BMC Cancer

SYSTEMATIC REVIEW Open Access

Accuracy of artificial intelligence in detecting tumor bone metastases: a systematic review and meta-analysis

Huimin Tao[1,4†], Xu Hui[2,3†], Zhihong Zhang[1], Rongrong Zhu[1,4], Ping Wang[4], Sheng Zhou[4*] and Kehu Yang[2,3*]

Abstract

Background Bone metastases (BM) represent a prevalent complication of tumors. Early and accurate diagnosis, however, is a significant hurdle for radiologists. Recently, artificial intelligence (AI) has emerged as a valuable tool to assist radiologists in the detection of BM. This meta-analysis was undertaken to evaluate the AI diagnostic accuracy for BM.

Methods Two reviewers performed an exhaustive search of several databases, including Wei Pu (VIP) database, China National Knowledge Infrastructure (CNKI), Web of Science, Cochrane Library, Ovid-Embase, Ovid-Medline, Wan Fang database, and China Biology Medicine (CBM), from their inception to December 2024. This search focused on studies that developed and/or validated AI techniques for detecting BM in magnetic resonance imaging (MRI) or computed tomography (CT). A hierarchical model was used in the meta-analysis to calculate diagnostic odds ratio (DOR), negative likelihood ratio (NLR), positive likelihood ratio (PLR), area under the curve (AUC), specificity (SP), and pooled sensitivity (SE). The risk of bias and applicability were assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST), while the Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis-artificial intelligence (TRIPOD-AI) was employed for evaluating the quality of evidence.

Result This review covered 20 articles, among them, 16 studies were included in the meta-analysis. The results revealed a pooled SE of 0.88 (0.82–0.92), a pooled SP of 0.89 (0.84–0.93), a pooled AUC of 0.95 (0.92–0.96), PLR of 8.1 (5.57–11.80), NLR of 0.14 (0.09–0.21) and DOR of 58 (31–109). When focusing on imaging algorithms. Based on ML, a pooled SE of 0.88 (0.77–0.92), SP 0.88 (0.82–0.92), and AUC 0.93 (0.91–0.95). Based on DL, a pooled SE of 0.89 (0.81–0.95), SP 0.89 (0.81–0.94), and AUC 0.95 (0.93–0.97).

Conclusion This meta-analysis underscores the substantial diagnostic value of AI in identifying BM. Nevertheless, in-depth large-scale prospective research should be carried out for confirming AI's clinical utility in BM management.

Keywords Bone metastases, Artificial intelligence, Diagnosis, Meta-analysis, Systematic review

†Huimin Tao and Xu Hui contributed equally to this work.

*Correspondence:
Sheng Zhou
Lzzs@sina.com
Kehu Yang
yangkh-ebm@lzu.edu.cn
[1]The First Clinical Medical College of Gansu, University of Chinese Medicine, Lanzhou, Gansu 730000, China
[2]Evidence-Based Medicine Centre, School of Basic Medical Science, Lanzhou University, Lanzhou 730000, China
[3]Centre for Evidence-Based Social Science/Center for Health Technology Assessment, School of Public Health, Lanzhou University, Lanzhou 730000, China
[4]Department of Radiology, Gansu Provincial Hospital, Lanzhou, Gansu 730000, China

Introduction

Bone metastases (BM) represent a common cancer complication [1]. Research indicates that approximately 350,000 people in the United States die each year due to BM [2]. Following metastases to the liver and lungs, bone is the next most common site for distant cancerous spread. BM primarily affects the axial skeleton, specifically the thoracic and lumbar vertebrae [3, 4]. Autopsy studies show that the BM incidence varies among different cancer types. Following prostate cancer at 68% and thyroid cancer at 42%, breast cancer has the highest prevalence rate at 73% [4–6]. BM triggers skeletal-related events (SREs), including spinal cord compression, hypercalcemia, pathological fractures, and pain [6, 7]. BM is widespread in patients with advanced cancer [8] and is challenging to treat [9]. However, anticancer treatments through integrated management typically control disease progression and mitigate the effects of BM on physical function. This approach should incorporate bone-targeting agents (BTAs) and appropriate local treatments including specialist palliative care, orthopaedic surgery, and radiation therapy [10, 11]. At present, studies [12–14] have reported immunotherapy related to tumors. The relationship between microbiota and disease is receiving increasing attention. As one of the diseases with frequent BM, the development and treatment of prostate cancer is closely related to the microbiota [15].

Given the aging population and increasing cancer incidence rates, the challenge of addressing BM is expected to significantly burden healthcare systems in the coming years [16, 17]. Accurate early BM diagnosis shows high importance in effective tumor staging, treatment planning, and prognosis assessment, placing a high importance on the role of radiologists [18]. Currently, CT and MRI are the predominant imaging methods used in clinical practice to detect BM. The primary advantages of CT and MRI include their relatively low cost and minimal radiation exposure. This study specifically focuses on the use of AI for identifying tumor BM through CT and MRI imaging. However, BM often presents similarly to other conditions on these imaging modalities, such as islands of bone, multiple myeloma (MM), and various osteolytic lesions, making the image evaluation process challenging and prone to errors, potentially leading to missed lesions and decreased SE [18]. AI is a branch of computer science that focuses on creating systems capable of executing activities usually needing human cognition [19]. There is a growing body of research exploring AI's potential as an adjunct tool in diagnosing tumor BM [20]. For instance, Noguchi et al. [21] developed a deep learning-based algorithm (DLA) that automatically detects bone metastases across all scanned regions. This DLA achieved an SE of 89.8% (44 of 49 cases) in the validation set and 82.7% (62 of 75 cases) in the test set. Additionally, the SE of radiologists in detecting BM increased from 51.7 to 71.7% in lesion-based analysis and from 74.4 to 91.1% in case-based analysis. Xiong et al. [22] introduced a machine learning (ML) classifier aimed at differentiating between vertebral metastasis and MM on lumbar spine MRI, achieving a diagnostic accuracy of 0.815. Despite several studies demonstrating AI's high accuracy in diagnosing BM, debates continue over whether ML or deep learning (DL) algorithms perform better, with mixed evidence from various studies. Despite the potential of automated approaches to reduce errors and enhance decision-making, there is still limited consensus on their reliability in real healthcare settings. To address these uncertainties, our research involved a comprehensive systematic review and meta-analysis to analyze the diagnostic capabilities of AI-based technologies in identifying BM, covering all pertinent studies.

Materials and methods

Protocol and registration

This systematic review was registered with PROSPERO (CRD: 42023452597). We conducted our study following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies guidelines [23–25].

Search strategy

We conducted a comprehensive search across multiple databases including Ovid-Medline, Ovid-Embase, Web of Science, Cochrane Library, CNKI, VIP database, Wan Fang database, and CBM database from their inception until December 2024. The objective was to identify studies that developed and/or validated AI algorithms and models for detecting BM from malignancies. Our search strategy utilized a combination of keywords and subject terms such as "Artificial Intelligence," "Random Forest," "Deep Learning," "Radiomics," "Support Vector Machine," "Decision Tree," "Machine Learning," "Bone and Bones," "Neoplasm Metastasis," "Sensitivity," and "Specificity." We also manually reviewed the reference lists of included studies to uncover any potentially overlooked publications. Table S1 presents the detailed search strategy.

Study selection

We included studies that applied AI and radiomics to diagnose and predict BM. The inclusion criteria were as follows:

1. Study Types: Cohort studies (both prospective and retrospective), case-control studies, and cross-sectional studies.
2. Participants: patients with tumor BM.
3. Exposure: Diagnosis of BM based on imaging data (CT/MRI) using AI algorithms.

Reference Standard: Patients with a diagnosis of BM confirmed by histopathology, radiology, and clinical assessments.

4. Outcomes: Studies must report diagnostic performance indices of the AI algorithms, such as SE, SP, positive predictive value (PPV), negative predictive value (NPV), PLR, NLR, DOR, or accuracy. These metrics should help estimate the true negative (TN), false negative (FN), false positive (FP), and true positive (TP) values.

Exclusion criteria were as follows

1. Study Type Discrepancies: Exclusion of case reports, reviews, conference abstracts, letters to the editor, guidelines, etc. 2. Participant Incompatibility: Exclusion of patients who did not undergo both CT and MRI. 3. Inconsistent Exposure Factors: Exclusion of studies that did not use CT/MRI imaging data for diagnosing bone metastases. 4. Inconsistent Reference Criteria: Exclusion of studies that did not use histopathology, radiology, and clinical diagnosis as reference standards.

Data extraction

The data extraction process was independently carried out by two reviewers (THM, ZZH) using predefined data extraction tables. Any conflicts that arose were resolved by a third reviewer (ZRR). The extracted information from the included studies comprised:

1. Basic characteristics of the study (study design, country, year of publication, authors),
2. Basic characteristics of the participants (sample size, gender, age),
3. Model characteristics (sample sizes of training, test, and validation sets, reference standards, type of AI model, medical image type),
4. Diagnostic performance metrics: TP, TN, FP, FN.

We utilized Review Manager 5.3 to estimate missing data based on the information available in the text or appendices of each study. For articles reporting multiple datasets (e.g., test sets, validation, and training) concurrently, we applied the method proposed by Liu et al. [26]. Method I involved treating each dataset as an independent study in the meta-analysis. Method II involved extracting the highest accuracy data from each study for inclusion in the meta-analysis.

Quality assessment

The risk of bias for each study was evaluated using the PROBAST [27]. This tool includes 20 questions divided into four domains: participant selection, predictors, outcomes, and analysis. Additionally, we assessed compliance with reporting guidelines using the TRIPOD-AI [28, 29] checklist for multivariate prediction models. TRIPOD-AI offers standardized guidance for reporting studies on prediction models, regardless of whether they employ regression modeling or ML methods. This checklist, comprising 27 items, is crucial for the transparent reporting of prediction model development and validation.

Meta-regression and subgroup analysis

We conducted meta-regression incorporating variables such as sample size, image quality, age distribution, proportion of female participants, imaging modality (CT/MRI), algorithm type (DL/ML), use of external validation data, and implementation of data augmentation techniques. The algorithms were categorized into two groups: DL algorithms, including Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN), and ML algorithms, including Voted Perceptron (VP), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Logistic Regression (LR), and Support Vector Machine (SVM).

Subgroup analyses were performed separately based on predetermined criteria and assessed for heterogeneity. These predefined subgroup analyses included factors such as imaging modality (CT vs. MRI), sample size, unit of data (number of patients or lesions), whether data enhancement was conducted, and type of study (single-center vs. multicenter).

