Journal of Imaging Informatics in Medicine (2024) 37:766–777
https://doi.org/10.1007/s10278-023-00945-3

Diagnostic Performance of Artificial Intelligence in Detection of Primary Malignant Bone Tumors: a Meta-Analysis

Mohammad Amin Salehi[1] · Soheil Mohammadi[1] · Hamid Harandi[1] · Seyed Sina Zakavi[2] · Ali Jahanshahi[3] · Mohammad Shahrabi Farahani[4] · Jim S. Wu[5]

Received: 1 August 2023 / Revised: 4 October 2023 / Accepted: 12 October 2023 / Published online: 12 January 2024

Abstract

We aim to conduct a meta-analysis on studies that evaluated the diagnostic performance of artificial intelligence (AI) algorithms in the detection of primary bone tumors, distinguishing them from other bone lesions, and comparing them with clinician assessment. A systematic search was conducted using a combination of keywords related to bone tumors and AI. After extracting contingency tables from all included studies, we performed a meta-analysis using random-effects model to determine the pooled sensitivity and specificity, accompanied by their respective 95% confidence intervals (CI). Quality assessment was evaluated using a modified version of Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) and Prediction Model Study Risk of Bias Assessment Tool (PROBAST). The pooled sensitivities for AI algorithms and clinicians on internal validation test sets for detecting bone neoplasms were 84% (95% CI: 79.88) and 76% (95% CI: 64.85), and pooled specificities were 86% (95% CI: 81.90) and 64% (95% CI: 55.72), respectively. At external validation, the pooled sensitivity and specificity for AI algorithms were 84% (95% CI: 75.90) and 91% (95% CI: 83.96), respectively. The same numbers for clinicians were 85% (95% CI: 73.92) and 94% (95% CI: 89.97), respectively. The sensitivity and specificity for clinicians with AI assistance were 95% (95% CI: 86.98) and 57% (95% CI: 48.66). Caution is needed when interpreting findings due to potential limitations. Further research is needed to bridge this gap in scientific understanding and promote effective implementation for medical practice advancement.

Keywords Bone tumor · Artificial intelligence · Deep learning · Machine learning · Bone malignancies

Introduction

Although relatively uncommon, bone tumors are the third most common cause of death for cancer patients under the age of 20 [1, 2]. It is essential to perform an accurate assessment of the presence and extent of bone tumors, both primary and metastatic, in order to perform the proper staging and treatment of these conditions. The accurate diagnosis of bone tumors hinges upon a comprehensive evaluation of three essential factors: clinical presentation, imaging appearance, and detailed histopathologic assessment [3]. Detecting bone abnormalities can be accomplished through a variety of imaging techniques, such as magnetic resonance imaging (MRI), positron emission tomography (PET) scan, X-ray, scintigraphy, and computed tomography (CT) scans [4]. However, MRI is considered the modality of choice for local staging of bone tumors [5]. Traditional methods, while still a valuable tool for initial bone tumor diagnosis, face challenges in accurately assessing lesions within complex anatomical regions [5].

Mohammad Amin Salehi and Soheil Mohammadi contributed equally.

✉ Soheil Mohammadi
soheil.mhm@gmail.com

1 School of Medicine, Tehran University of Medical Sciences, Pour Sina St, Keshavarz Blvd, Tehran 1417613151, Iran

2 School of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran

3 School of Medicine, Guilan University of Medical Sciences, Rasht, Iran

4 Medical Students Research Committee, Shahed University, Tehran, Iran

5 Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, MA 02215, USA

It has been shown that artificial intelligence (AI) algorithms, particularly deep learning, have made great strides in image recognition [6]. Radiological images are no exception to this role. In the realm of machine learning, deep learning is a specialized subset characterized by the integration of three or more neural network, layers that mimic the functions of human neurons [7].

The application of AI in radiology dates back to 1992 when it was first utilized for the detection of microcalcifications in mammography. However, in recent times, there has been a surge in interest and attention towards it [8].

The use of AI in radiology has been subject to considerable discussion over the past few years, with a focus on its advantages and disadvantages. Despite the potential benefits of incorporating AI into radiology practice, radiologists must take certain barriers into consideration. These include ethical dilemmas, privacy concerns, regulatory compliance, and the need for proper training of AI systems, and safe integration of AI technologies into routine clinical workflows [9]. Additionally, the accuracy and precision of implementing the system in clinical practice needs to be carefully evaluated. This includes assessing its reliability and potential limitations to ensure appropriate utilization in real-world scenarios.

In the domain of AI applications for bone tumor detection, there has been a growing interest in recent years. This surge in interest is driven by the potential advantages AI offers in improving the accuracy and efficiency of bone tumor diagnosis [10]. However, to effectively harness the power of AI in this specialized field, it is crucial to recognize the current state of AI approaches, their associated results, and the existing gaps and limitations. These aspects are pivotal in justifying the need for a comprehensive meta-analysis.

In this study, we aim to conduct a meta-analysis on studies that evaluated the diagnostic performance of AI algorithms in the detection of primary bone tumors, distinguishing them from other bone lesions, and comparing them with clinician assessment. Furthermore, we attempted to compare the diagnostic precision of AI algorithms with that of experienced radiologists.

Materials and Methods

Protocol and Registration

The authors submitted this review to the International Prospective Register of Systematic Reviews (PROSPERO) website (CRD42022321526).

Search Strategy and Study Selection

To find articles related to AI algorithms in bone tumor diagnosis, we searched PubMed, Scopus, Web of Science core collection, CINAHL, EBSCO, IEEE, Medline, and ACM digital library. The initial search was conducted in March 2022 and was updated in January 2023. We used a combination of keywords related to bone tumors and AI (Table E3-E7). We included original studies investigating the diagnostic performance of artificial intelligence algorithms in detection of bone tumors in comparison of a control group (consisting of healthy controls, patients with benign bone lesions, or metastatic bone lesions). Language, publication time, and age of group participants are not limited. We excluded studies that met the following criteria: non-original studies, studies on metastasis outcome, and studies that were not written in English. Our meta-analysis exclusively incorporated studies providing the complete set of four numbers in a contingency table.

