European Radiology (2022) 32:7196–7216
https://doi.org/10.1007/s00330-022-08956-4

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Diagnostic accuracy and potential covariates of artificial intelligence for diagnosing orthopedic fractures: a systematic literature review and meta-analysis

Xiang Zhang[1] · Yi Yang[1] · Yi-Wei Shen[1] · Ke-Rui Zhang[1] · Ze-kun Jiang[2] · Li-Tai Ma[1] · Chen Ding[1] · Bei-Yu Wang[1] · Yang Meng[1] · Hao Liu[1]

Received: 11 March 2022 / Revised: 7 May 2022 / Accepted: 8 June 2022 / Published online: 27 June 2022

Abstract

Objectives To systematically quantify the diagnostic accuracy and identify potential covariates affecting the performance of artificial intelligence (AI) in diagnosing orthopedic fractures.

Methods PubMed, Embase, Web of Science, and Cochrane Library were systematically searched for studies on AI applications in diagnosing orthopedic fractures from inception to September 29, 2021. Pooled sensitivity and specificity and the area under the receiver operating characteristic curves (AUC) were obtained. This study was registered in the PROSPERO database prior to initiation (CRD 42021254618).

Results Thirty-nine were eligible for quantitative analysis. The overall pooled AUC, sensitivity, and specificity were 0.96 (95% CI 0.94–0.98), 90% (95% CI 87–92%), and 92% (95% CI 90–94%), respectively. In subgroup analyses, multicenter designed studies yielded higher sensitivity (92% vs. 88%) and specificity (94% vs. 91%) than single-center studies. AI demonstrated higher sensitivity with transfer learning (with vs. without: 92% vs. 87%) or data augmentation (with vs. without: 92% vs. 87%), compared to those without. Utilizing plain X-rays as input images for AI achieved results comparable to CT (AUC 0.96 vs. 0.96). Moreover, AI achieved comparable results to humans (AUC 0.97 vs. 0.97) and better results than non-expert human readers (AUC 0.98 vs. 0.96; sensitivity 95% vs. 88%).

Conclusions AI demonstrated high accuracy in diagnosing orthopedic fractures from medical images. Larger-scale studies with higher design quality are needed to validate our findings.

Key Points

- *Multicenter study design, application of transfer learning, and data augmentation are closely related to improving the performance of artificial intelligence models in diagnosing orthopedic fractures.*
- *Utilizing plain X-rays as input images for AI to diagnose fractures achieved results comparable to CT (AUC 0.96 vs. 0.96).*
- *AI achieved comparable results to humans (AUC 0.97 vs. 0.97) but was superior to non-expert human readers (AUC 0.98 vs. 0.96, sensitivity 95% vs. 88%) in diagnosing fractures.*

Keywords Fractures, bone · Artificial intelligence · Meta-analysis

Abbreviations

AI	Artificial intelligence
CI	Confidence interval
cML	Classical machine learning
CNN	Convolutional neural networks
DEXA	Dual-energy X-ray absorptiometry

Xiang Zhang and Yi Yang contributed equally to this work.

✉ Hao Liu
liuhao6304@126.com

[1] Department of Orthopedics, Orthopedic Research Institute, West China Hospital, Sichuan University, No. 37 Guo Xue Rd, Chengdu 610041, China

[2] West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610000, China

DL	Deep learning
FN	False negative
FP	False positive
SE	Sensitivity
SP	Specificity
TN	True positive
TP	True positive

Introduction

Owing to an increase in the aging population, orthopedic fractures have become a major health issue. It has an estimated global incidence of 9.0–22.8 cases per 1,000 people per year [1–3]. Although the radiological examination is the main method for diagnosing fractures, misinterpretation of images leading to misdiagnoses is not uncommon and could be attributed to the lack of experience of radiologists [4] or excessive workloads [5, 6]. A misdiagnosis of fracture could directly affect patients' outcomes and lead to serious complications such as malunion or arthritis, due to delayed surgical treatments [7, 8]. From a clinical perspective, it is important to formulate a user-friendly diagnostic model that could be easily interpreted, even by less-experienced doctors, for early and accurate diagnosis of orthopedic fractures on medical images.