Sensitivity analysis

To further mitigate the impact of confounding bias, we conducted sensitivity analyses using four distinct approaches: employing the "leave-one-out" method to individually exclude each study, excluding a study with low adherence to TRIPOD-AI guidelines (< 50%), excluding 3 studies with low risk of bias, and combining data using two different methodologies proposed by Liu [26].

Data synthesis and analysis

We utilized MetaDiSc 1.4 (XI Cochrane Colloquium, Barcelona, Spain) and the MIDAS module of STATA version 16 (Stata Corp LP, College Station, USA) to compute 2×2 contingency tables in our study. MetaDiSc 1.4 was employed to detect any threshold effects. Initially, we assessed the threshold effect using Spearman's correlation coefficient. As no threshold effect was observed, we opted for the bivariate mixed-effect model [30]. Subsequently, the following parameters were determined along with their 95% confidence intervals (CIs): the Summary Receiver Operating Characteristics curve (SROC), SP, PLR, NLR, DOR, and SE. The SROC curves serve as meta-analysis tools that aggregate SE and SP data from

various studies to evaluate the efficacy of diagnostic tests comprehensively [31]. The AUC is a crucial measure of the curve's overall performance, ranging from 0 to 1, where 0 indicates a complete failure of the test to correctly diagnose any cases, and 1 indicates flawless distinction between all cases and non-cases [32–34]. We used Q-tests and I^2 statistics to assess heterogeneity across studies, with $I^2 > 50\%$ and/or $P < 0.05$ indicating high heterogeneity. Lastly, Fagan's nomogram was employed to examine the pre-test and post-test probability relationship [35].

Publication bias

When the analysis included more than ten studies, publication bias was detected through Deek's funnel plot asymmetry test. The funnel plot illustrates publication bias by depicting the relationship between the effect size of each study and its precision, often expressed as the reciprocal of the sample size or effective sample size (ESS) [36]. Ideally, the funnel plot should be symmetrical; an asymmetrical funnel plot may indicate the presence of publication bias. We utilized Deeks' Test—a regression-based method—to assess the statistical symmetry of the funnel plot. An absence of publication bias can be indicated by a p-value greater than 0.05.

Result

Selection of studies and characteristics

We initially retrieved a total of 5,300 articles, from which 1,239 were identified as duplicates. After screening the remaining 4,061 articles by reviewing their titles and abstracts, we excluded 3,970. We then assessed 91 full-text articles in terms of eligibility, excluding 71 of them. Consequently, 20 articles were included in the systematic review. Of these, 16 studies (80%) [18, 20, 37–48] offered sufficient data for constructing contingency tables and were covered in the meta-analysis. Figure 1 presents the process of literature screening.

Tables 1 and 2 present specific characteristics of the 20 studies published between 2019 and 2024. All studies were retrospective and utilized supervised learning methods for analysis. Fourteen (70%) studies [18, 20, 37–46, 49–53] specifically focused on using AI methods to diagnose tumor BM. The remaining 6 (30%) studies explored AI's ability to differentiate BM from other diseases; 4 (20%) studies [22, 45, 54, 55] addressed

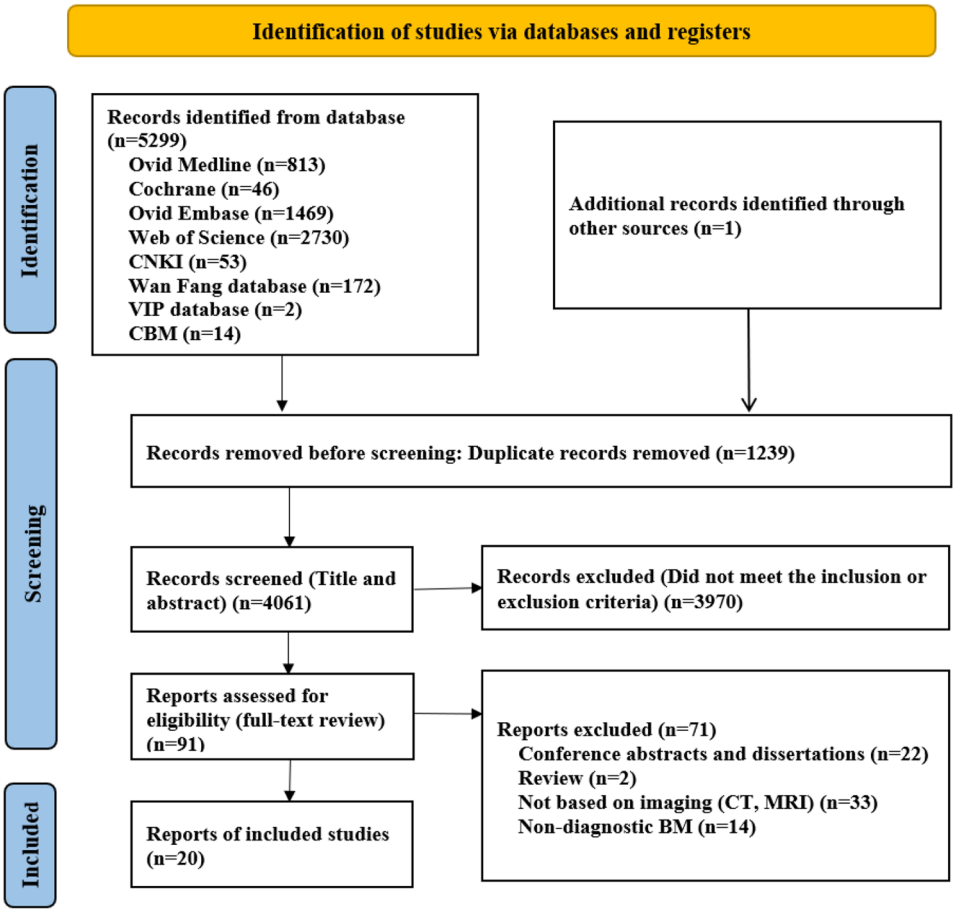

Fig. 1 PRISMA flow chart of the literature retrieval

Table 1 Characteristics of the 17 included studies

Author	Year	Country	Primary tumor(*n*)	Scan area	Reference standard	Positive patients (*n*)	Negative patients (*n*)	Female (%)	Age (mean)	Pictures (*n*)	No. of Images per Set			Proportion of data sets
											Train	Test	Validation	
Xiong	2021	China	lung (30), nasopharyngeal (13), breast (11), others (6)	lumbar	core needle or excisional biopsy	60	47	38	60.6	178	75	NR	32	7:NR:3
Wang	2023	China	NR	spinal	NR	636	305	NR	NR	941	NR	NR	NR	8:2:NR
Shi	2023	China	lung (10), breast (13)	spine and vertebral	pathology or MRI	23	28	53	60.3	137	102	35	NR	3:1:NR
Özgül	2023	Turkey	lung (44), kidneys (21), thyroid gland (12), bladder (9), others (16)	peripheral skeleton	radiology, nuclear medicine	102	70	40	60.2	172	120	52	NR	7:3:NR
Noguchi	2022	Japan	lung (57), prostate (33), breast (25), others (54)	vertebra, Pelvis, rib, scapula, limb, sternum, clavicle	bone scintigraphy or FDG-PET	219	513	47	64.1	1375	1375	75	49	55:3:2
Liu	2021	China	prostate (230)	pelvic	biopsy	168	166	NR	61.6	334	266	34	34	8:1:1
Koike	2023	Japan	NR	spine	MR images	79		42	69.2	2125	1782	343	NR	8:2:NR
Huo	2023	China	lung (126)	spine, pelvis, limb, sternum, clavicle	pathologically	57	69	42	61.5	126	76	38	12	6:3:1
Hong	2021	Korea	prostate (32), lung (4), others (5)	abdominal	pathologically	112	129	71	67.6	241	177	64	NR	7:3:NR
Hinzpeter	2022	Switzerland	NR	thoracic and/or lumbar spine and/or pelvic bones	68 Ga-PSMA PET imaging	67		NR	71.0	410	328	82	NR	8:2:NR
Filograna	2019	Italy	Lung (3), prostatic (1), others (5)	NR	NR	8		37	NR	58	NR	NR	NR	NR
Duan	2023	China	lung (37), breast (8), other (22)	cervical vertebrae, thoracic vertebrae, lumbar vertebrae	pathological	67	54	33	54.5	121	74	29	18	8: external:2
Chen	2022	China	NSCLC (144)	chest	biopsy	144	51	19	60.5	195	NR	NR	NR	7:3:NR
Chang	2022	USA	NR	chest, abdomen, pelvis	follow-up or MRI	242		NR	NR	600	540	60	1104	9:1: external
Dong	2021	China	prostatic (40)	pelvis	PET/CT	40	88	NR	NR	128	79	NR	49	6:NR:4
Lee	2023	Korea	lung (59), breast (19), hepatocellular (19), renal cell (10), others (36)	chest, abdomen, and spine	pathologically	161	64	39	64.3	510	175	50	NR	7:3:NR
Kim	2024	Korean	NR	whole-spine	NR	322		50	63.5	11,419	242	60	20	8:2:external
Duan2	2024	China	Lung (68), breast (30), melanoma (11), prostatic (9), rectum (24), liver (10), kidney (11), thyroid (10)	cervical, thoracic, lumbar	pathologically	173		62	61.3	209	NR	NR	24	8:2:external
Park	2024	Switzerland.	Gastric (96)	iliac	pathologically	96		45	58.4	96	NR	NR	14	6:4:external
Zhang	2024	China	Prostatic (414)	pelvic	pathologically	106		105	0	73	211	169	NR	42