Data Extraction

We extracted the following variables in each study: author and publication year; country of study; the number of participants for each trait; imaging modality; utilized algorithm and architecture; imaging dimensions; evaluation metrics; level; training size; type of validation and validation size; imaging view; the number of TN, FP, TP, FN; sensitivity and specificity; positive predictive value, the area under the curve (A.U.C.), and accuracy. We created a contingency table for each study from the nominal data provided. To guarantee the precision of the data for analysis, two reviewers independently assessed each table.

Quality Assessment

We employed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist to assess adherence to reporting guidelines in the studies under review. This checklist consists of 22 recommendations that aim to promote transparent reporting of research that involves the development and/or validation of prediction models [11]. However, since not all items on this checklist were relevant for AI studies (e.g., reporting follow-up time in diagnostic accuracy studies), we used a modified version of TRIPOD (Table E6). To evaluate bias and applicability, we utilized the Prediction Model Study Risk of Bias Assessment Tool (PROBAST) checklist, which is designed to assess papers based on four key areas: participants, predictors, outcomes, and analysis (Table E7) [12]. We evaluated the training and testing images for bias and applicability in the first area.

Statistical Analysis

We performed a meta-analysis of studies to assess the diagnostic performance of AI algorithms and clinicians. If no

less than three eligible studies were identified, we used a random-effects model, in order to consider the heterogeneity between studies, to determine the combined sensitivity and specificity, accompanied by their respective 95% confidence intervals (CI). Additionally, summary receiver operating characteristic (ROC) curves were made to calculate the pooled sensitivities and specificities.

Furthermore, we performed a meta-regression analysis to investigate possible reasons for differences among the studies, analyzing factors such as the utilization of data augmentation, imaging view, imaging modality, usage of transfer learning technique, localization of pathology in model output, and the type of approach used in each study. Statistical significance was determined using a significance level of P less than 0.05. We performed all data analysis using Stata version 17 software (Stata Corp, College Station, TX) [13, 14].

Publication Bias

To reduce the potential impact of publication bias, we checked the reference lists of the studies included in our analysis. Furthermore, we thoroughly assessed publication bias by conducting a regression analysis involving diagnostic log odds ratios and asymmetry testing [15].

Results

Study Selection and Characteristics

Based on a database search, we found 3406 articles, of which 794 were retrieved from PubMed and 1406 from Scopus. The same numbers for other databased can be observed in Fig. 1. A total of 1280 were removed by automated and manual deduplication before the screening. The title and abstract screening process started with 2126 articles, 462 of them not being original, nine of which were not written in English, and 1583 were irrelevant to our topic.

The full-text screening was conducted on the remaining 72 records, which resulted in 45 studies being excluded based on the following criteria: full text not found ($n=13$), not enough data to be extracted ($n=17$), and only performed segmentation ($n=15$). To resolve any disagreements in each phase, the research team engaged in thorough discussions and deliberations among themselves (Fig. 1).

In accordance with Tables 1 and 2, 24 studies were included in the qualitative and quantitative synthesis. Twenty-three studies evaluated their algorithm through internal validation [10, 16–37], nine through external validation [1, 10, 18, 21, 24, 25, 38–40] (five used both types of validation [10, 18, 21, 24, 25]). Internal validation involves setting aside a portion of the initial patient dataset to evaluate the model's performance, whereas external validation employs an entirely distinct patient population to assess the model's performance [41]. Fourteen studies implemented the algorithm on X-ray images [1, 10, 16, 19, 20, 24, 25, 29–33, 36], eleven utilized MRI [18, 21–23, 27, 28, 34, 35, 37, 38, 40], two used CT-scan, and one used panoramic radiograph [26] (Arana et al. used both CT and X-ray[17]). Gitto et al. also performed PET-CT scan along with a CT scan [39]. In machine learning concept, a tuning set, often referred to as a validation set or a development set, represents a subset of your dataset employed for the purpose of fine-tuning the hyperparameters of a machine learning algorithm. It also plays a crucial role in making decisions concerning the model's architecture and configuration [42].

In most of the included studies, tumors have been detected in multiple body parts simultaneously. The standard reference for validation of the diagnosis of the lesion was also histopathology (in 20 studies) [1, 10, 17–25, 27–30, 34, 37, 39, 40, 43] or expert radiologists and biopsy at the same time in two studies [26, 31]. Five studies did not mention their standard reference [16, 32, 33, 35, 36].

Study Participants and Algorithm Development

The total number of study participants or bone lesions could not be determined since some articles provided insufficient data. The mean age of the participants differed widely between studies (median, 33.1; IQR, 26–53), the same as the percentage of disease-positive participants (median, 46.6; IQR, 30.6–75.8). Across all experiments, a cumulative quantity of 37,501 images was employed to train algorithms to detect lesions on 19,130 images as testing sets. A total of 43.2% contained primary bone malignancies, while the remaining was considered as control group. Eleven studies used data augmentation, and six used transfer learning. (Tables E1-3).

Model Performance

Tables E2 and E3 supplementary presents the results of various studies that evaluated model output using different metrics. Specifically, 21 studies used sensitivity and specificity, all 24 used accuracies, 14 used the area under the receiver operating characteristic curve, 19 used positive and negative predictive values, and 18 used F1 as the metric for assessment.

Quality Assessment

Figure 2 illustrates the compliance of each study to the modified version of the TRIPOD statement tool. None of the studies included in the analysis comply with items 8 and 9 of the protocol, pertaining to sample size and description of the handling of missing data, respectively. Furthermore, the adherence percentages were also notably low for some other key items, such