Artificial intelligence (AI) has shown remarkable promise in detecting, localizing, and identifying abnormity in medical imaging fields, such as screening of breast cancer [9–11], analysis of retinal images [12, 13], detection of brain metastasis [14, 15], and classification of skin lesions [16, 17]. The amount of research in AI for fracture detection and localization on medical images has greatly increased. Studies are consistently showing that AI can automatically detect varying sizes or types of fractures via different AI algorithms. To compare the results of these studies and identify the optimal AI algorithm for fracture detection, a comparative study is needed.

However, AI algorithms are also reported to be inherently vulnerable to overfitting and spectrum bias [18–20]. Further, algorithm accuracy further depends on a variety of factors such as the type of study, i.e., multicenter or single-center study [21–23], the use of transfer learning [24–27] or data augmentation [28–30], whether the training dataset is well-balanced [31–33], adoption of DL or cML [34–36], and the types of the medical image used [37–39]. Transfer learning means that a convolutional neural network (CNN) is trained starting from the weights of a pretrained network to accomplish a different but similar task, thus requiring fewer image data. The data augmentation technique is used to amplify the data, which involved making a number of non-exact copies, or transformations of each image. This served to provide the CNNs with more training examples. A balanced test dataset means that the dataset has approximately the same number of fractures as non-fractures and imbalanced datasets may cause the model to learn insufficiently from less of that type of data. The DL group was defined as the studies that utilized CNNs as their main algorithm. Otherwise, the studies were classified into the cML group. The main difference between them is that DL replaces the process of feature extraction, but requires large datasets. Thus, adequate comparisons of the technical details used in such studies are also required.

Therefore, this comprehensive systematic review and meta-analysis aimed to determine the diagnostic accuracy of AI-based systems at detecting fractures in radiological images and explore factors affecting the performance of these models, and guide future research.

Materials and methods

This systematic review was conducted following the Preferred Reported Items for Systematic Reviews and Meta-Analysis guidelines [40], and the study protocol was registered in the international open-access Prospective Register of Systematic Reviews (PROSPERO, number: CRD42021254618) prior to data retrieval.

Literature search

A comprehensive literature search was conducted on PubMed, Embase, Web of Science, and Cochrane Library from inception to September 29, 2021, to retrieve all relevant studies concerning AI in the diagnosis of fracture from medical images. Search terms included both entry terms and medical descriptors/MeSH terms such as "artificial intelligence," "machine learning," "deep learning," "neural network," and "fracture." Supplementary File 1 summarizes the search strategy used in each database.

Study selection

Studies satisfying the following criteria were included: (1) Population type—patients with orthopedic fractures; (2) index test—diagnostic accuracy evaluated with computational models and algorithms; (3) reference standard—radiologists' conclusions based on CT or MRI; (4) design—prospective or retrospective studies.

The following studies were excluded: (1) letters, editorials, conference abstracts, systematic reviews or meta-analyses, consensus statements, guidelines; (2) non-English publications; (3) contained patients with confounding factors such as bone-related diseases, i.e., osteoporosis; (4) had insufficient data on 2 × 2 contingency tables; (5) involved fracture prediction rather than diagnosis; (6) not included orthopedic fractures such as dental fracture, and (7) full text was not available.

Data extraction

Data extraction was conducted by two independent reviewers using a piloted and standardized data extraction form. Any disagreements were resolved by mutual consensus. The following data from each included study were retrieved: (1) study characteristics—authors' information, study design (multicenter or single-center, prospective or retrospective), type of radiological images (X-ray or CT or DEXA), study cohort and image sources, gold standard, sample size; (2) patients' characteristics—mean age, male-to-female ratio, fracture location; (3) algorithms characteristics—specific type or name of the algorithm of AI, data augmentation, and transfer learning information; (4) DIAGNOSTIC accuracy of test results—TP, FP, FN, and TN calculated from the provided data.