Table 2 Algorithm and data for the 20 included studies

Study	Year	Imaging Modality	Model Output	TP	FP	FN	TN
Xiong [56]	2021	MRI	ANN; RF; SVM; NB; KNN	108	10	5	55
Shi [47]	2023	CT	XGBoost	73	1	3	60
Özgül [57]	2023	CT	LR; KNN; NB; ANN; RF; VP; SVM; DT	87	8	5	72
Noguchi [18]	2022	CT	CNN	50	12	0	38
Liu [46]	2021	MRI	U-net	14	3	2	15
Koike [20]	2023	CT	InceptionV3	109	6	38	190
Huo [44]	2023	CT	DCNN	21	4	3	22
Hong [43]	2021	CT	RF	33	1	8	22
Hinzpeter [42]	2022	CT	Gradient-boosted tree	32	4	9	37
Duan1 [40]	2023	MRI	Resnet34; Resnet101; EfficientNet-B3; MVITV2	49	15	10	43
Chang [37]	2022	CT	DCNN	57	3	3	57
Dong [39]	2021	MRI	RF	13	8	3	25
Lee [45]	2023	CT	RF	21	3	14	12
Kim [50]	2024	MRI	U-net	150	39	25	490
Park [51]	2024	CT	RF	33	7	7	49
Zhang [53]	2024	MRI	SVM	28	1	10	3
Wang [52]	2023	MRI	CNN	/	/	/	/
Filograna [41]	2019	MRI	LR	/	/	/	/
Chen [38]	2022	CT	SVM	/	/	/	/
Duan2 [49]	2024	MRI	DL	/	/	/	/

differentiation between BM and multiple myeloma (MM), one (5%) study [43] compared BM with bone islands, and one (5%) study [40] differentiated spinal metastases from spinal tuberculosis. Two (9.52%) studies [38, 44] identified primary tumors as lung cancer, while 3 (15%) studies [39, 46, 53] identified prostate cancer. Four (20%) studies [20, 37, 42, 50]did not specify the primary tumor site. One (5%) study [51] reported gastric cancer as the primary tumor. The remaining 10 (50%) studies reported a variety of primary tumors, including those of the lung, breast, thyroid, prostate, and stomach. Only 7 (35%) studies [18, 40, 43–45, 49, 50]compared the performance of AI models with healthcare professionals on the same test set. Nine (45%) studies [39–41, 46, 49, 50, 52, 53, 56] employed AI techniques for diagnosing BM using MRI, while the remaining 11 (55%) studies [18, 20, 37, 38, 42–45, 47, 51, 57] used CT scans. Regarding model validation, 4 (20%) studies [37, 49, 50, 51] utilized out-of-sample data for external validation, and 7 (35%) studies [18, 39, 40, 44, 46, 53, 56] conducted internal validation. All reference standard forms were accepted for metastasis diagnosis, predominantly pathologic diagnosis, with a few studies using imaging tests such as PET-CT, MRI, SPECT, and follow-up imaging as the reference standard.

Quality assessment

Adherence to TRIPOD reporting standards was variable (Fig. 2a). Only three studies reported more than 70% of the items, and 3 studies reported less than 50% of items (Table S2). None of the studies reported on the following items: 3c, 11,14,18d, 19,27a, 27b. More than 50% of studies did not report on the following items: 8c,9b,12 g,13, 18E,18f,22. Only 3 studies had low bias and the rest had high bias according to PROBAST (Fig. 2b). The primary factors contributing to this assessment were the absence of external validation and the internal validation of models using small sample sizes (Table S3). Additionally, poor management of missing data was identified as the most prevalent risk of bias (item 4.4), which was not reported in any of the included studies.

Meta-analysis

Pooled detectability of AI performance in diagnosing tumor BM

We applied Method I to extract data from 80 contingency tables across 16 studies (refer to Table S4) for assessing AI accuracy for diagnosing tumor BM. These tables provided comprehensive details necessary for our analysis. We further analyzed data from these tables to generate SROC curves (Fig. 3a). Diagnostic accuracy improves as the summary estimate moves closer to the graph's upper left corner. The pooled SE was 0.87 (0.84–0.89), the pooled SP was 0.81 (0.76–0.84), and the pooled AUC was 0.91 (0.88–0.93). Forest plots of SE and SP are presented in Figure S1.

Using Method II, we obtained another ROC curve (Fig. 3b), showing a pooled SE of 0.88 (0.82–0.92), SP of 0.89 (0.84–0.93), AUC of 0.95 (0.92–0.96), PLR of 8.11 (5.57–11.80), NLR of 0.14 (0.09–0.21), and DOR of 58.64 (31.34–109.70) as shown in Figures S3-S4. However, this analysis indicated significant heterogeneity in both SE and SP, with I^2 values of 80.06 (70.90-89.22) and

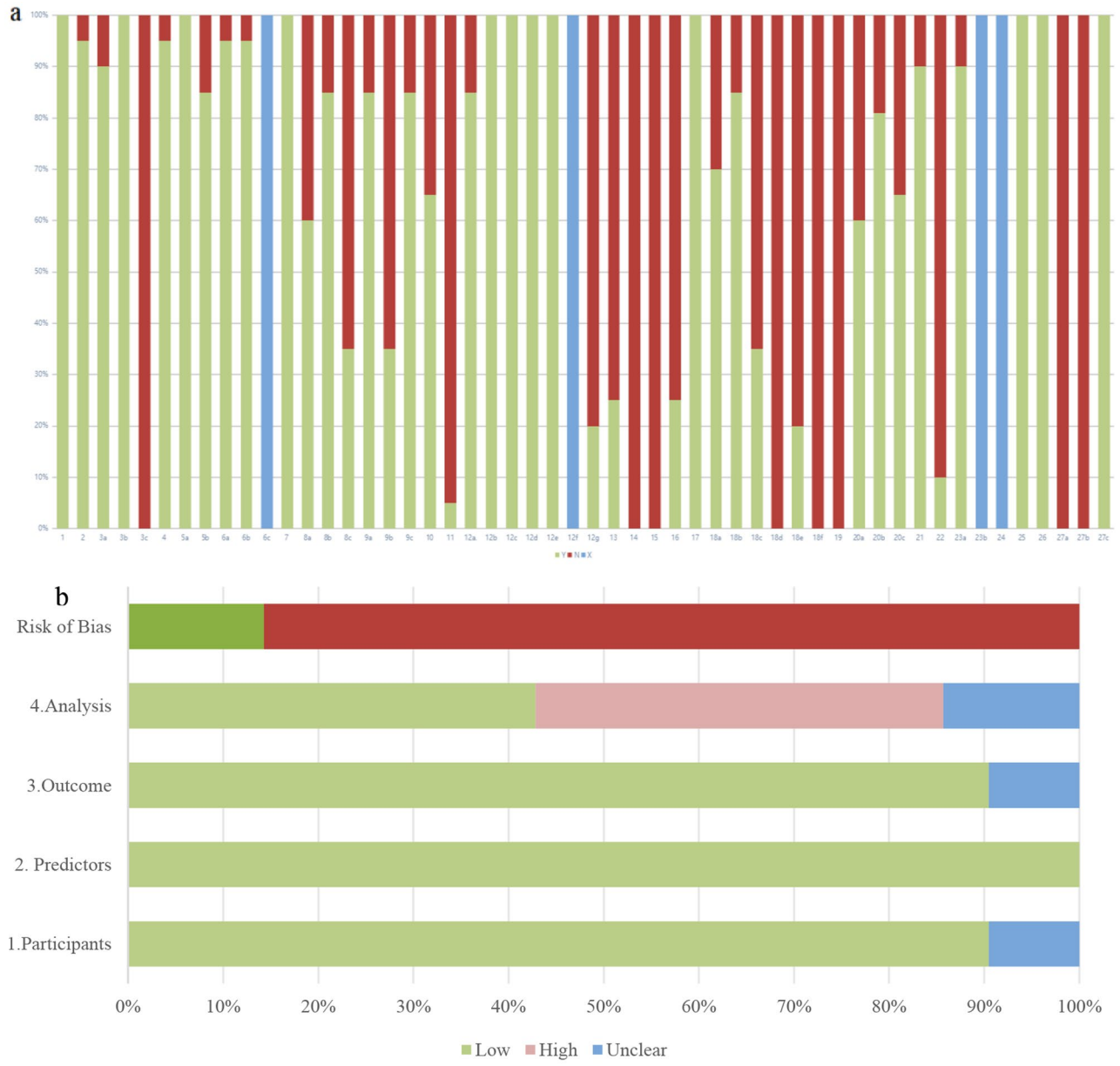

Fig. 2 Risk of bias assessment and quality assessment. **a**: Methodological quality assessment of the included studies using the TRIPOD-AI tool. Y, Yes for reported; N, Not reported; X, not applicable. **b**: Risk of Bias for each study using PROBAST. Green, low; blue, unclear; and red high risk of bias

75.71 (63.95–87.47) respectively, and a p-value < 0.01. To evaluate the clinical utility of AI, we constructed a Fagan nomogram (Figure S5). Assuming a 50% prevalence of BM, the Fagan plots illustrated that the posterior probability of having BM is 89% following a positive test result, and the probability of the absence of BM is 12% with a negative test result.

Pooled detectability of AI and healthcare professionals in diagnosing tumor BM

Seven studies compared the diagnostic performance of AI with that of healthcare professionals; however, only 6 (30%) [18, 40, 43, 44, 45, 50] provided sufficient data (TP, FP, FN, TN) to conduct a meta-analysis. In these studies, the diagnostic abilities of radiologists were assessed against those of AI models using the same datasets (Figure S6). The AI models demonstrated a pooled AUC of 0.92 (0.89–0.94), SE of 0.86 (0.72–0.94), and SP of 0.86 (0.77–0.92). In comparison, radiologists achieved a pooled AUC of 0.90 (0.88–0.93), SE of 0.83 (0.78–0.83), and SP of 0.88 (0.80–0.93). The findings indicate that the diagnostic performance of AI models was comparable to, or better than, that of radiologists.