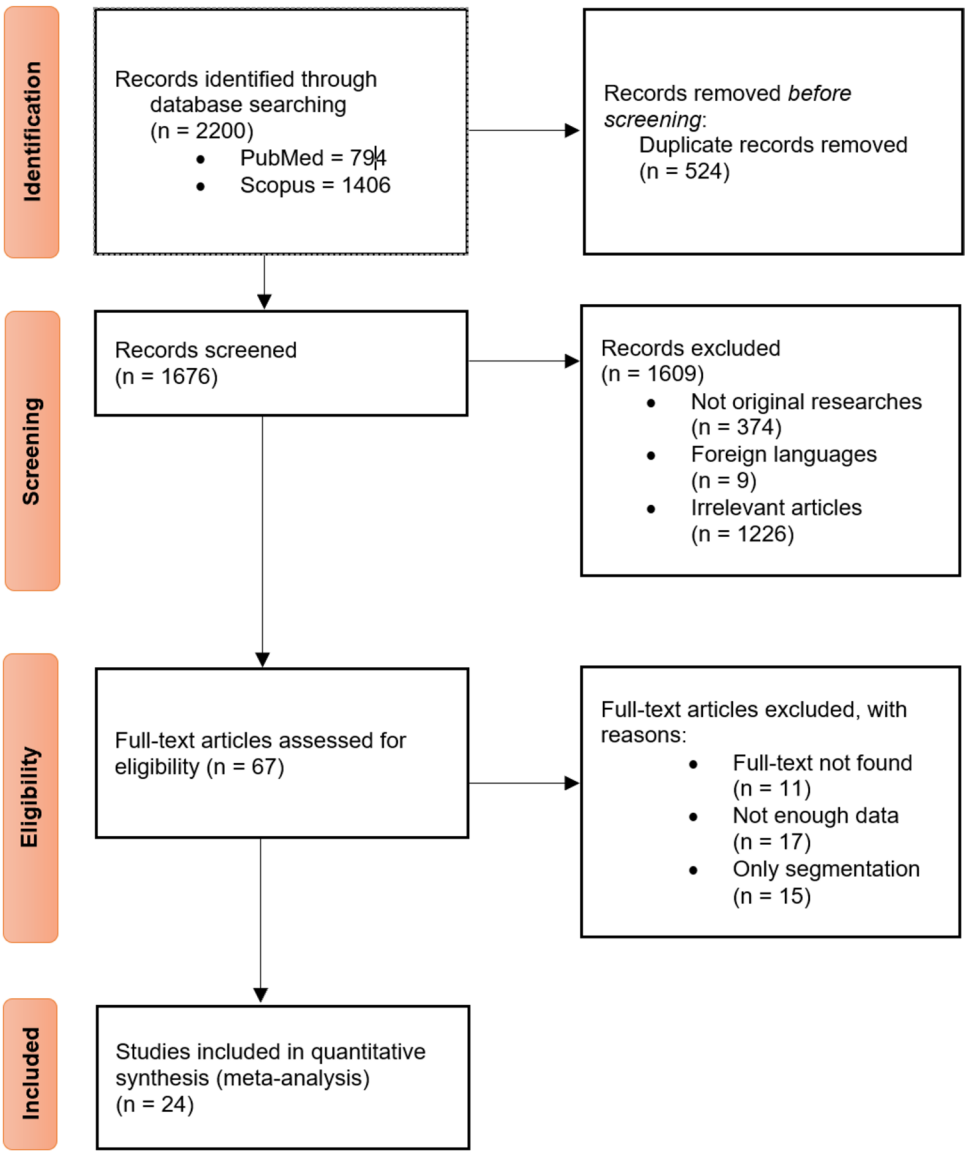

Fig. 1 Flowchart of study selection

as the description of participant flow through the study (24%), explanation of the usage of the prediction model (24%), and dates of image collection (30%). Of the 27 items evaluated, 17 received adherence ratings equal to or above 50%.

PROBAST was also evaluated on the included studies, resulting in a total rating of 37 (55.22%) studies with a low risk of bias and 31 (67.39%) studies with low concerns for applicability, as depicted in Fig. 2. Notably, 23 (95.83%) studies were classified as having a high risk of patient selection bias due to the predetermined inclusion and exclusion criteria. However, five studies (20.83%) raised concerns regarding the applicability of the results.

Meta-Analysis

A total of 27 studies containing adequate data were included in our study, from which we extracted 86 contingency tables for binary bone tumor detection. Of these, 42 contingency tables from 21 studies were related to AI algorithms with internal validation, 11 contingency tables from eight studies were related to AI algorithms with external validation, ten contingency tables from two studies were related to clinicians' performance with internal validation, and 16 contingency Tables from three studies were related to clinicians' performance with external validation. An additional seven contingency tables were extracted from a single study investigating clinicians who used AI as an assistant tool [34]. A meta-analysis was conducted for all five groups of studies; however, due to insufficient data, the groups of clinicians with AI assistance and clinicians with internal validation had less than three studies each, necessitating further research to validate the results in these groups. Figure 3 is a forest plot displaying the effect estimates of four groups of our study and their confidence intervals. Each row in the plot corresponds to an AI model, with some studies evaluating multiple

Table 1 Study characteristics, internal validation

First author	Year	Country	Imaging modality	Target condition	View	Comparison group	No. of images per set			Reference standard	Model output
							Training	Tuning	Testing		
Reinus	1994	United States	X-ray	Focal bone lesions	NR	N/A	NR	NR	NR	NR	Binary classification and differential diagnosis of tumors
Arana	1998	Spain	X-ray, CT scan	Calvarial lesions	NR	Comparison with other algorithm	NR	NR	NR	Histologic diagnosis	Binary classification
Do	2017	USA	X-ray	Bone tumors	AP view or image with best visualization of the tumor	N/A	NR	NR	710	Pathologic diagnosis and pathognomonic features	Pathological diagnosis and differential diagnosis of tumors
Ho	2019	South Korea	X-ray	Knee bone tumors	NR	Comparison of multiple architectures	NR	NR	963	Bone tumor specialists	Binary and ternary classification
Gitto	2020	Italy	1.5-T MRI	Cartilaginous tumors of the bone	T1W & T2W	Expert clinician	NR	NR	NR	Histological diagnosis	Binary classification
He	2020	China	X-ray	Primary bone tumors	NR	Expert clinicians	NR	NR	NR	Histological diagnosis	Binary and ternary classification
Altameem	2020	Saudi Arabia	X-ray	Bone cancer	NR	Comparison of multiple algorithms	NR	NR	NR	NR	Bone cancer detection
Chianca	2021	Italy	1.5-T MRI	Spinal lesions	T1W, T2W & DWI	Comparison of other algorithms and expert clinicians	NR	NR	NR	Post-operative pathological diagnosis	Binary and ternary classification
Eweje	2021	United States	MRI	Bone and soft tissue tumors	T1W, T2W & T1C	Comparison of other algorithms and expert clinicians	NR	NR	NR	Histological diagnosis	Binary classification
Lee	2021	Korea	Panoramic radiographs	jaw lesions	Panoramic	Comparison of multiple algorithms	322	NR	134	CT and histologically confirmed diagnosis	Binary classification
Pan	2021	China	X-ray	Bone tumors	AP & lateral	N/A	NR	NR	NR	Pathological diagnosis	Binary and ternary classification
Sharma	2021	India	X-ray	Bone cancer	NR	Comparison with other algorithm	65	NR	40	NR	Detection of cancerous bone from healthy bone
Liu	2021	China	X-ray	Bone tumors	AP & lateral	Five radiologists	784	97	101	Pathological diagnosis	Binary and ternary classification

