjam R et al 2012 and Zhang M et al 2020 defined the degree of enhancement as follows: contrast to intensity variation ratio (CVIR) and contrast-to-background ratio (CBR), respectively.; †The study yielded detectability per voxel-wise proportion, whereas the others presented per nodule-wise proportion.

Table 3 Quality assessment according to the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) criteria

Source	Risk of Bias											Concern of Applicability		
	Patient selection			Index Test (Machine Learning)		Reference Standard		Flow and Timing				Patient Selection	Index Test§	Reference Standard
	Consecutive	Case-control	Inappropriate Exclusion	Blinded to Ref/Stan.	Prespecified-threshold	Correctly Classified Target Condition	Blinded to Index test	Adequate Interval: Index to Ref/Stan.	All patients with Ref/Stan.	Same Ref/Stan.	All patients analyzed			
Ambrosini RD et al 2010[9]	Uncelar	No	No	Yes	Yes	Yes†	Yes	Yes	Yes	Yes	Yes	Low	Unclear	Low
Farjam R et al 2012[13]	Uncelar	No	Yes*	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Uncelar‡	Low	Unclear	Low
Pérez-Ramírez Ú et al 2016[15]	Uncelar	No	No	Yes	Yes	Yes†	Yes	Yes	Yes	Yes	Yes	Low	Unclear	Low
Sunwoo L et al 2017[16]	Yes	No	No	Yes	Yes	Yes†	Yes	Yes	Yes	Yes	Yes	Low	Low	Low
Yang S et al 2013[18]	Uncelar	No	Yes*	Yes	Yes	Yes†	Yes	Yes	Yes	Yes	Unclear‡	Low	Unclear	Low
Bousabarah K et al 2020[10]	Uncelar	No	No	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Low	Unclear	Low
Charron O et al 2018[11]	Uncelar	No	No	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Low	Unclear	Low
Dikici E et al 2020[12]	Uncelar	No	Yes*	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Low	Unclear	Low
Grøvik E et al 2020[14]	Yes	No	No	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Low	Unclear	Low
Xue J et al 2019[17]	Uncelar	No	No	Yes	Yes	Yes†	Yes	Yes	Yes	Yes	Yes	Low	Low	Low
Zhang M et al 2020[19]	Unclear	No	No	Yes	Yes	Yes†	Yes	Yes	Yes	Yes	Yes	Low	Unclear	Low
Zhou Z et al 2020[20]	Uncelar	No	No	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Low	Unclear	Low

* The studies excluded nodules using a size criteria; † The studies assessed the metastasis on follow up MRI.; ‡ Several patients were excluded during the size thresholding process.; § The temporal separation and external validation data sets were considered as a low concern of applicability.

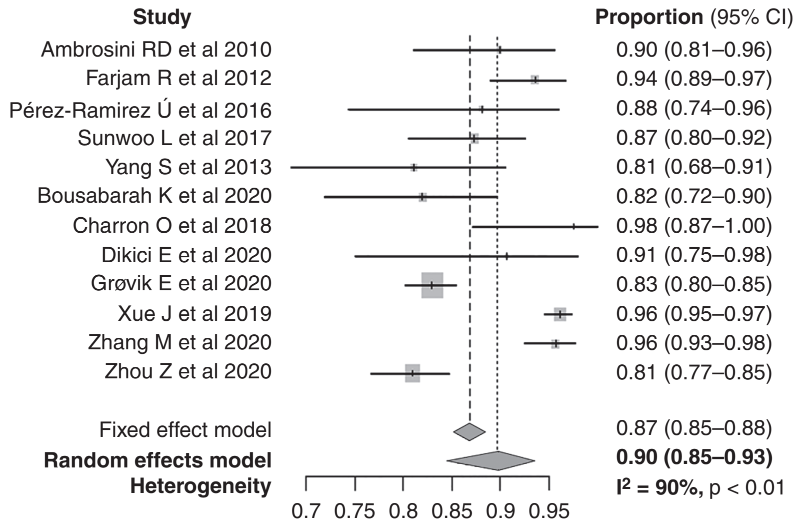

Fig. 3 Forest plot of the pooled proportion of detectability.