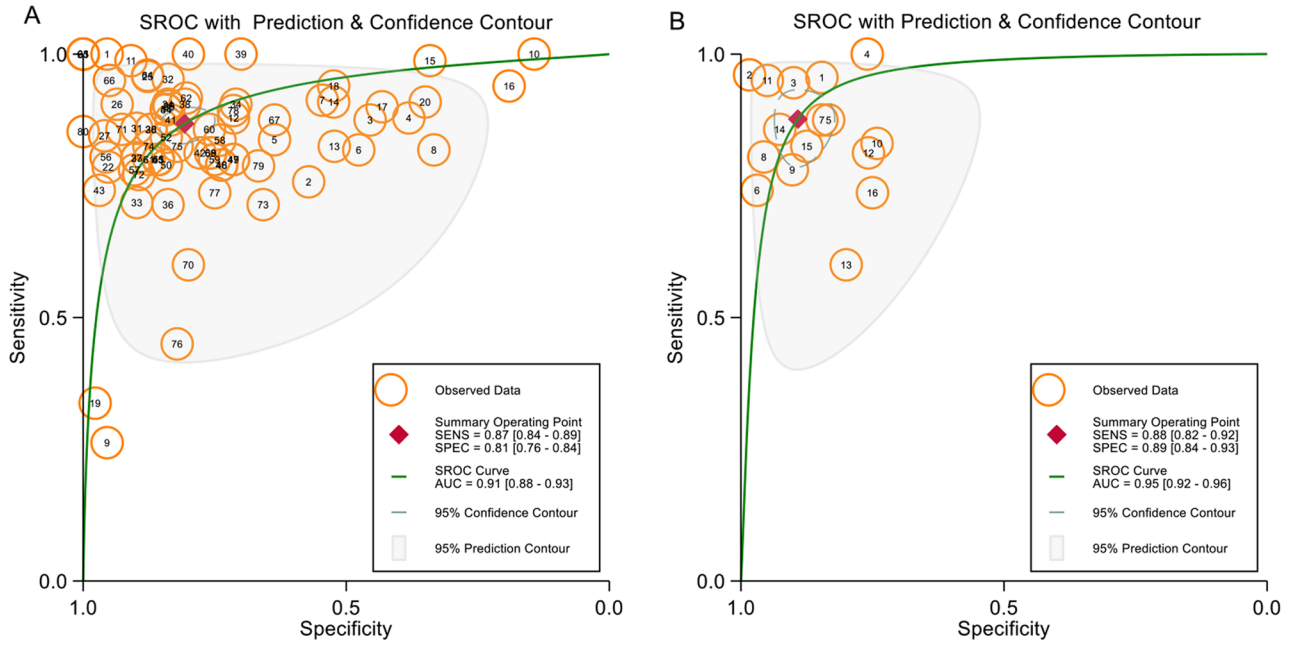

Fig. 3 ROC curves of all studies included in the meta-analysis (16 studies). **a**: ROC curves of all studies included in the meta-analysis (16 studies with 80 tables). **b**: ROC curves of studies when selecting contingency tables reporting the highest accuracy (16 studies with 16 tables)

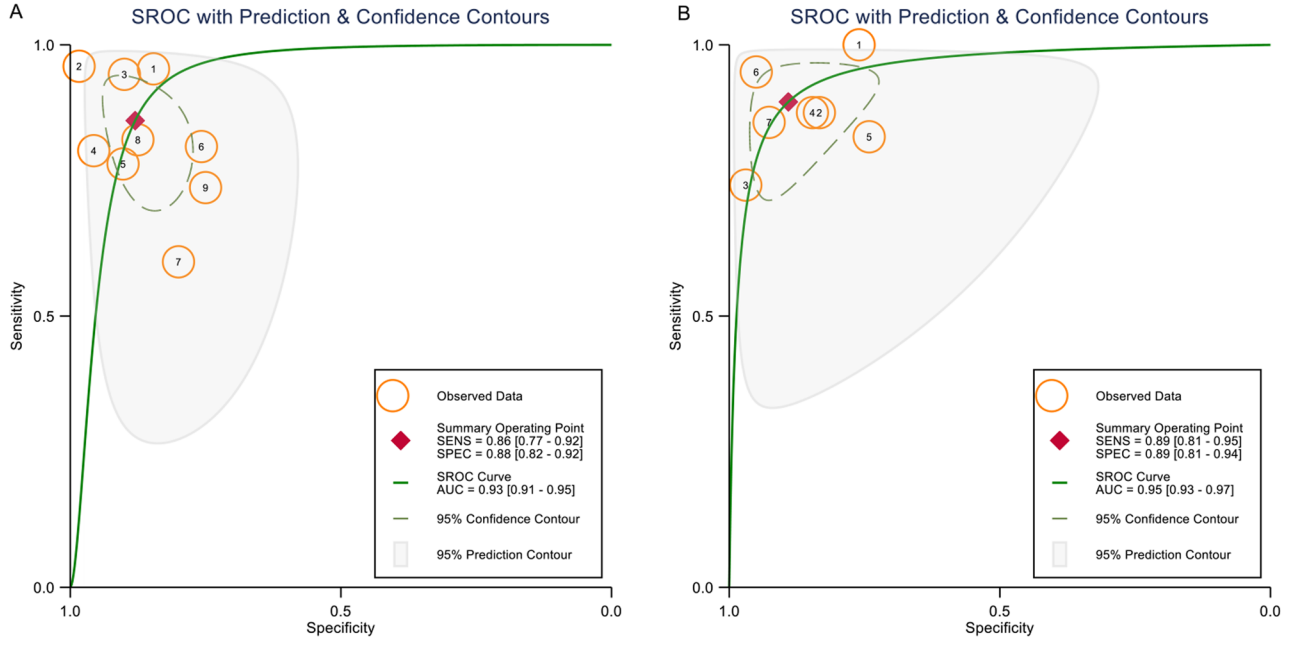

Fig. 4 ROC curves of subgroup analyses. **a**: ROC curves of ML algorithm (9 studies). **b**: ROC curves of DL algorithm (7 studies)

Pooled AI performance in diagnosing tumor BM using DL or ML

The meta-analysis results for different algorithms (DL or ML) are illustrated in Fig. 4. The DL algorithm achieved a pooled SE of 0.89 (0.81–0.95), with I^2 of 77.83%, and SP of 0.89 (0.81–0.94), with I^2 of 87.27%. The AUC was 0.95 (0.93–0.97). Forest plots for SE and SP are shown in Figure S7. The ML algorithm exhibited a pooled SE of 0.86 (0.77–0.92), with I^2 of 83.60%, and SP of 0.88 (0.86–0.92), with I^2 of 47.49%. The AUC was 0.93 (0.91–0.95), with corresponding forest plots displayed in Figure S8.

Pooled AI performance in diagnosing tumor BM using CT or MRI

The meta-analysis of different imaging modalities (CT and MRI) is depicted in Fig. 5. CT imaging showed a

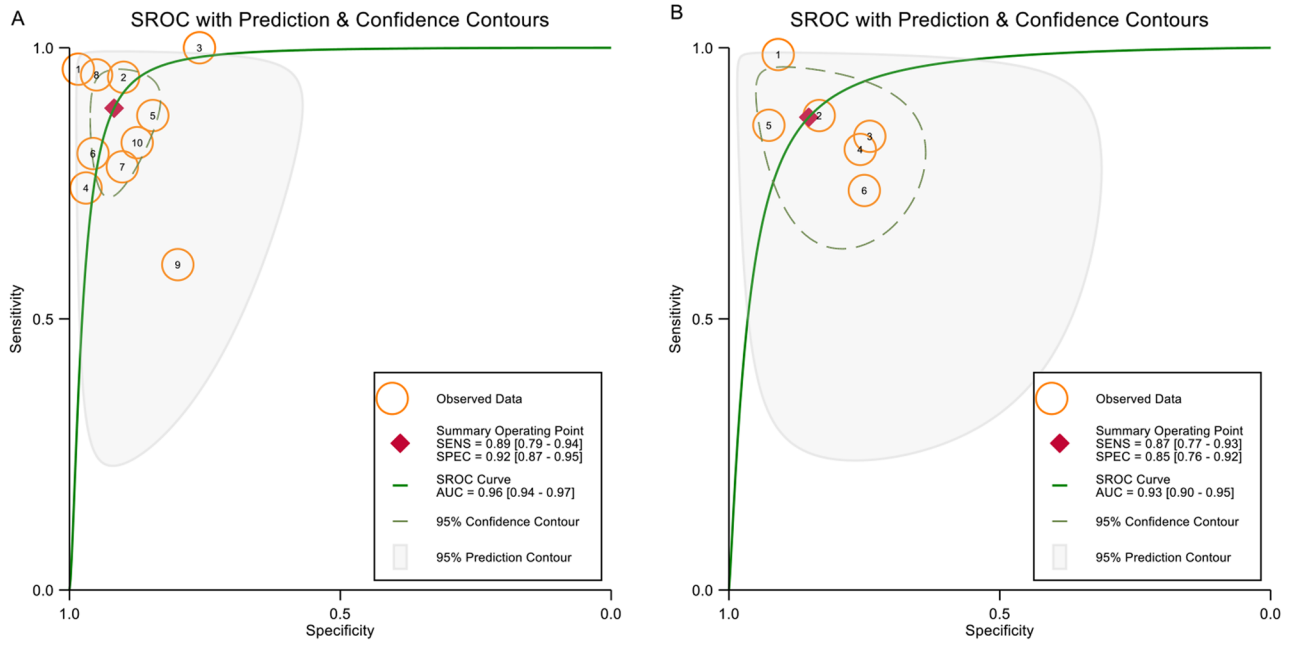

Fig. 5 ROC curves of subgroup analyses. **a**: ROC curves of CT imaging (10 studies). **b**: ROC curves of MRI imaging (6 studies)

pooled SE of 0.89 (0.79–0.94), with I^2 of 85.46%, and SP of 0.92 (0.87–0.95), with I^2 of 75.07%. The AUC was 0.96 (0.94–0.97), and forest plots for SE and SP are presented in Figure S9. MRI imaging reported a pooled SE of 0.87 (0.77–0.93), with I^2 of 68.90%, and SP of 0.85 (0.76–0.92), with I^2 of 81.43%. The AUC was 0.93 (0.90–0.95), with forest plots shown in Figure S10.

Meta-regression

The meta-regression analysis examined various covariates including sample size, number of images, age, imaging modalities, data enhancement, external validation, and algorithm type, as shown in Fig. 6. Notably, imaging modality and algorithm type were significant factors that contributed to reduced specificity ($p < 0.01$). Additionally, imaging modality was a significant factor in reducing SE ($p < 0.05$).

Subgroup analyses

The results of the subgroup analyses are presented in Table 3. For studies involving fewer than 100 images, the pooled SE was 0.84 (0.71–0.92) with an I^2 of 83.04%, SP was 0.95 (0.88–0.98) with an I^2 of 79.23%, and the AUC was 0.96 (0.94–0.98). For studies with more than 100 images, the pooled SE was 0.89 (0.81–0.93) with an I^2 of 79.50%, SP was 0.86 (0.81–0.90) with an I^2 of 73.13%, and the AUC was 0.93 (0.90–0.95). The ROC curves and forest plots for SE and SP are shown in Figure S11.

In terms of data augmentation, studies employing this technique reported a pooled SE of 0.80 (0.74–0.85) with an I^2 of 86.00%, and a pooled SP of 0.96 (0.93–0.97) with an I^2 of 35.10%. The ROC curves and forest plots for these measures are displayed in Figure S12. Studies without data augmentation reported a pooled SE of 0.88 (0.82–0.93) with an I^2 of 78.77%, SP of 0.87 (0.81–0.91) with an I^2 of 72.51%, and an AUC of 0.93 (0.91–0.95). The corresponding ROC curves and forest plots are shown in Figure S13.