Table 1 (continued)

First author	Year	Country	Imaging modality	Target condition	View	Comparison group	No. of images per set			Reference standard	Model output
							Training	Tuning	Testing		
Gitto-A	2022	Italy	1.5-T MRI	Spine bone tumors	T2W & DWI	Models with different feature numbers	NR	NR	NR	Histological diagnosis	Binary classification
Park	2022	Korea	X-ray	Proximal femur tumors	AP	Expert clinicians and CNN model without applying image pre-processing procedures	538	NR	NR	Confirmed using MRI and tissue specimens	Ternary classification
Liu-A	2022	China	MRI	Primary spine tumors	Axial and Sagittal	Expert clinicians	14,072	NR	6161	Pathological diagnosis	Binary classification
Consalvo	2022	Germany	X-ray	Ewing sarcoma and acute osteomyelitis	NR	NR	127.4	27.3	27.3	Histopathological diagnosis	Binary classification
Liu-B	2022	China	MRI	Spinal tumors	Sagittal T1W, T2W, and FS-T2W	Expert clinicians	15,778	NR	4815	Pathologically confirmed by trocar biopsy or surgery	Binary classification
Von Schacky	2022	Germany	X-ray	Primary bone tumors	NR	Expert clinicians	NR	NR	NR	Histopathological diagnosis	Binary classification
Keyang Zhao	2022	China	3-T MRI	Musculoskeletal tumors	Axial and sagittal, and coronal T1W, T2W, DWI, and CET1-W	Expert clinicians	180	62	62	Pathological diagnosis	Binary classification
Shen Zhao	2022	China	MRI	Vertebrae tumor	NR	NR	NR	NR	NR	NR	Detection of vertebral tumor

Postero-anterior (PA), Magnetic resonance imaging (MRI), Tesla (T), T1-weighted (T1W), T2-weighted (T2W), fat saturated T2-weighted (FS-T2W), post contrast T1-weighted (T1C), contrast-enhanced T1-weighted (CET1-W), diffusion-weighted imaging (DWI), computed tomography (CT), not reported (NR), not applicable (N/A)

Table 2 Study characteristics, External Validation

First author	Year	Country	Imaging modality	Target condition	View	Comparison group	No. of images per set			Reference standard	Model output
							Training	Tuning	Testing		
Ho	2019	South Korea	X-ray	Knee bone tumors	NR	Comparison of multiple architectures	NR	NR	963	Bone tumor specialists	Ternary classification
He	2020	China	X-ray	Primary bone tumors	NR	Expert clinicians	NR	NR	NR	Histological diagnosis	Binary and ternary classification
Chianca	2021	Italy	1.5-T MRI	Spinal lesions	T1W, T2W, and DWI	Comparison of other algorithms	NR	NR	NR	Post-operative pathological diagnosis	Binary and ternary classification
Eweje	2021	United States	MRI	Bone & soft tissue tumors	T1W, T2W, and T1C	Comparison of other algorithms	NR	NR	NR	Histological diagnosis	Binary classification
Gitto	2021	Italy	CT and PET-CT scan	Atypical cartilaginous tumor or appendicular chondrosarcoma	NR	Musculoskeletal radiologist	84	NR	36	Histological diagnosis	Binary classification
Gitto-B	2022	Italy	1.5-T and 3-T MRI	ACT & CST of long bones	T1W and T2W	Expert bone tumor radiologist	93	NR	65	Post-operative pathological diagnosis	Binary classification
Li	2022	China	X-ray	Bone tumors	NR	Expert clinicians	740	262	428	Histological diagnosis	Four-way classification
Von Schacky	2022	Germany	X-ray	Primary bone tumors	NR	Expert clinicians	NR	NR	NR	Histopathological diagnosis	Binary Classification

Magnetic resonance imaging (MRI), Tesla (T), T1-weighted (T1W), T2-weighted (T2W), post contrast T1-weighted (T1C), diffusion-weighted imaging (DWI), computed tomography (CT), positron emission tomography (PET), atypical cartilaginous tumor (ACT), grade II chondrosarcoma (CS2), not reported (NR)

Table 3 Pooled sensitivities, specificities, area under the curve, positive likelihood ratio, negative likelihood ratio, and diagnostic odds ratio

Parameter	Sensitivity (%)	Specificity (%)	AUC	Positive likelihood ratio	Negative likelihood ratio	Diagnostic odds ratio	No. of contingency tables
Algorithms, internal validation	84 (79–88)	86 (81–90)	0.92 (0.89–0.94)	6.0 (4.2–8.6)	0.18 (0.13–0.25)	33 (17–61)	42
Algorithms, external validation	84 (75–90)	91 (83–96)	0.93 (0.91–0.95)	9.6 (4.9–19.0)	0.18 (0.11–0.28)	54 (25–117)	11
Clinicians, internal validation	76 (64–85)	64 (55–72)	0.74 (0.70–0.78)	2.1 (1.7–2.6)	0.38 (0.26–0.54)	6 (4–9)	10
Clinicians, external validation	85 (73–92)	94 (89–97)	0.96 (0.94–0.97)	13.6 (7.2–25.7)	0.316 (0.09–0.3)	85 (29–250)	16
Clinicians with AI assistance	95 (86–98)	57 (48–66)	0.66 (0.62–0.70)	2.2 (1.8–2.7)	0.09 (0.03–0.26)	24 (8–75)	7

Area under the curve (AUC), artificial intelligence (AI)

AI algorithms. Figure 4 demonstrates the hierarchical summary receiver operating characteristic (HSROC) curves for internal and external validation test sets, respectively.