methodologically common features, such as the use of development and validation datasets, the need for data labeling, and preprocessing. In both groups, the development dataset is used to train the model, which needs to be sufficiently large so that it consists of a sufficient number of observations to reach the optimal performance.[41,42] However, the sample size of the development dataset required for training the model sufficiently differs in both groups. In most cML methods, the features are predetermined by an expert with domain knowledge. Thus, the data are mainly used to determine thresholds or parameters of the model in cML; hence, the required data for training tend to be relatively limited. On the other hand, the features are not predefined in DL. Instead, they are subsequently learned by the algorithm itself.[43] In addition, because DL is generally built with much more complex network architecture than is cML, reaching up to several millions of hyperparameters for recent forms of convolutional neural networks,[44] inevitably requires much bigger data. As such, we found that the number of enrolled patients and the metastatic nodules in this study were relatively larger in the DL group, ranging from 62 to 584 and 932 to 3264, respectively, compared with the cML group, in which it ranged from 26 to 140 and 158 to 1632, respectively. Furthermore, it was noted that only studies in the DL group performed data augmentation during the model development. Data augmentation refers to boosting the training data by rotation, resizing, or signal modulation to compensate for the relative shortage of data, thereby facilitating the training and reducing the risk of overfitting.[45] On the other hand, in terms of the test set, there is relatively little difference between cML and DL in its size. Consequently, the sample size ratio of developmental/test was higher than 4 in the majority of DL studies, whereas the ratio was less than 1 in most cML studies.[46,47]

In this study, we found that the false-positive rate per person was significantly reduced in the DL group compared with the cML group, while the pooled proportion of BM detectability remained comparable. This suggests that overall, the DL group tended to perform better than cML group. At a given performance level, if we choose to set the algorithm to be more sensitive (towards higher detectability), the algorithm will become less specific (towards higher false-positive rate), and vice versa. Thus, since the detectability was fairly high for cML studies, recent DL studies might have put more effort into lowering the false-positive rate per person while maintaining the detectability. Due to the limited number of analyzed studies and the “black-box” nature of DL,[21–23] it is challenging to assess the reasons for improved performance of DL studies compared with cML studies. Although the handcrafted features selected by the experts with domain knowledge in classical ML seems to be quite effective, the combination of more complex network architecture with bigger training data might have enabled the DL model to learn formerly unknown clues in detecting a nodule and differentiating true metastasis from false-positive lesions. In addition, the use of data augmentation would have contributed to enriching the development dataset. We believe that the reduction of the false-positive rate has clinical implications as this would help alleviate the increased workload of human readers using CAD systems. Future studies need to be carried out to clarify the reason for the improved performance and to enhance the explainability of DL.

Given that we are undergoing a transition period from cML to DL, it would be particularly meaningful to perform a quality assessment of research regarding the detection of BM on brain MRI. Unfortunately, the quality assessment tool for reporting the detection performance of machine learning studies has not yet been fully established. Recently, to address applications of artificial intelligence in medical imaging that include classification, image reconstruction, text analysis, and workflow optimization, CLAIM has been proposed as an extension of the Standards for Reporting and Diagnostic Accuracy Studies (STARD) guideline.[25,48] Therefore, we assessed the included studies using the QUADAS-2 criteria framework,[26] adjusted by CLAIM. One concern is whether the dataset reflects the real-world data in terms of the clinical aspect. 3 of the enrolled studies excluded nodules during data handling due to their size. Only two of the enrolled studies declared that patients were "consecutively collected" during the patient enrollment process, and only two performed data sampling with a temporal split or external validation. Non-consecutive sampling has a risk to obscure the spectrum of the disease in the dataset.[22,49] Many reviewers of machine learning research recommend consecutive rather than convenient sampling, split rather than random sampling, temporal split in internal sampling studies, and external or global sampling methods to reduce the overfitting and spectrum bias, as well as enhance the clinical impact and generalizability.[21–23] Many existing guidelines, such as the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD), are a work-in-progress with regard to machine learning studies.[50] These guidelines will help to improve the robustness of the methodology used, lower the overfitting and spectrum bias, and ultimately enhance the reproducibility of machine learning studies.

This study has some limitations. First, a relatively small number of studies are included. However, the numbers of enrolled studies in the cML and DL groups were well balanced to offer parallel evaluation. Second, we could not compare the performance of BM detection between a human performer and CAD due to the limited number of available studies. A further comparative study is warranted to confirm the added value of CAD.

In conclusion, a comparable detectability of BM with a low false-positive rate per person is noted in the DL group compared with the cML group. Improvements are required in terms of quality and study design.

Supplementary Material

Supplementary data are available at *Neuro-Oncology* online.