Analyzing the data based on the number of patients resulted in a pooled SE of 0.86 (0.79–0.91) with an I^2 of 47.45%, SP of 0.82 (0.76–0.88) with an I^2 of 27.56%, and an AUC of 0.91 (0.88–0.93). Based on the number of lesions, the pooled SE was 0.90 (0.79–0.95) with an I^2 of 88.15%, SP of 0.92 (0.87–0.95) with an I^2 of 79.63%, and an AUC of 0.96 (0.94–0.98). The ROC curves and forest plots for these metrics are shown in Figure S14.

Subgroup analyses of multicenter and single-center studies revealed that multicenter studies had a pooled SE of 0.79 (0.70–0.87) with an I^2 of 77.08%, and a pooled SP of 0.87 (0.70–0.94) with an I^2 of 88.49%. Single-center studies reported a pooled SE of 0.90 (0.84–0.94) with an I^2 of 81.85%, a pooled SP of 0.90 (0.84–0.93) with an I^2 of 73.46%, and an AUC of 0.96 (0.93–0.97). The ROC curves and forest plots for these measurements are displayed in Figure S15.

Studies with and without external validation also showed differences. Studies with external validation had a pooled SE of 0.87 (0.83–0.91) with an I^2 of 61.4%, and a pooled SP of 0.92 (0.90–0.94) with an I^2 of 13.7%. The corresponding ROC curves and forest plots are shown in Figure S16. Studies without external validation had a pooled SE of 0.88 (0.80–0.93) with an I^2 of 82.49%, a

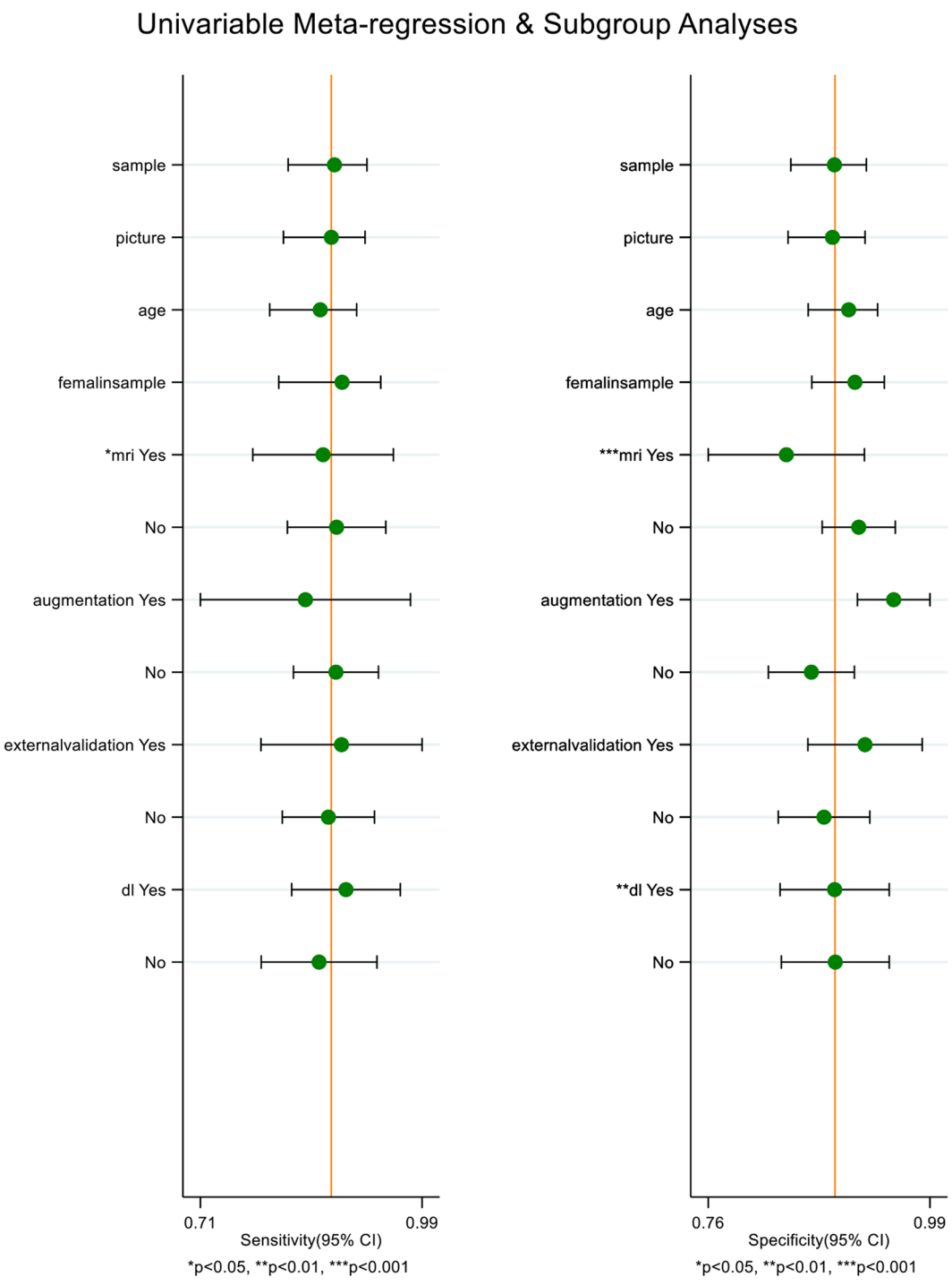

Fig. 6 The forest includes studies in the meta-regression

pooled SP of 0.88 (0.82–0.92) with an I^2 of 74.97%, and an AUC of 0.94 (0.92–0.96). The ROC curves and forest plots are displayed in Figure S17.

Finally, subgroup analysis based on different imaging modalities indicated that in CT-based studies, the pooled SE was 0.89 (0.79–0.94) with an I^2 of 85.46%, and SP was 0.92 (0.87–0.95) with an I^2 of 75.07%. In MRI-based studies, the pooled SE was 0.87 (0.77–0.93) with an I^2 of 68.90%, and SP was 0.85 (0.76–0.92) with an I^2 of 81.43%.

When we conducted the subgroup analysis, we observed several notable trends. First, single-center studies demonstrated higher SE, SP, and DOR compared to multi-center studies. Additionally, studies evaluating the number of lesions reported higher SE, SP, and DOR than

Table 3 Subgroup analyses

Subgroup	No. of trials	SE	I^2	SP	I^2	PLR	NLR	DOR
Picture								
< 100	4	0.84 (0.71, 0.92)	83.04	0.95 (0.88, 0.98)	79.23	16.80 (6.80, 41.49)	0.16 (0.08, 0.33)	102 (25, 410)
≥ 100	12	0.89 (0.81, 0.93)	79.50	0.86 (0.81, 0.90)	73.13	6.41 (4.41, 9.18)	0.13 (0.08, 0.22)	119 (38, 364)
Data augmentation								
With	3	0.80 (0.74, 0.85)	86.00	0.96 (0.93, 0.97)	35.10	15.75 (7.85, 31.61)	0.18 (0.09, 0.37)	93 (29, 293)
Without	13	0.88 (0.82, 0.93)	78.77	0.87 (0.81, 0.91)	72.51	6.64 (4.58, 9.62)	0.14 (0.08, 0.22)	119 (39, 364)
Data unit								
Based on number of patients	6	0.86 (0.79, 0.91)	47.45	0.82 (0.76, 0.88)	27.56	4.88 (3.34, 7.12)	0.18 (0.11, 0.27)	28 (13, 59)
Based on number of lesions	9	0.90 (0.79–0.95)	88.15	0.92 (0.87, 0.95)	79.63	11.70 (7.23, 18.95)	0.11 (0.05, 0.23)	104(45, 237)
Type of study								
Single-center	12	0.90 (0.84,0.94)	81.85	0.90 (0.84,0.93)	73.46	8.81 (5.77, 13.45)	0.05 (0.02, 0.11)	81 (41, 162)
Multicenter	4	0.79 (0.70,0.87)	77.08	087 (0.74,0.94)	88.49	5.94 (2.80, 12.59)	0.24(0.15, 0.38)	25 (8, 77)
External validation								
With	3	0.87 (0.83,0.91)	61.40	0.92 (0.90,0.94)	13.70	10.68 (6.84,16.67)	0.14 (0.08,0.24)	82 (30,224)
Without	13	0.88 (0.80,0.93)	82.49	0.88 (0.82,0.92)	74.97	7.42 (4.77,11.54)	0.14 (0.09,0.23)	52 (25,109)
Imaging modality								
CT	10	0.89 (0.79,0.94)	85.46	0.92 (0.87,0.95)	75.07	10.79 (6.66,17.48)	0.12 (0.06,0.23)	89 (38,204)
MRI	6	0.87 (0.77,0.93)	68.90	0.85 (0.76,0.92)	81.43	5.92 (3.33,10.92)	0.15 (0.08,0.30)	39 (13,120)

those assessing the number of patients. Likewise, CT-based studies exhibited greater SE, SP, and DOR levels than MRI-based studies.

We also discovered that studies employing data enhancement techniques exhibited lower SE and higher SP relative to those not using data enhancement. Notably, in studies utilizing data enhancement, the heterogeneity of specificity was significantly reduced to 35.10%. Furthermore, externally validated studies showed a higher DOR (82 vs. 53) and greater specificity (0.92 vs. 0.88) compared to non-validated studies. These validated studies also significantly reduced heterogeneity to almost negligible levels (I^2 = 13.70% vs. 74.97%).

Sensitivity analysis

The sensitivity analysis results of AI diagnostic performance are presented in Table S5. The analysis indicates that the impact of omitting any individual study from the overall estimation is relatively minor. Moreover, even when 3 studies [37, 50, 51] with low risk of bias were excluded, or one study with poor adherence to TRIPOD-AI guidelines was excluded, the results remained stable. Combining the data using two different methods, the overall results were largely unaffected, indicating the robustness of this meta-analysis.

Publication bias

The assessment using a funnel plot (Fig. 7) revealed no evidence of publication bias ($p = 0.29 > 0.05$).

Discussion

This systematic review and meta-analysis represents the first effort to evaluate the accuracy of AI in diagnosing tumor BM. The meta-analysis synthesized data from 16 studies, revealing a pooled SE of 0.88 (0.82, 0.92), SP of 0.89 (0.83, 0.93), and AUC of 0.95 (0.92, 0.96). The diagnostic performance between radiologists and AI, using the same datasets, was found to be comparable. Specifically, AI models demonstrated a pooled AUC of 0.92 (0.89–0.94), SE of 0.86 (0.72–0.94), and SP of 0.86 (0.77–0.92). Conversely, radiologists achieved a pooled AUC of 0.90 (0.88–0.93), SE of 0.83 (0.78–0.83), and SP of 0.88 (0.80–0.93). Despite AI's high diagnostic accuracy, this meta-analysis identified considerable heterogeneity among the included studies.