The pooled sensitivities for AI algorithms and clinicians on internal validation test sets were 84% (95% CI: 79,88) and 76% (95% CI: 64,85), and pooled specificities were 86% (95% CI: 81,90), and 64% (95% CI: 55,72), respectively. At external validation, the pooled sensitivity and specificity for AI algorithms were 84% (95% CI: 74,90) and 91% (95% CI: 83,96), and also for clinicians were 85% (95% CI: 73,92) and 94% (95% CI: 89,97) respectively (Table 3). The pooled sensitivity and specificity from seven contingency tables of clinicians with AI assistance are 95% (95% CI: 86.98) and 57% (95% CI: 48,66). The positive and negative likelihood ratio, AUC, and diagnostic odds ratio of each group is also demonstrated in Table 3.

The findings from our analyses of studies using the internal validation demonstrate that the use of MRI was associated with lower sensitivity and specificity (75%, CI: 67, 83, and 74%, CI: 66, 83, respectively) compared to other imaging modalities (90%, CI: 86, 94, *P*-value = 0.00 and 92%, CI: 89, 95, *P*-value = 0.00). It was also found that the use of data augmentation resulted in a lower sensitivity (82%, CI:74–90 versus 86%, CI: 80–91, *P*-value:0.00) and higher specificity (90%, CI: 85–96 versus 82%, CI: 75–89, *P*-value:0.10). Furthermore, transfer learning had a positive impact on both sensitivity and specificity (Sensitivity: 97%, CI: 94–100 versus 81%, CI: 76,86, *P*-value: 0.36) (specificity: 97%, CI: 93,100 versus 83%, CI:78–89, *P*-value: 0.43). The study found that there was no significant difference in sensitivity and specificity between studies that utilized patients without bone lesions as a control group versus those that used patients with other types of tumors as a control group (excluding the specific tumor under investigation). The *P*-values for sensitivity and specificity were 0.26 and 0.77, respectively. In Tables E9-12, other parameters and results from meta-regression can be found for other groups as well.

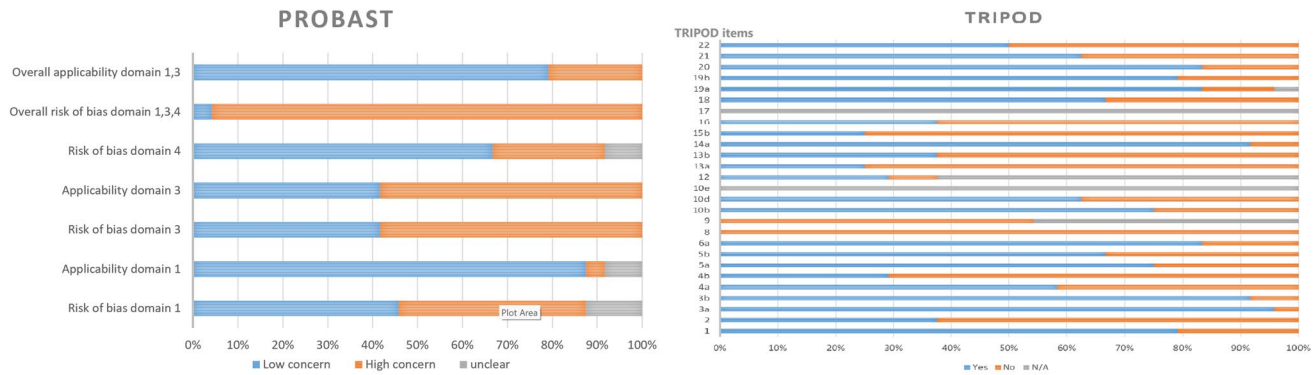

Fig. 2 Adherence to Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD), and Prediction Model Study Risk of Bias Assessment Tool (PROBAST) reporting guidelines