Risk of bias and applicability

The quality and risk of bias were assessed by two independent reviewers using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool [41]. This tool included four domains (patient selection, index test, reference standard, flow and timing) for risk of bias assessment and three domains (patient selection, index test, reference standard) for applicability concerns. Each domain was assessed as low, unclear, or high risk. Risk of bias graphs were plotted using the Revman software (version 5.3).

Statistical analysis

The Stata (version16) and MetaDiSc (version1.4) software were used to perform statistical analysis. The random-effects model was used in all the combinations. Pooled SE and SP, AUC, and corresponding 95% confidence intervals (CIs) were calculated. Forest plots were drawn to assess the heterogeneity in sensitivity and specificity. ROC curves comparing AI and human readers in diagnosing fractures were drawn using the Review Manager software (version 5.3).

Statistical heterogeneity was assessed using the I^2 test. The I^2 statistic describes the percentage of variation in each study due to heterogeneity rather than chance, while I^2 values of 0–25%, 25–50%, 50–75%, and > 75% represent very low, low, medium, and high heterogeneity, respectively [42].

Spearman correction coefficient test was used to evaluate the threshold effect. In addition, Deek's funnel plot asymmetry test was used to determine the potential presence of publication bias. p values > 0.1 indicated a low publication bias.

In addition, a subgroup analysis of studies was performed to further evaluate the effects of heterogeneity. The six covariates considered were as follows: (a) multicenter or single-center study; (b) deep learning or classical machine learning; (c) balanced or unbalanced training set; (d) with or without transfer learning; (e) with or without data augmentation; (f) medical image type (X-ray or CT or DEXA); (g) risk of bias; (h) presence or absence of localization of fractures; (i) one vs more than one type of fracture.

Results

Selection of studies

The systematic literature search initially identified 8335 potentially eligible articles from PubMed, Embase, Web of Science, and Cochrane Library (Fig. 1). After excluding 1685 duplicates, screening of the remaining 6650 titles and abstracts yielded 127 potentially eligible articles. After full-text reviews of the 127 provisionally eligible articles, 88 articles were excluded due to no access to full text (3), contained insufficient data (54), not written in English (3), fracture prediction was not related to diagnosis (3), absence of orthopedic fractures (8), and fracture classification was not related to diagnosis (17). Finally, 39 articles were included in this present systematic review and meta-analysis.

Characteristics of the included studies

Tables 1 and 2 show the detailed study characteristics of the 39 studies (53 trials), which were published between 2013 and 2020. X-rays [43–75] and CT [76–81] were used as inputs for medical images while DEXA was only used in some X-ray studies [51, 61, 65, 66]. Thirteen of 17 trials were multicenter studies [43, 45, 46, 48, 50, 52, 54, 56, 62, 71, 72, 74] while the remaining 26 of 36 trials were single-center studies [43, 44, 47, 49, 51, 53, 55, 57–61, 63–70, 73, 75–81], of which one study included both single-center and multicenter trials [43]. In terms of the applied algorithm, 36 of 39 studies focused on deep learning [43–64, 66–75, 78–81] and 3 used classical machine learning [65, 76, 77]. Fifteen studies had balanced training sets [46–48, 55, 57–60, 63, 68, 71–73, 75, 77] while the rest had unbalanced training sets [43–45, 49–54, 56, 61, 62, 64–67, 69, 70, 74, 76, 78–81]. Seventeen studies applied transfer learning [46, 47, 49, 50, 55, 57, 59–61, 63, 64, 68, 71–73, 75, 81] while the remaining 22 studies did not [43–45, 48, 51–54, 56, 58, 62, 65–67, 69, 70, 74, 76–80]. Twenty-two studies used data augmentation [43–49, 52, 55–59, 63, 64, 66, 67, 69–71, 73, 75] while the remaining studies used only raw and unamplified data [50, 51, 53, 54, 60–62, 65, 68, 72, 74, 76–81]. The number of enrolled patients across all studies was