While the overall diagnostic performance of AI for diagnosing BM is promising and, in some measures, comparable to or even superior to that of radiologists, the adoption of this technology in clinical practice remains limited. A significant barrier is the "black box" nature of AI, which obscures the decision-making process, making it difficult for clinicians to fully understand and trust the AI-driven insights [58]. To mitigate this issue, efforts have been made to enhance transparency using techniques such as the CAM, which utilizes a globally averaged pooling layer to generate heat maps that visually explain AI decisions [59]. One popular approach, Gradient-weighted Class Activation Mapping (Grad-CAM), improves the interpretability of CNNs by emphasizing the regions of the input image that are most influential for predictions [60]. Grad-CAM has been applied across various domains, including image classification, object detection, and semantic segmentation, to integrate AI

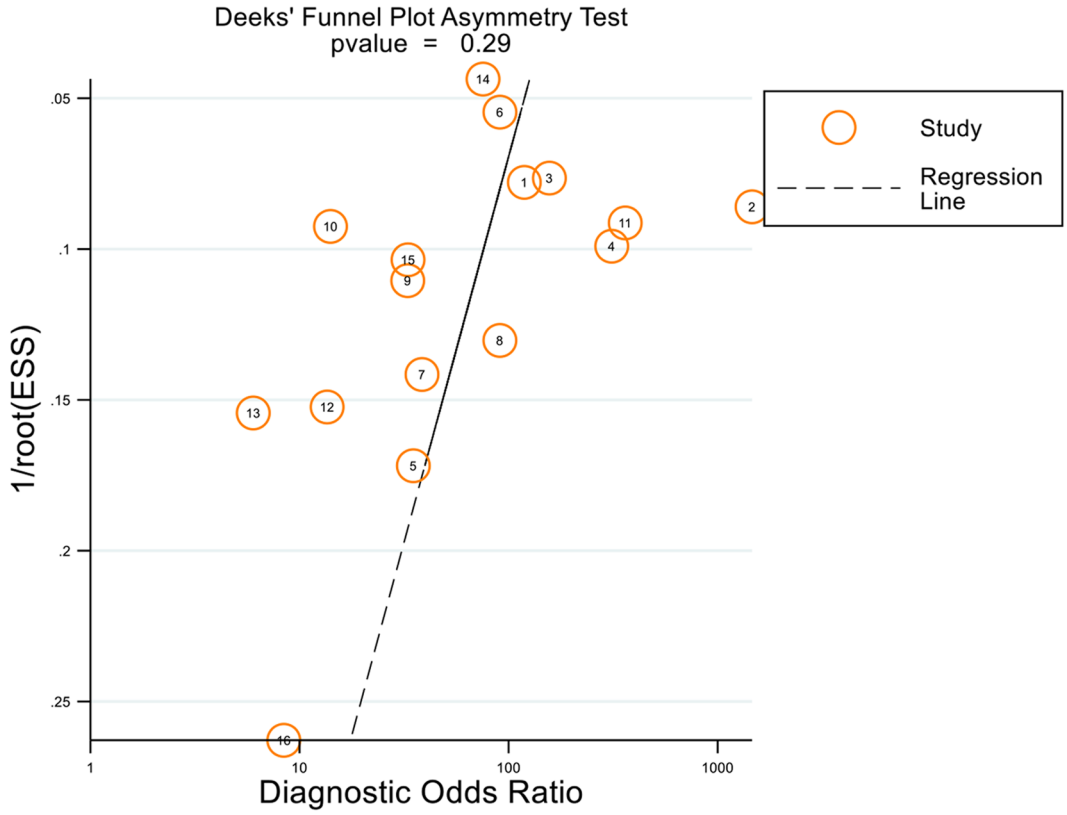

Fig. 7 Deeks' funnel plot

outputs with existing healthcare systems and allow clinicians to visually assess AI-driven predictions by highlighting suspected areas of BM (heat maps) [61]. To enable the systematic application of Grad-CAM, it is recommended that heat maps be generated after each model prediction and provided to the doctor along with the final diagnosis.

We conducted separate analyses of ML and DL algorithms and discovered that the SE, SP, and AUC were higher with the DL algorithm than with the ML algorithm. DL autonomously extracts discriminative features from input data using CNNs, which are composed of multiple layers of nonlinear functions [62]. Unlike traditional ML, which requires often problematic feature engineering, DL uses these layers to progressively extract higher-level features from the original inputs, thereby demonstrating superior performance [63]. The U-Net architecture, introduced in 2015, remains one of the most popular CNN architectures for medical image segmentation. In our study, the CNN architecture with the highest AUC was MVITV2, achieving 0.95 in the test set and 0.98 in the validation set. However, the limited number of studies included in our analysis restricted our ability to examine the more commonly used DL models (CNN, DNN) and ML models (ANN, KNN, DT, RF, VP, DT).

Although the study that used data augmentation showed a lower SE than the study without it, the results for SP and PLR were better. The application of DL in medical image analysis often encounters challenges such as insufficient training data and imbalanced classification [64, 65, 66]. To combat these issues, data augmentation is frequently used to increase the size and diversity of training sets in DL, which helps prevent overfitting due to limited data and ultimately improves performance on test sets. It is crucial for researchers dealing with small training datasets to effectively implement data augmentation to overcome these limitations [67].

Through subgroup analysis, we observed that single-center studies exhibited higher SE, SP, and DOR compared to multicenter studies, contradicting our initial hypothesis. Generally, using images from different healthcare institutions in multicenter studies increases dataset diversity, which enhances model generalization capability and yields more reliable outcomes [68]. Despite the use of data from various institutions in the multicenter studies included in our analysis, the small overall sample size likely contributed to the unexpected results. Additionally, studies that assessed the number of lesions rather than the number of patients showed higher SE, SP, and DOR, which can be attributed to the increase in sample size resulting from multiple lesions per patient.

We conducted a detailed subgroup analysis of studies based on whether they used external datasets for validation. The findings revealed that studies incorporating external datasets demonstrated a significant improvement in SE and SP compared to those that did not use such datasets. Additionally, heterogeneity was greatly diminished or even negligible in studies with external validation ($I^2 = 13.70\%$), indicating that external validation promotes uniformity in diagnostic criteria and procedures across studies, thereby stabilizing and enhancing the consistency of results. Furthermore, the DOR also improved significantly (82 vs. 52), further underscoring the crucial role of external validation in boosting the diagnostic performance of models. External validation is essential for assessing a model's generalizability [69]. Given that the objective of validation is to evaluate performance across diverse patient populations, it is feasible to gather new datasets from various centers [70]. Prospective validation mimics the actual clinical usage scenario, allowing for the identification of potential issues and challenges beforehand and providing a more solid foundation for the clinical application of the model. By integrating multicenter and prospective validation, we anticipate a substantial enhancement in the relevance and utility of our findings.

This study's findings suggest that the diagnostic efficacy of CT surpasses that of MRI, which is contrary to typical clinical expectations. Firstly, the studies included exhibit a high risk of bias, potentially compromising the reliability of the results. Secondly, the distribution of focus within these studies was uneven; 10 studies concentrated on CT-based diagnosis of BM, while only 6 focused on MRI-based diagnosis of BM. This disparity in sample sizes could result in an overestimation of CT's diagnostic capabilities. Additionally, adherence to the TRIPOD-AI guidelines was generally low among the included studies, which could introduce further biases in outcomes. The TRIPOD-AI statement provides a comprehensive reporting framework for research on AI-based diagnostic models. In order to improve the transparency and reproducibility of the research, future research should strictly follow the 27 items in the TRIPOD-AI guidelines. The most critical elements include: first, a clear description of the source and quality of the data, especially the steps of data processing. Second, the model development process is reported in detail, including feature selection, algorithm selection and the rationale behind it, as well as hyperparameter tuning and cross-validation methods. Third, provide a detailed description of model validation, including the use of internal and external validation datasets, and the reporting of performance metrics and results. In addition, the software and code used to develop and validate the model should be provided.

This study represents the first systematic review and meta-analysis aimed at evaluating the accuracy of AI in diagnosing tumor BM. It highlights several potential benefits. Firstly, recent research [42] has demonstrated AI's superiority in detecting small lesions that are often missed by radiologists using traditional methods. Secondly, AI's ability to efficiently process and analyze image data not only reduces diagnostic times but also provides rapid support for clinical decision-making [18]. Additionally, AI-assisted tools have lightened the workload of radiologists, enhancing their efficiency and enabling better management of their time and focus. Despite conservatively affirming the high accuracy of AI in diagnosing BM, the study acknowledges existing knowledge gaps. One significant issue is the high heterogeneity among the included studies, likely caused by variations in dataset sizes, types of imaging equipment, and AI algorithm choices. To address this, future research will focus on standardized data processing and quality control, establishing uniform protocols for image data acquisition, annotation, and preprocessing to ensure data consistency and comparability across different studies. Another limitation is the current focus of AI applications in BM diagnosis, which is predominantly at the imaging level. The integration of relevant clinical information, such as patient history and laboratory test results, remains insufficient. Going forward, efforts will be directed toward developing a multi-modal and multi-dimensional diagnostic model that deeply integrates image data with clinical data. This approach aims to provide a more comprehensive evaluation of patients' conditions and enhance overall diagnostic accuracy. Furthermore, the absence of large-scale, multi-center prospective validation studies restricts the broader clinical adoption of AI models. Future initiatives will involve multi-center collaborations to gather extensive data on BM from hospitals across various regions and levels of medical care and conduct rigorous external validations to assess the stability and generalizability of AI models in diverse clinical settings. With ongoing advancements and enhancements in deep learning algorithms, such as the expansion of Transformer architectures in medical imaging, AI models are expected to exhibit improved feature extraction and learning capabilities [71]. These developments will likely enhance the diagnostic accuracy, SE, and SP of AI applications in medical settings.