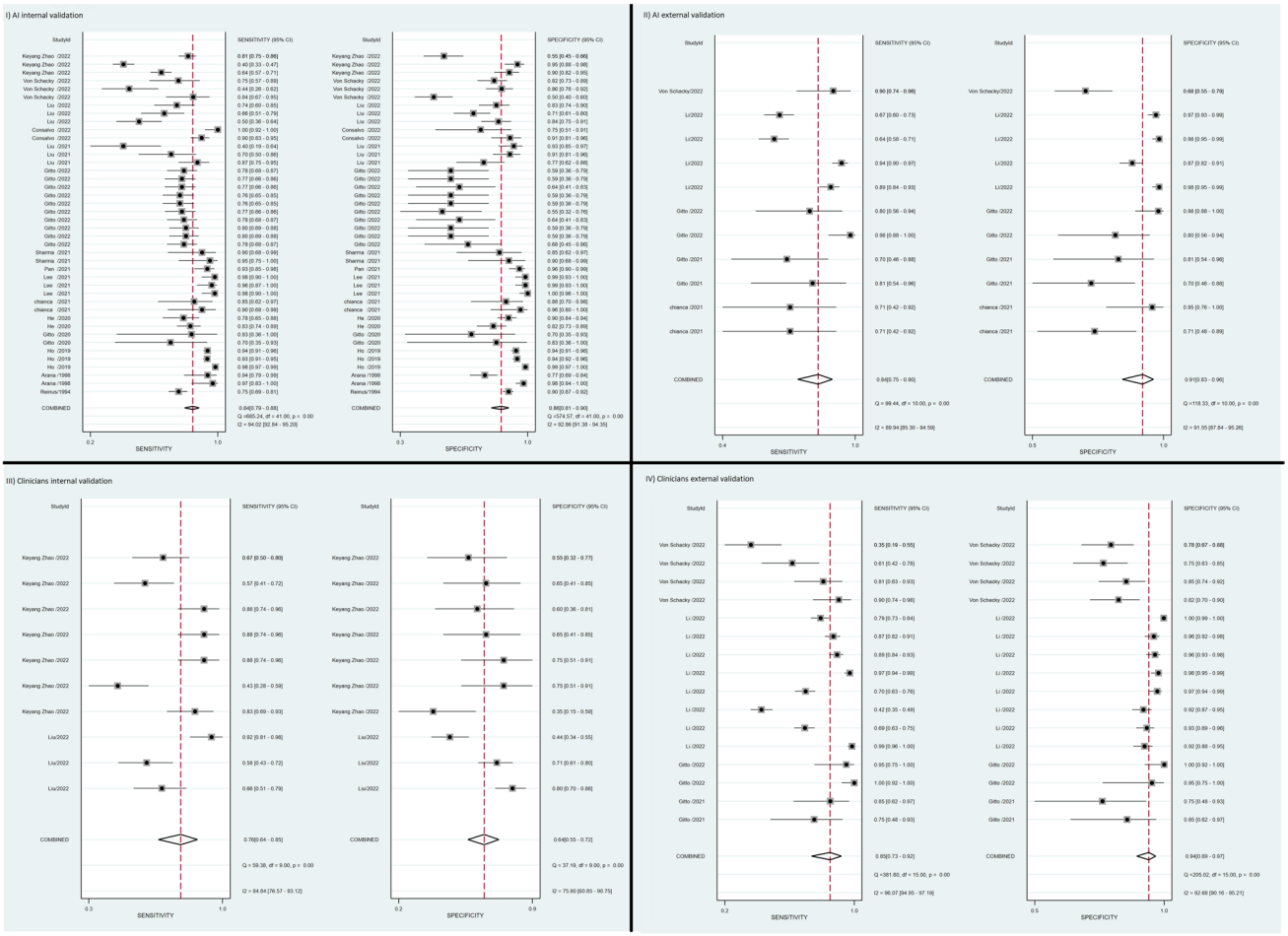

Fig. 3 Forest plots of algorithms and clinicians' performance. The plot shows the effect size estimate (diamond symbol), confidence interval (horizontal line), and the individual study effect sizes (square symbols) with their corresponding weights

Publication Bias

Publication bias assessment was performed individually for all five groups, including studies with internally validated AI, externally validated AI, internally validated clinician, externally validated clinician, and clinicians with AI assistance performance. The slope coefficient was − 31.15 (95%CI: − 48.92, − 13.38; $P = 0.00$), − 13.91 (95%CI: − 31.17, 3.35; $P = 0.10$), − 3.26 (95%CI: − 27.74, 21.22; $P = 0.77$), − 24.45 (95%CI: − 65.62, 16.72; $P = 0.22$), and − 985.26 (95%CI: − 1311.26, − 659.23; $P = 0.00$) respectively, which indicates a weak association between sample size and effect size for the first four groups, and low potential for publication bias (Table E8).

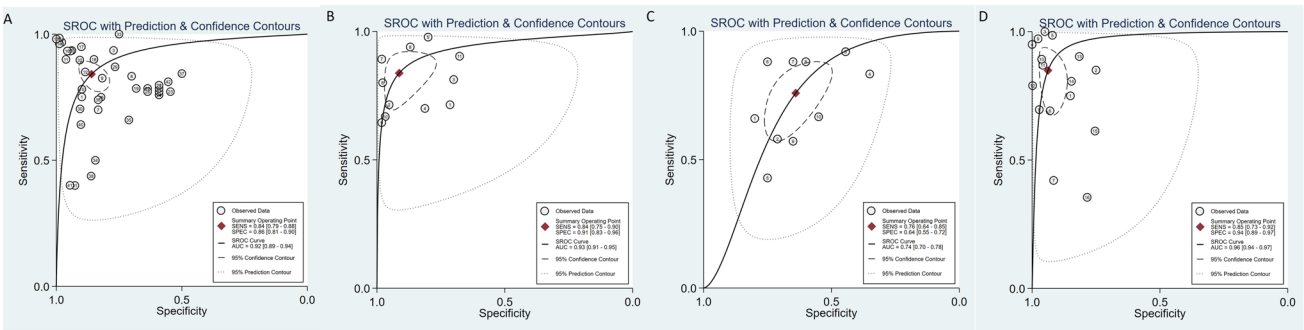

Fig. 4 Hierarchical summary receiver operating characteristic (HSROC) curves for **A** algorithms on internal validation, **B** algorithms on external validation, **C** clinicians on internal validation, **D** clinicians on external validation

Discussion

To the best of our knowledge, there appears to be an absence of meta-analytical investigations into the performance of artificial intelligence in detecting primary bone lesions via imaging. There is, however, a growing body of original literature on this subject. Zheng and colleagues conducted a meta-analysis, examining the use of artificial intelligence in detecting metastasis in various parts of the body, including lymph nodes, bone, and other types, through imaging [44]. The authors reported a diagnostic odds ratio of 22.14 (95% CI: 18.52, 26.46) for all metastases, significantly less than the diagnostic odds ratios derived from both our internal and external validation studies, which are 33 (95% CI: 17,61) and 54 (95% CI: 25,117).

In certain cases, the machines outperformed human experts, demonstrating robust performance. Based on these findings, automated methods may offer an alternative to traditional expert-driven approaches for the diagnosis of certain medical conditions. Despite the promising results of our study, caution must be exercised when interpreting the findings due to the potential for underreporting of radiologists' performance. This may be attributed to a range of factors, including the following:

1. The studies under review lack specification regarding the level of experience exhibited by the radiologists.
2. The modalities used in different studies to detect bone tumor may not be the modality of choice for this purpose.
3. The lack of sufficient number of studies that have simultaneously examined the quality of algorithms and the accuracy of radiologists on a dataset.

In a meta-analysis of 21 studies performed in 2020, Younis et al. measured the pooled sensitivity and specificity of 18-fluorodeoxyglucose-positron emission tomography (18 F-FDG-PET) combined with CT, modalities used by expert radiologists, for the detection of bone and soft tissue sarcomas, 89.2 and 76.3%, respectively [45]. A meta-analysis of 31 studies examined expert radiologists' performance in detecting Ewing sarcoma (ES) using 18 F-FDG PET as well as PET/computed tomography (PET/CT). The pooled sensitivity and specificity for total ES lesions were found to be 92.6% and 74.1%, respectively [46].