Keywords

artificial intelligence | brain metastasis | deep learning | machine learning | magnetic resonance imaging

Funding

This study was supported by a grant from the National Research Foundation of Korea (grant number: NRF-2018R1C1B6007917).

Conflict of interest statement. The authors report no conflict of interest, and the present study has not been presented elsewhere.

Authorship statement. S.J.C.: experimental design, the implementation, analysis, and interpretation of the data, writing of the draft manuscript, and approval of the final version. L.S.: experimental design, the implementation, analysis, and interpretation of the data, writing of the manuscript at the revision stage, and approval of the final version. S.H.B., Y.J.B., B.S.C., and J.H.K.: analysis and interpretation of the data, writing of the manuscript at the revision stage, and approval of the final version.

References

1. Fink KR, Fink JR. Imaging of brain metastases. *Surg Neurol Int.* 2013;4(Suppl 4):S209–S219.
2. Achrol AS, Rennert RC, Anders C, et al. Brain metastases. *Nat Rev Dis Primers.* 2019;5(1):5.
3. Nieder C, Grosu AL, Gaspar LE. Stereotactic radiosurgery (SRS) for brain metastases: a systematic review. *Radiat Oncol.* 2014;9:155.
4. Suh JH, Kotecha R, Chao ST, Ahluwalia MS, Sahgal A, Chang EL. Current approaches to the management of brain metastases. *Nat Rev Clin Oncol.* 2020;17(5):279–299.
5. Pope WB. Brain metastases: neuroimaging. *Handb Clin Neurol.* 2018;149:89–112.
6. Patel SH, Robbins JR, Gore EM, et al; Expert Panel on Radiation Oncology–Brain Metastases. ACR Appropriateness Criteria® follow-up and retreatment of brain metastases. *Am J Clin Oncol.* 2012;35(3):302–306.
7. Triebel KL, Gerstenecker A, Meneses K, et al. Capacity of patients with brain metastases to make treatment decisions. *Psychooncology.* 2015;24(11):1448–1455.
8. Nishikawa RM, Schmidt RA, Linver MN, Edwards AV, Papaioannou J, Stull MA. Clinically missed cancer: how effectively can radiologists use computer-aided detection? *AJR Am J Roentgenol.* 2012;198(3):708–716.
9. Ambrosini RD, Wang P, O'Dell WG. Computer-aided detection of metastatic brain tumors using automated three-dimensional template matching. *J Magn Reson Imaging.* 2010;31(1):85–93.
10. Bousabarah K, Ruge M, Brand JS, et al. Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data. *Radiat Oncol.* 2020;15(1):87.
11. Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Comput Biol Med.* 2018;95:43–54.
12. Dikici E, Ryu JL, Demirer M, et al. Automated brain metastases detection framework for T1-weighted contrast-enhanced 3D MRI. *IEEE J Biomed Health Inform.* 2020;24(10):2883–2893.
13. Farjam R, Parmar HA, Noll DC, Tsien CI, Cao Y. An approach for computer-aided detection of brain metastases in post-Gd T1-W MRI. *Magn Reson Imaging.* 2012;30(6):824–836.
14. Grøvik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *J Magn Reson Imaging.* 2020;51(1):175–182.