To the best of our knowledge, this is the first systematic review and meta-analysis to evaluate the accuracy of AI-based systems in diagnosing BM. However, this meta-analysis faced several limitations. Firstly, selection bias could not be fully eliminated, as indicated by the PROBAST tool assessment, due to the retrospective nature of all included studies. Secondly, the majority of the studies were conducted at single centers. Many experts in AI

research advocate for the adoption of continuous, multicenter study designs or external validation methods to enhance the clinical relevance and generalizability of the findings [67]. Thirdly, the lack of external validation is a concern, as only one included study utilized external validation, thus challenging the generalizability of the model. Additionally, some studies reported that both training and testing data were derived from the same cohort, raising concerns about potential overfitting of the AI systems [72]. Finally, although many manuscripts adhered to the TRIPOD-AI guidelines, they often omitted crucial details such as the training procedures, model tuning, and the size of the test sets.

Conclusion

The application of AI in the diagnosis of tumor-related BM shows substantial promise. This meta-analysis tentatively suggests that AI could perform comparably to medical professionals in diagnosing tumor BM via medical radiography, providing a foundation for its clinical use. The broader clinical implementation of AI could potentially address the shortage of medical resources, enhance the detection rates and accuracy of tumor BM diagnoses, and thereby improve patient outcomes. Nonetheless, it is important to recognize that further high-quality research is necessary to ensure that AI applications in healthcare can be effectively integrated into clinical practice and adhere to standardized research methodologies.

Abbreviations

AI	Artificial intelligence
BM	Bone metastasis
CNKI	China National Knowledge Infrastructure
VIP	Wei Pu database
CBM	China Biology Medicine
CT	Computed tomography
MRI	Magnetic resonance imaging
SE	Sensitivity
SP	Specificity
AUC	Area under the curve
PROBAST	Prediction model Risk of Bias Assessment Tool
TRIPOD-AI	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis-Artificial intelligence
PLR	Positive likelihood ratio
NLR	Negative likelihood ratio
DOR	Diagnostic odds ratio
SREs	Skeletal-related events
PET/CT	Positron emission tomography/computed tomography
MM	Multiple myeloma
DLA	Dl-based algorithm
ML	Machine learning
DL	Deep learning
PPV	Positive predictive value
NPV	Negative predictive value
TP	True positive
FP	False positive
FN	False negative
TN	True negative
CNN	Convolutional Neural Networks
DNN	Deep Neural Networks
SVM	Support Vector Machine
LR	Logistic Regression
ANN	Artificial Neural Network
KNN	K-Nearest Neighbor
DT	Decision Tree
NB	Naive Bayes
RF	Random Forest
VP	Voted Perceptron
sROC	Summary receiver operating characteristics curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13631-0 .

Supplementary Material

Acknowledgements

I would like to thank Professors ZS and YKH for their invaluable guidance and support throughout the research process. I also want to acknowledge my collaborators, who provided important insights and suggestions during data collection, analysis, and the writing of the paper.

Author contributions

In this meta-analysis, THM served as the principal investigator, responsible for the study design, methodological development, and paper writing; ZRR conducted the literature search and selection, and evaluated the quality of the included studies; ZZH was primarily responsible for data extraction and statistical analysis; WP interpreted the results and discussed their clinical significance; HX, ZS, and YKH collaborated on drafting and revising the manuscript to ensure fluency of language. All authors actively participated in the final review of the manuscript and collectively agreed on the accuracy and completeness of its content.

Funding

National Natural Science Foundation of China (NO: 82360358) and Gansu Provincial People's Hospital Intramural Research Fund Program (NO: 21GSSYB-20).

Data availability

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

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 18 November 2024 / Accepted: 3 February 2025
Published online: 18 February 2025

References

1. Lin Q, Li T, Cao C, Cao Y, Man Z, Wang H. Deep learning based automated diagnosis of bone metastases with SPECT thoracic bone images. Sci Rep. 2021;11(1). https://doi.org/10.1038/s41598-021-83083-6.
2. Huang J-F, Shen J, Li X, Rengan R, Silvestris N, Wang M, et al. Incidence of patients with bone metastases at diagnosis of solid tumors in adults: a large population-based study. Annals Translational Med. 2020;8(7):482–482.
3. Shih J-T, Yeh T-T, Wang S-H, Shen P-H, Wang C-C, Chien W-C, et al. Incidence of bone metastases in patients with organ-specific cancers: a nationwide population-based cohort study. Int J Clin Pract. 2021;75(5):e13997–13997.

4. Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. Clin Cancer Res. 2006;12(20 Pt 2):6243s-6249s.https://doi.org/10.1158/1078-0432.Ccr-06-0931
5. Knapp BJ, Devarakonda S, Govindan R. Bone metastases in non-small cell lung cancer: a narrative review. J Thorac Dis. 2022;14(5):1696–712. https://doi.org/10.21037/jtd-21-1502.
6. Kosteva J, Langer C. The changing landscape of the medical management of skeletal metastases in nonsmall cell lung cancer. Curr Opin Oncol. 2008;20(2):155–61. https://doi.org/10.1097/CCO.0b013e3282f54cf2.
7. Cook RJ, Major P. Methodology for treatment evaluation in patients with cancer metastatic to bone. J Natl Cancer Inst. 2001;93(7):534–8. https://doi.org/10.1093/jnci/93.7.534.
8. Macedo F, Ladeira K, Pinho F, Saraiva N, Bonito N, Pinto L, et al. Bone metastases: an overview. Oncol Rev. 2017;11(1):321. https://doi.org/10.4081/oncol.2017.321.
9. Roodman GD. Mechanisms of disease: mechanisms of bone metastasis. New Engl J Med. 2004;350(16):1655–64. https://doi.org/10.1056/NEJMra030831.
10. Coleman RE, Croucher PI, Padhani AR, Clézardin P, Chow E, Fallon M, et al. Bone metastases. Nat Rev Dis Primers. 2020;6(1):83. https://doi.org/10.1038/s41572-020-00216-3.
11. Mollica V, Rizzo A, Rosellini M, Marchetti A, Ricci AD, Cimadamore A, et al. Bone Targeting agents in patients with metastatic prostate Cancer: state of the art. Cancers (Basel). 2021;13(3). https://doi.org/10.3390/cancers13030546.
12. Guven DC, Erul E, Kaygusuz Y, Akagunduz B, Kilickap S, De Luca R, et al. Immune checkpoint inhibitor-related hearing loss: a systematic review and analysis of individual patient data. Support Care Cancer. 2023;31(12):624. https://doi.org/10.1007/s00520-023-08083-w.
13. Rizzo A, Santoni M, Mollica V, Logullo F, Rosellini M, Marchetti A, et al. Peripheral neuropathy and headache in cancer patients treated with immunotherapy and immuno-oncology combinations: the MOUSEION-02 study. Expert Opin Drug Metab Toxicol. 2021;17(12):1455–66. https://doi.org/10.1080/17425255.2021.2029405.
14. Sahin TK, Ayasun R, Rizzo A, Guven DC. Prognostic Value of Neutrophil-to-Eosinophil ratio (NER) in Cancer: a systematic review and Meta-analysis. Cancers (Basel). 2024;16(21). https://doi.org/10.3390/cancers16213689.
15. Rizzo A, Santoni M, Mollica V, Fiorentino M, Brandi G, Massari F. Microbiota and prostate cancer. Semin Cancer Biol. 2022;86(Pt 3):1058–65. https://doi.org/10.1016/j.semcancer.2021.09.007.
16. Schulman KL, Kohles J. Economic burden of metastatic bone disease in the U.S. Cancer. 2007;109(11):2334–42. https://doi.org/10.1002/cncr.22678.
17. Svendsen ML, Gammelager H, Sværke C, Yong M, Chia VM, Christiansen CF, et al. Hospital visits among women with skeletal-related events secondary to breast cancer and bone metastases: a nationwide population-based cohort study in Denmark. Clin Epidemiol. 2013;5:97–103. https://doi.org/10.2147/clep.S42325.
18. Noguchi S, Nishio M, Sakamoto R, Yakami M, Fujimoto K, Emoto Y, et al. Deep learning-based algorithm improved radiologists' performance in bone metastases detection on CT. Eur Radiol. 2022;32(11):7976–87. https://doi.org/10.1007/s00330-022-08741-3.
19. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial Intelligence in fracture detection: a systematic review and Meta-analysis. Radiology. 2022;304(1):50–62. https://doi.org/10.1148/radiol.211785.
20. Koike Y, Yui M, Nakamura S, Yoshida A, Takegawa H, Anetai Y, et al. Artificial intelligence-aided lytic spinal bone metastasis classification on CT scans. Int J Comput Assist Radiol Surg. 2023. https://doi.org/10.1007/s11548-023-02880-8.
21. Noguchi S, Nishio M, Sakamoto R, Yakami M, Fujimoto K, Emoto Y, et al. Deep learning-based algorithm improved radiologists' performance in bone metastases detection on CT. Eur Radiol. 2022;32(11):7976–87. https://doi.org/10.1007/s00330-022-08741-3.
22. Xiong X, Wang J, Hu S, Dai Y, Zhang Y, Hu C. Differentiating between multiple myeloma and metastasis subtypes of lumbar vertebra lesions using machine learning-based Radiomics. Front Oncol. 2021;11:601699. https://doi.org/10.3389/fonc.2021.601699.
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71.
24. Ge L, Tian JH, Li YN, Pan JX, Li G, Wei D, et al. Association between prospective registration and overall reporting and methodological quality of systematic reviews: a meta-epidemiological study. J Clin Epidemiol. 2018;93:45–55. https://doi.org/10.1016/j.jclinepi.2017.10.012.
25. Wang X, Chen Y, Yao L, Zhou Q, Wu Q, Estill J, et al. Reporting of declarations and conflicts of interest in WHO guidelines can be further improved. J Clin Epidemiol. 2018;98:1–8. https://doi.org/10.1016/j.jclinepi.2017.12.021.
26. Haining L, Hao W, Ningping Z, Yu L. Methods of data extraction in meta-analysis of diagnostic accuracy study. Chin J Evidence-Based Med. 2018;18(9). https://doi.org/10.7507/1672-2531.201805153.
27. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to assess risk of Bias and Applicability of Prediction Model studies: explanation and elaboration. Ann Intern Med. 2019;170(1):W1–33. https://doi.org/10.7326/m18-1377.
28. 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. https://doi.org/10.1136/bmjopen-2020-048008.
29. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. https://doi.org/10.1136/bmj.g7594.
30. Li Z, Li Y, Li N, Shen L. Positron emission tomography/computed tomography outperforms MRI in the diagnosis of local recurrence and residue of nasopharyngeal carcinoma: an update evidence from 44 studies. Cancer Med. 2019;8(1):67–79. https://doi.org/10.1002/cam4.1882.
31. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20(19):2865–84. https://doi.org/10.1002/sim.942.
32. Fu J, Li L, Wang X, Zhang M, Zhang Y, Li Z. Clinical utility of arterial spin labeling for preoperative grading of glioma. Biosci Rep. 2018;38(4). https://doi.org/10.1042/bsr20180507.
33. Fu J, Li Y, Li N, Li Z. Comprehensive analysis of clinical utility of three-dimensional ultrasound for benign and malignant breast masses. Cancer Manag Res. 2018;10:3295–303. https://doi.org/10.2147/cmar.S176494.
34. Fu J, Li Y, Li Z, Li N. Clinical utility of decarboxylation prothrombin combined with α-fetoprotein for diagnosing primary hepatocellular carcinoma. Biosci Rep. 2018;38(5). https://doi.org/10.1042/bsr20180044.
35. Zhang G, Li Y, Li C, Li N, Li Z, Zhou Q. Assessment on clinical value of prostate health index in the diagnosis of prostate cancer. Cancer Med. 2019;8(11):5089–96. https://doi.org/10.1002/cam4.2376.
36. Lin L, Chu H. Quantifying publication bias in meta-analysis. Biometrics. 2018;74(3):785–94. https://doi.org/10.1111/biom.12817.
37. Chang CY, Buckless C, Yeh KJ, Torriani M. Automated detection and segmentation of sclerotic spinal lesions on body CTs using a deep convolutional neural network. Skeletal Radiol. 2022;51(2):391–9. https://doi.org/10.1007/s00256-021-03873-x.
38. Chen L, Yu L, Li X, Tian Z, Lin X. Value of CT Radiomics and clinical features in Predicting Bone metastases in patients with NSCLC. Contrast Media Mol Imaging. 2022;2022. https://doi.org/10.1155/2022/7642511.
39. Dong H, Guangming LU. The predictive value of bp-MRI radiomics model in the first diagnosis of bone metastasis for prostate cancer. J Practical Radiol. 2021;37(6):968–71. https://doi.org/10.3969/j.issn.1002-1671.2021.06.023.
40. Duan S, Dong W, Hua Y, Zheng Y, Ren Z, Cao G, et al. Accurate differentiation of spinal tuberculosis and spinal metastases using MR-Based deep learning algorithms. Infect Drug Resist. 2023;16:4325–34. https://doi.org/10.2147/IDR.S417663.
41. Filograna L, Lenkowicz J, Cellini F, Dinapoli N, Manfrida S, Magarelli N, et al. Identification of the most significant magnetic resonance imaging (MRI) radiomic features in oncological patients with vertebral bone marrow metastatic disease: a feasibility study. Radiologia Med. 2019;124(1):50–7. https://doi.org/10.1007/s11547-018-0935-y.
42. Hinzpeter R, Baumann L, Guggenberger R, Huellner M, Alkadhi H, Baessler B. Radiomics for detecting prostate cancer bone metastases invisible in CT: a proof-of-concept study. Eur Radiol. 2022;32(3):1823–32. https://doi.org/10.1007/s00330-021-08245-6.
43. Hong JH, Jung J-Y, Jo A, Nam Y, Pak S, Lee S-Y, et al. Development and validation of a Radiomics Model for differentiating Bone islands and Osteoblastic Bone metastases at Abdominal CT. Radiology. 2021;299(3):626–32. https://doi.org/10.1148/radiol.2021203783.
44. Huo T, Xie Y, Fang Y, Wang Z, Liu P, Duan Y, et al. Deep learning-based algorithm improves radiologists' performance in lung cancer bone metastases detection on computed tomography. Front Oncol. 2023;13:1125637. https://doi.org/10.3389/fonc.2023.1125637.