Furthermore, several secondary outcomes have been identified from the data in addition to the general findings. The meta-regression findings in Table E9 illustrate that applying AI to the identification of bone tumors using X-ray and CT pictures gave much better results than MRI images, which was predictable based on previous studies revealing that CT is a superior modality for visualizing bone structure and abnormalities [47]. Furthermore, the application of transfer learning, a machine learning approach for enhancing learning in new tasks by leveraging knowledge gained from a related task (48), has significantly improved the performance of the automated method. It should be noted, however, that data augmentation, a strategy for constructing incremental algorithms or sampling algorithms using unobserved data or latent variables [48], adversely affects the sensitivity of machine learning algorithms [43, 49].

For a proper evaluation of the results of this study, consideration must be given to the limitations encountered during the investigation. It is difficult to achieve the utmost precision in the conclusions of this study due to a lack of data available for analysis. As can be observed in Table E1, the assessment of the accuracy of AI on external validation was predicated solely on data derived from five studies [1, 10, 18, 38, 39]. One of the main challenges was analyzing the performance of radiologists. Table E1 also demonstrates that just two studies reported the data from radiologist on internal validation [34, 43] and three on external validation [1, 10, 38]. Presently, a scarcity of research on this topic creates a remarkable chance for diligent researchers to conduct further investigations and bridge the existing gap in scientific understanding. Such endeavors would offer valuable insights into the viability of utilizing this technology in real-world contexts and thus promote its effective implementation for the advancement of medical practice. The other limitations we encountered in this review article were the absence of precise participant and image quantity details in some studies, particularly regarding those with or without a bone lesion. Additionally, the limited number of studies for each specific body region hindered our ability to conduct meaningful sub-group analyses based on different areas containing bone malignancies. Consequently, future research endeavors in this field hold the potential to provide valuable insights, enabling researchers to more accurately assess AI performance in detecting bone tumors within distinct body regions. We intended to perform a sensitivity analysis to minimize the impact of publication bias measured by modified-PROBAST method. However, the results of quality assessment showed that the number of studies in low risk of bias subgroup did not reach a suitable number (four studies) to perform analysis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-023-00945-3.

Author Contribution Soheil Mohammadi and Mohammad Amin Salehi designed the project, contributed to protocol development, literature search, screening, data extraction, and writing of the original draft. Hamid Harandi contributed to writing of the original draft. Seyed Sina Zakavi and Ali Jahanshahi contributed to screening and protocol development. Mohammad Shahrabi Farahani contributed to the writing of the original draft. Jim S. Wu encouraged and supervised the project, reviewed the manuscript, and contributed to the writing of the original and final draft.

Data Availability The data that support the findings of this study are available from the authors upon reasonable request.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Li J, Li S, Li X, Miao S, Dong C, Gao C, et al. Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model. European Radiology. 2022.
2. Kerr DL, Dial BL, Lazarides AL, Catanzano AA, Lane WO, Blazer DG, 3rd, et al. Epidemiologic and survival trends in adult primary bone tumors of the spine. Spine J. 2019;19(12):1941-9.
3. Georgeanu VA, Mămuleanu M, Ghiea S, Selișteanu D. Malignant Bone Tumors Diagnosis Using Magnetic Resonance Imaging Based on Deep Learning Algorithms. Medicina (Lithuania). 2022;58(5).
4. Salazar C, Leite M, Sousa A, Torres J. Correlation between imagenological and histological diagnosis of bone tumors. A retrospective study. Acta Ortop Mex. 2019;33(6):386-90.
5. Goyal N, Kalra M, Soni A, Baweja P, Ghonghe NP. Multi-modality imaging approach to bone tumors - State-of-the art. J Clin Orthop Trauma. 2019;10(4):687-701.
6. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500-10.
7. Han SH, Kim KW, Kim S, Youn YC. Artificial Neural Network: Understanding the Basic Concepts without Mathematics. Dement Neurocogn Disord. 2018;17(3):83-9.
8. Driver CN, Bowles BS, Bartholmai BJ, Greenberg-Worisek AJ. Artificial Intelligence in Radiology: A Call for Thoughtful Application. Clin Transl Sci. 2020;13(2):216-8.
9. Nair AV, Ramanathan S, Sathiadoss P, Jajodia A, Blair Macdonald D. Barriers to artificial intelligence implementation in radiology practice: What the radiologist needs to know. Radiologia (Engl Ed). 2022;64(4):324-32.
10. von Schacky CE, Wilhelm NJ, Schäfer VS, Leonhardt Y, Jung M, Jungmann PM, et al. Development and evaluation of machine learning models based on X-ray radiomics for the classification and differentiation of malignant and benign bone tumors. European Radiology. 2022;32(9):6247-57.
11. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Annals of internal medicine. 2015;162(1):55-63.
12. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine. 2019;170(1):51-8.
13. Dwamena B. MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies. 2009.
14. Harbord RM, Whiting P. Metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression. The Stata Journal. 2009;9(2):211-29.
15. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. Journal of clinical epidemiology. 2005;58(9):882-93.
16. Altameem T. Fuzzy rank correlation-based segmentation method and deep neural network for bone cancer identification. Neural Computing and Applications. 2020;32(3):805-15.
17. Arana E, Marti-Bonmati L, Bautista D, Paredes R. Calvarial eosinophilic granuloma: Diagnostic models and image feature selection with a neural network. Academic Radiology. 1998;5(6):427-34.
18. Chianca V, Cuocolo R, Gitto S, Albano D, Merli I, Badalyan J, et al. Radiomic Machine Learning Classifiers in Spine Bone Tumors: A Multi-Software, Multi-Scanner Study. Eur J Radiol. 2021;137:109586.
19. Consalvo S, Hinterwimmer F, Neumann J, Steinborn M, Salzmann M, Seidl F, et al. Two-Phase Deep Learning Algorithm for Detection and Differentiation of Ewing Sarcoma and Acute Osteomyelitis in Paediatric Radiographs. Anticancer Research. 2022;42(9):4371-80.
20. Do BH, Langlotz C, Beaulieu CF. Bone Tumor Diagnosis Using a Naïve Bayesian Model of Demographic and Radiographic Features. J Digit Imaging. 2017;30(5):640-7.
21. Eweje FR, Bao B, Wu J, Dalal D, Liao WH, He Y, et al. Deep Learning for Classification of Bone Lesions on Routine MRI. EBioMedicine. 2021;68:103402.
22. Gitto S, Cuocolo R, Albano D, Chianca V, Messina C, Gambino A, et al. MRI radiomics-based machine-learning classification of bone chondrosarcoma. Eur J Radiol. 2020;128:109043.
23. Gitto S, Cuocolo R, van Langevelde K, van de Sande MAJ, Parafioriti A, Luzzati A, et al. MRI radiomics-based machine learning classification of atypical cartilaginous tumour and grade II chondrosarcoma of long bones. EBioMedicine. 2022;75.
24. He Y, Pan I, Bao B, Halsey K, Chang M, Liu H, et al. Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. EBioMedicine. 2020;62:103121.
25. Ho NH, Yang HJ, Kim SH, Jung ST, Joo SD. Regenerative semi-supervised bidirectional w-network-based knee bone tumor classification on radiographs guided by three-region bone segmentation. IEEE Access. 2019;7:154277-89.
26. Lee A, Kim MS, Han SS, Park PG, Lee C, Yun JP. Deep learning neural networks to differentiate Stafne's bone cavity from pathological radiolucent lesions of the mandible in heterogeneous panoramic radiography. PLoS ONE. 2021;16(7 July).
27. Liu H, Jiao M, Yuan Y, Ouyang H, Liu J, Li Y, et al. Benign and malignant diagnosis of spinal tumors based on deep learning and weighted fusion framework on MRI. Insights Imaging. 2022;13(1):87.
28. Liu H, Jiao ML, Xing XY, Ou-Yang HQ, Yuan Y, Liu JF, et al. BgNet: Classification of benign and malignant tumors with MRI multi-plane attention learning. Front Oncol. 2022;12:971871.
29. Liu Y, Yang P, Pi Y, Jiang L, Zhong X, Cheng J, et al. Automatic identification of suspicious bone metastatic lesions in bone scintigraphy using convolutional neural network. BMC Med Imaging. 2021;21(1):131.
30. Pan D, Liu R, Zheng B, Yuan J, Zeng H, He Z, et al. Using Machine Learning to Unravel the Value of Radiographic Features for the Classification of Bone Tumors. Biomed Res Int. 2021;2021:8811056.
31. Park CW, Oh SJ, Kim KS, Jang MC, Kim IS, Lee YK, et al. Artificial intelligence-based classification of bone tumors in the proximal femur on plain radiographs: System development and validation. PLoS One. 2022;17(2):e0264140.
32. Reinus WR, Wilson AJ, Kalman B, Kwasny S. Diagnosis of focal bone lesions using neural networks. Investigative Radiology. 1994;29(6):606-11.
33. Sharma A, Yadav DP, Garg H, Kumar M, Sharma B, Koundal D. Bone Cancer Detection Using Feature Extraction Based Machine Learning Model. Comput Math Methods Med. 2021;2021:7433186.
34. Zhao K, Zhang M, Xie Z, Yan X, Wu S, Liao P, et al. Deep Learning Assisted Diagnosis of Musculoskeletal Tumors Based on Contrast-Enhanced Magnetic Resonance Imaging. Journal of Magnetic Resonance Imaging. 2022;56(1):99-107.
35. Zhao S, Chen B, Chang H, Chen B, Li S. Reasoning discriminative dictionary-embedded network for fully automatic vertebrae tumor diagnosis. Med Image Anal. 2022;79:102456.
36. Do NT, Jung ST, Yang HJ, Kim SH. Multi-Level Seg-Unet Model with Global and Patch-Based X-ray Images for Knee Bone Tumor Detection. Diagnostics (Basel). 2021;11(4).
37. Fouad H, Hassanein AS, Soliman AM, Al-Feel H. Internet of Medical Things (IoMT) Assisted Vertebral Tumor Prediction