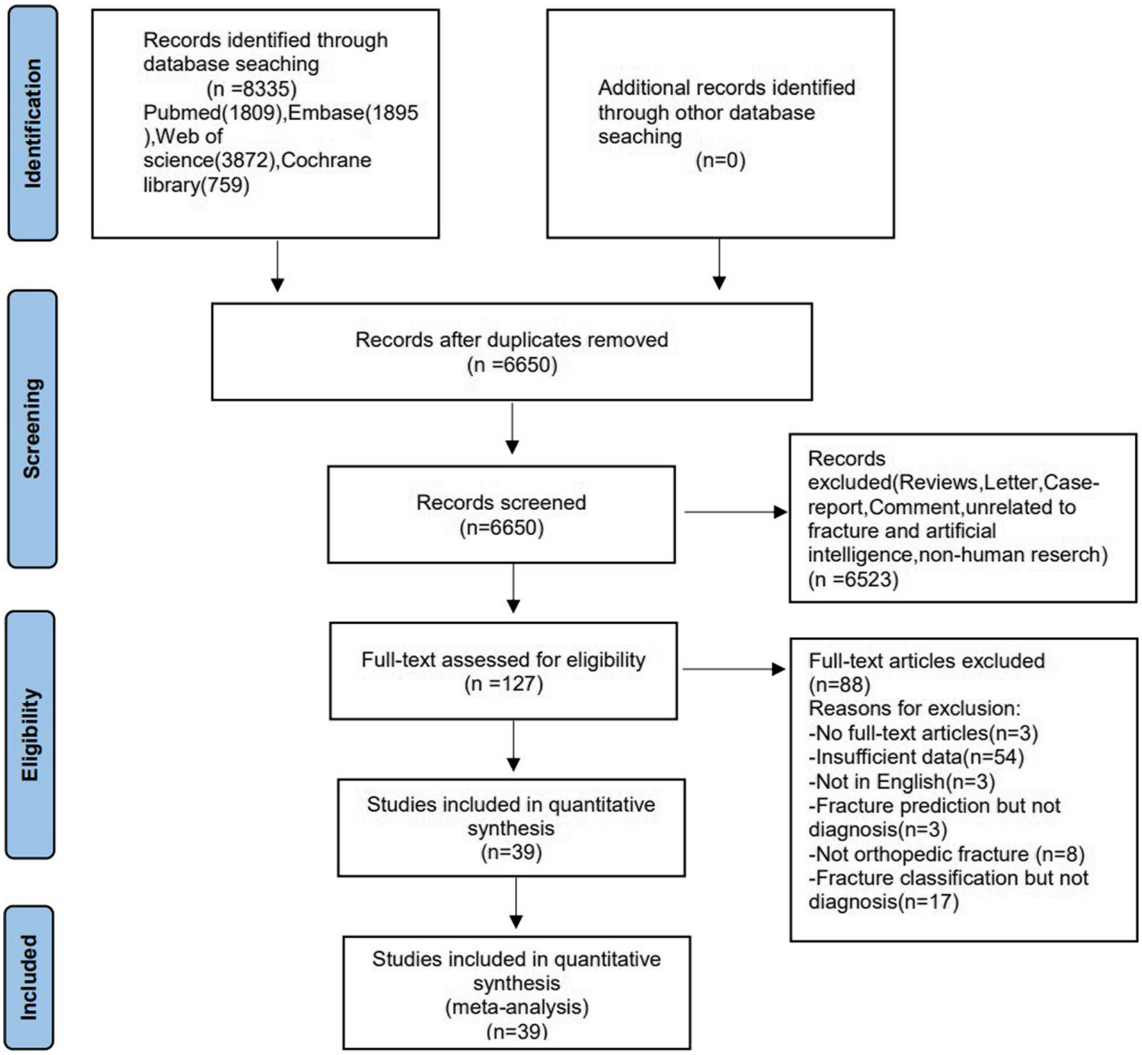

Identification
Records identified through database seaching (n =8335) Pubmed(1809),Embase(1895),Web of science(3872),Cochrane library(759)
Additional records identified through othor database seaching (n=0)
Screening
Records after duplicates removed (n =6650)
Records screened (n=6650)
Records excluded(Reviews,Letter,Case-report,Comment,unrelated to fracture and artificial intelligence,non-human reserch) (n =6523)
Eligibility
Full-text assessed for eligibility (n =127)
Full-text articles excluded (n=88)
Reasons for exclusion:
-No full-text articles(n=3)
-Insufficient data(n=54)
-Not in English(n=3)
-Fracture prediction but not diagnosis(n=3)
-Not orthopedic fracture (n=8)
-Fracture classification but not diagnosis(n=17)
Studies included in quantitative synthesis (n=39)
Included
Studies included in quantitative synthesis (meta-analysis) (n=39)

Fig. 1 PRISMA flow chart of the literature retrieval. Flow diagram of the study selection process for this systematic review and meta-analysis

464,478, ranging from 50 to 327,612 patients across individual studies.

Quality assessment of the studies

The risk of bias and applicability concerns was assessed using the QUADAS-2 criteria (Fig. 2). In the patient selection domain, 11 studies were considered to have a high risk of bias due to non-consecutive patient selection [60, 61], case-control designs [58, 64], and inappropriate exclusions [43, 47, 48, 53, 63, 72, 75]. In the index test domain, all studies were considered to have a low risk of bias because the ground truth was blinded to the machine and a prespecified threshold was used. In the reference standard domain, 34 studies were considered to have a low risk of bias because they used the opinions based on CT or MRI of radiologists [43, 45–47, 49–61, 63–75, 78–81], whereas the others were considered unclear because they did not mention the gold standard [44, 48, 62, 76, 77]. In the flow and timing domain, all studies were considered to have a low risk of bias [43–81]. In the index test domain for concern of applicability, 37 studies that performed internal validation with a temporal split or external validation were considered to have a low concern of applicability [43, 45–76, 78–81], whereas the others that used internal validation with a random split were considered to have an unclear concern of applicability [44, 77]. In the patient selection and reference standard domains, all studies were considered to have low concern of applicability. The overall risk of bias of the included studies was determined to be low.

Pooled detectability of AI performance in diagnosing fractures

For all of the 39 included studies, Spearman's correlation coefficient of heterogeneity caused by the threshold effect was 0.11, meaning that the threshold effect was not significant and the data could be combined.

Table 1 Study characteristics, patient demographics, and diagnostic test criteria of the included studies