45. Lee S, Lee SY, Kim S, Huh YJ, Lee J, Lee KE, et al. Differentiating multiple myeloma and osteolytic bone metastases on contrast-enhanced computed Tomography scans: the feasibility of Radiomics Analysis. Diagnostics (Basel). 2023;13(4). https://doi.org/10.3390/diagnostics13040755.
46. Liu X, Han C, Cui Y, Xie T, Zhang X, Wang X. Detection and segmentation of pelvic bones metastases in MRI images for patients with prostate Cancer based on deep learning. Front Oncol. 2021;11:773299. https://doi.org/10.3389/fonc.2021.773299.
47. Shi J, Huang H, Xu S, Du L, Zeng X, Cao Y, et al. XGBoost-based multiparameters from dual-energy computed tomography for the differentiation of multiple myeloma of the spine from vertebral osteolytic metastases. Eur Radiol. 2023;33(7):4801–11. https://doi.org/10.1007/s00330-023-09404-7.
48. !!!. INVALID CITATION!!!.
49. Duan S, Cao G, Hua Y, Hu J, Zheng Y, Wu F, et al. Identification of origin for spinal metastases from MR images: comparison between Radiomics and Deep Learning methods. World Neurosurg. 2023;175:E823–31. https://doi.org/10.1016/j.wneu.2023.04.029.
50. Kim DH, Seo J, Lee JH, Jeon E-T, Jeong D, Chae HD, et al. Automated detection and segmentation of bone metastases on spine MRI using U-Net A Multicenter. Korean J Radiol. 2024;25(4):363–73. https://doi.org/10.3348/kjr.2023.0671.
51. Park J, Jung M, Kim SK, Lee YH. Prediction of bone marrow metastases using computed tomography (CT) Radiomics in patients with gastric Cancer: uncovering invisible metastases. Diagnostics 2024; 14.
52. Wang D, Sun Y, Tang X, Liu C, Liu R. Deep learning-based magnetic resonance imaging of the spine in the diagnosis and physiological evaluation of spinal metastases. J Bone Oncol. 2023;40:100483. https://doi.org/10.1016/j.jbo.2023.100483.
53. Zhang Y-F, Zhou C, Guo S, Wang C, Yang J, Yang Z-J, et al. Deep learning algorithm-based multimodal MRI radiomics and pathomics data improve prediction of bone metastases in primary prostate cancer. J Cancer Res Clin Oncol. 2024;150(2). https://doi.org/10.1007/s00432-023-05574-5.
54. Shi J, Huang H, Xu S, Du L, Zeng X, Cao Y, et al. XGBoost-based multiparameters from dual-energy computed tomography for the differentiation of multiple myeloma of the spine from vertebral osteolytic metastases. Eur Radiol. 2023;33(7):4801–11. https://doi.org/10.1007/s00330-023-09404-7.
55. Ozgul HA, Akin IB, Mutlu U, Balci A. Diagnostic value of machine learning-based computed tomography texture analysis for differentiating multiple myeloma from osteolytic metastatic bone lesions in the peripheral skeleton. Skeletal Radiol. 2023. https://doi.org/10.1007/s00256-023-04333-4.
56. Xiong X, Wang J, Hu S, Dai Y, Zhang Y, Hu C. Differentiating between multiple myeloma and metastasis subtypes of lumbar vertebra lesions using machine learning-based Radiomics. Front Oncol. 2021;11:601699. https://doi.org/10.3389/fonc.2021.601699.
57. Ozgul HA, Akin IB, Mutlu U, Balci A. Diagnostic value of machine learning-based computed tomography texture analysis for differentiating multiple myeloma from osteolytic metastatic bone lesions in the peripheral skeleton. Skeletal Radiol. 2023;52:1703–11.
58. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, et al. Artificial Intelligence and Machine Learning in Radiology: opportunities, challenges, pitfalls, and Criteria for Success. J Am Coll Radiol. 2018;15(3 Pt B):504–8. https://doi.org/10.1016/j.jacr.2017.12.026.
59. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 2921–9.
60. Abas Mohamed Y, Ee Khoo B, Shahrimie Mohd Asaari M, Ezane Aziz M, Rahiman Ghazali F. Decoding the black box: explainable AI (XAI) for cancer diagnosis, prognosis, and treatment planning-A state-of-the art systematic review. Int J Med Inf. 2025;193:105689. https://doi.org/10.1016/j.ijmedinf.2024.105689.
61. Livieris IE, Pintelas E, Kiriakidou N, Pintelas P. Explainable image similarity: integrating siamese networks and Grad-CAM. J Imaging. 2023;9(10). https://doi.org/10.3390/jimaging9100224.
62. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, et al. AI applications to medical images: from machine learning to deep learning. Phys Med. 2021;83:9–24. https://doi.org/10.1016/j.ejmp.2021.02.006.
63. Cui S, Tseng HH, Pakela J, Ten Haken RK, El Naqa I. Introduction to machine and deep learning for medical physicists. Med Phys. 2020;47(5):e127–47. https://doi.org/10.1002/mp.14140.
64. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol. 2021;65(5):545–63. https://doi.org/10.1111/1754-9485.13261.
65. Khalifa NE, Loey M, Mirjalili S. A comprehensive survey of recent trends in deep learning for digital images augmentation. Artif Intell Rev. 2022;55(3):2351–77. https://doi.org/10.1007/s10462-021-10066-4.
66. Sorin V, Barash Y, Konen E, Klang E. Creating Artificial images for Radiology Applications using Generative Adversarial Networks (GANs) - a systematic review. Acad Radiol. 2020;27(8):1175–85. https://doi.org/10.1016/j.acra.2019.12.024.
67. Zhang X, Yang Y, Shen YW, Zhang KR, Jiang ZK, Ma LT, et al. Diagnostic accuracy and potential covariates of artificial intelligence for diagnosing orthopedic fractures: a systematic literature review and meta-analysis. Eur Radiol. 2022;32(10):7196–216. https://doi.org/10.1007/s00330-022-08956-4.
68. Beyaz S, Açıcı K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg. 2020;31(2):175–83. https://doi.org/10.5606/ehc.2020.72163.
69. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the external validation for checking learned Model Interpretability and Generalizability. Patterns (N Y). 2020;1(8):100129. https://doi.org/10.1016/j.patter.2020.100129.
70. Zheng Q, Yang L, Zeng B, Li J, Guo K, Liang Y, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: a systematic review and meta-analysis. EClinicalMedicine. 2021;31:100669. https://doi.org/10.1016/j.eclinm.2020.100669.
71. Liu Z, Lv Q, Yang Z, Li Y, Lee CH, Shen L. Recent progress in transformer-based medical image analysis. Comput Biol Med. 2023;164:107268. https://doi.org/10.1016/j.compbiomed.2023.107268.
72. Gao L, Jiao T, Feng Q, Wang W. Application of artificial intelligence in diagnosis of osteoporosis using medical images: a systematic review and meta-analysis. Osteoporos Int. 2021;32(7):1279–86. https://doi.org/10.1007/s00198-021-05887-6.

Publisher's note