Using Heuristic Hock Transformation Based Gautschi Model–A Numerical Approach. IEEE Access. 2020;8:17299-309.
38. Gitto S, Bologna M, Corino VDA, Emili I, Albano D, Messina C, et al. Diffusion-weighted MRI radiomics of spine bone tumors: feature stability and machine learning-based classification performance. Radiol Med. 2022.
39. Gitto S, Cuocolo R, Annovazzi A, Anelli V, Acquasanta M, Cincotta A, et al. CT radiomics-based machine learning classification of atypical cartilaginous tumours and appendicular chondrosarcomas. EBioMedicine. 2021;68:103407.
40. Ouyang H, Meng F, Liu J, Song X, Li Y, Yuan Y, et al. Evaluation of Deep Learning-Based Automated Detection of Primary Spine Tumors on MRI Using the Turing Test. Front Oncol. 2022;12:814667.
41. Shung D, Simonov M, Gentry M, Au B, Laine L. Machine Learning to Predict Outcomes in Patients with Acute Gastrointestinal Bleeding: A Systematic Review. Dig Dis Sci. 2019;64(8):2078-87.
42. Montesinos-López O, Montesinos A, Crossa J. Overfitting, Model Tuning, and Evaluation of Prediction Performance. 2022. p. 109–39.
43. Liu S, Feng M, Qiao T, Cai H, Xu K, Yu X, et al. Deep Learning for the Automatic Diagnosis and Analysis of Bone Metastasis on Bone Scintigrams. Cancer Manag Res. 2022;14:51-65.
44. 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.
45. Younis MH, Abu-Hijleh HA, Aldahamsheh OO, Abualruz A, Thalib L. Meta-Analysis of the Diagnostic Accuracy of Primary Bone and Soft Tissue Sarcomas by 18F-FDG-PET. Med Princ Pract. 2020;29(5):465-72.
46. Seth N, Seth I, Bulloch G, Siu AHY, Guo A, Chatterjee R, et al. (18) F-FDG PET and PET/CT as a diagnostic method for Ewing sarcoma: A systematic review and meta-analysis. Pediatr Blood Cancer. 2022;69(3):e29415.
47. Zimmer WD, Berquist TH, McLeod RA, Sim FH, Pritchard DJ, Shives TC, et al. Bone tumors: magnetic resonance imaging versus computed tomography. Radiology. 1985;155(3):709-18.
48. van Dyk DA, Meng X-L. The Art of Data Augmentation. Journal of Computational and Graphical Statistics. 2001;10(1):1-50.
49. Torrey L, Shavlik J. Transfer learning. Handbook of Research on Machine Learning Applications. 2009.