15. Pérez-Ramírez Ú, Arana E, Moratal D. Brain metastases detection on MR by means of three-dimensional tumor-appearance template matching. *J Magn Reson Imaging.* 2016;44(3):642–652.
16. Sunwoo L, Kim YJ, Choi SH, et al. Computer-aided detection of brain metastasis on 3D MR imaging: observer performance study. *PLoS One.* 2017;12(6):e0178265.
17. Xue J, Wang B, Ming Y, et al. Deep learning-based detection and segmentation-assisted management of brain metastases. *Neuro Oncol.* 2020;22(4):505–514.
18. Yang S, Nam Y, Kim MO, Kim EY, Park J, Kim DH. Computer-aided detection of metastatic brain tumors using magnetic resonance black-blood imaging. *Invest Radiol.* 2013;48(2):113–119.
19. Zhang M, Young GS, Chen H, et al. Deep-learning detection of cancer metastases to the brain on MRI. *J Magn Reson Imaging.* 2020;52(4):1227–1236.
20. Zhou Z, Sanders JW, Johnson JM, et al. Computer-aided detection of brain metastases in T1-weighted MRI for stereotactic radiosurgery using deep learning single-shot detectors. *Radiology.* 2020;295(2):407–415.
21. Gregory J, Welliver S, Chong J. Top 10 reviewer critiques of radiology Artificial Intelligence (AI) articles: qualitative thematic analysis of reviewer critiques of machine learning/deep learning manuscripts submitted to JMRI. *J Magn Reson Imaging.* 2020;52(1):248–254.
22. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology.* 2018;286(3):800–809.
23. Park SH, Kressel HY. Connecting technological innovation in artificial intelligence to real-world medical practice through rigorous clinical validation: what peer-reviewed medical journals could do. *J Korean Med Sci.* 2018;33(22):e152.
24. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med.* 2009;151(4):W65–W94.
25. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiology: Artificial Intelligence.* 2020;2(2):e200029.
26. Whiting PF, Rutjes AW, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529–536.
27. Kim KW, Lee J, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers—part I. general guidance and tips. *Korean J Radiol.* 2015;16(6):1175–1187.
28. Lee J, Kim KW, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers—part II. Statistical methods of meta-analysis. *Korean J Radiol.* 2015;16(6):1188–1196.
29. Suh CH, Park SH. Successful publication of systematic review and meta-analysis of studies evaluating diagnostic test accuracy. *Korean J Radiol.* 2016;17(1):5–6.
30. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557–560.
31. Ainsworth NL, McLean MA, McIntyre DJO, et al. Quantitative and textural analysis of magnetization transfer and diffusion images in the early detection of brain metastases. *Magn Reson Med.* 2017;77(5):1987–1995.
32. Azimi P, Shahzadi S, Sadeghi S. Use of artificial neural networks to predict the probability of developing new cerebral metastases after radiosurgery alone. *J Neurosurg Sci.* 2020;64(1):52–57.
33. El Kader Isselmou A, Xu G, Zhang S. Improved methods for brain tumor detection and analysis using MR brain images. *Biomed Pharmacol J.* 2019;12(4):1621–1631.
34. Jun Y, Eo T, Kim T, et al. Deep-learned 3D black-blood imaging using automatic labelling technique and 3D convolutional neural networks for detecting metastatic brain tumors. *Sci Rep.* 2018;8(1):9450.
35. Shearkhani O, Khademi A, Eilaghi A, et al. Detection of volume-changing metastatic brain tumors on longitudinal MRI using a semiautomated algorithm based on the Jacobian operator field. *AJNR Am J Neuroradiol.* 2017;38(11):2059–2066.
36. Sikpa D, Fouquet JP, Lebel R, Diamandis P, Richer M, Lepage M. Automated detection and quantification of breast cancer brain metastases in an animal model using democratized machine learning tools. *Sci Rep.* 2019;9(1):17333.
37. Szwarc P, Kawa J, Rudzki M, Pietka E. Automatic brain tumour detection and neovasculature assessment with multiseries MRI analysis. *Comput Med Imaging Graph.* 2015;46(Pt 2):178–190.
38. Perez-Ramirez U, Arana E, Moratal D. Computer-aided detection of brain metastases using a three-dimensional template-based matching algorithm. *Annu Int Conf IEEE Eng Med Biol Soc.* 2014;2014:2384–2387.
39. Noguchi T, Uchiyama F, Kawata Y, et al. A fundamental study assessing the diagnostic performance of deep learning for a brain metastasis detection task. *Magn Reson Med Sci.* 2020;19(3):184–194.
40. Ettinger DS, Wood DE, Aisner DL, et al. Non-small cell lung cancer, version 5.2017, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw.* 2017;15(4):504–535.
41. Alwosheel A, van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J Choice Model.* 2018;28: 167–182.
42. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402–2410.
43. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *Radiographics.* 2017;37(2):505–515.
44. Gowda SN, Yuan C. ColorNet: investigating the importance of color spaces for image classification. *Lect Notes Comput Sci.* 2019;11364:581–596.
45. Lemley J, Bazrafkan S, Corcoran P. Smart augmentation learning an optimal data augmentation strategy. *IEEE Access.* 2017;5: 5858–5869.
46. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. *arXiv.* 2016;1511.06348
47. Linjordet T, Balog K. *Impact of Training Dataset Size on Neural Answer Selection Models.* 2019;(14–18):828–835.
48. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open.* 2016;6(11):e012799.
49. Sica GT. Bias in research studies. *Radiology.* 2006;238(3):780–789.
50. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet.* 2019;393(10181):1577–1579.