Author year	Country	Study design	Medical images	Study cohort and ray sources		Population describe		Fracture part	Model	Gold standard
				Train set	Test set	Mean age	Male:female			
Al-Helo et al (2013) [76]	USA	Single-center retro	CT	From a collaborating radiology center		Unclear	Unclear	Lumber	K-means	Unclear
Bae et al (2021) [43]	Korea	Multicenter retro	X-ray	Hospital A: Seoul		Fracture: 75.7 Normal: 46.4	1796:2395	Femoral neck	Resnet-18	Two emergency medicine specialists
				Hospital A	Hospital B: Gyeonggi-do					
				Hospital A + hospital B, 1/2005–12/2018						
Beyaz et al (2020) [44]	Turkey	Single-center retro	X-ray	Baskent University Adana Turgut Noyan, 1/2013–1/2018		74.9	32:33	Femoral neck	CNN+Gas	Unclear
Blüthgen et al (2020) [45]	Switzerland	Single-center retro	X-ray	University Hospital Zurich, 4/2017–7/2017	University Hospital Zurich, 4/2017–7/2017	Unclear	Unclear	Radius	CNN	Two radiology residents
					MURA dataset	Unclear	Unclear		CNN	
Burns et al (2017) [77]	USA	Single-center retro	CT	University of California, 2012–2015		73	98:42	Lumber	Unclear	Manually annotated data set
Cheng et al (2019) [47]	China	Single-center pro	X-ray	Chang Gung Memorial Hospital, Taiwan, 1/2012–12/2016	Chang Gung Memorial Hospital, Taiwan, 2017	Fracture: 72.34 Normal: 44.88	81941:1664	Hip	DenseNet-121	Radiologists
Cheng et al (2020) [46]	China	Multicenter pro	X-ray	CGMH, Linkou, 8/2008–12 2016	CGMH, Linkou CGMH, and Kaohsiung CGMH, 3/2019–8/2019	Unclear	Unclear	Hip	DenseNet-121	Clinical information
Choi et al (2020) [49]	Korea	Multicenter retro	X-ray	Seoul National University Hospital, 1/2013–12/2017	Seoul National University Hospital, 1/2018–12/20-17 Gyeongsang National University Changwon Hospital, 1/2016–12/20-18	Unclear	Unclear	Pediatric supracon-dylar fracture	Resnet-50	Two pediatric radiologists
Choi et al (2021) [48]	China	Multicenter retro	X-ray	Chang Gung Memorial Hospital's (CGMH), 8/2008–12/2016	CGMH (n = 250)	Unclear	Unclear	Hip	X-ception	Unclear

Table 1 (continued)

Author year	Country	Study design	Medical images	Study cohort and ray sources		Population describe		Fracture part	Model	Gold standard
				Train set	Test set	Mean age	Male:female			
					Stanford ($n = 250$)	Unclear	Unclear			
Chung et al (2018) [50]	Korea	Multicenter retro	X-ray	Several large hospitals in Korea		65	Unclear	Proximal humerus	Resnet-152	Two shoulder orthopedic specialists
Derkatch et al (2019) [51]	Canada	Single-center retro	DEXA	Province of Manitoba BMD Program, February 2010–12/2017		Fracture: 76.9 Normal: 75.4	226:3596	Hip	InceptionRes-NetV2+ DenseNet	Four physicians
Gan et al (2019) [52]	China	Multicenter retro	X-ray	Medical Center of Ningbo City, Lihuili Hospital, of the Ningbo University School, 1/2010–9/2017		48	1366:974	Radius	Faster R-CNN	Orthopedists
Guy et al (2021) [53]	USA	Single-center retro	X-ray	Nstitut du Mouvement et de l'appareil Locomoteur, 270, boulevard de Sainte Marguerite, 13009 Marseille, 1/2015–July 2018		Unclear	Unclear	Femoral neck	Lobe neuronal network	An orthopedic surgeon
						Unclear	Unclear	Trochanteric		
Hendrix et al (2021) [54]	Netherland	Multicenter retro	X-ray	Jeroen Bosch Ziekenhuis,12/2018–03/2019, Radboudumc, 01/200–04/2019	Jeroen Bosch Ziekenhuis, 03/2011–4/2020	Unclear	1287:1519	Scaphoid	DenseNet-121	A specialist and a musculoskeletal radiologist
Hu et al (2021) [78]	China	Single-center retro	CT	Ningbo Third Hospital		Unclear	Unclear	Rib	SGANet	Two attending doctors
Jiménez-Sánchez et al (2020) [55]	Spain	Single-center retro	X-ray	Rechts derIsar Hospital in Munich, 2007–2017		75.7	242:538	Hip	CNN	Two trauma surgeons and one senior radiologist
Jones et al (2020) [56]	USA	multicenter retro	X-ray	15 hospitals and outpatient care centers		Train set: 54 Test set: 75.4	143449:184163	Fractures all over the body	Dilated Residual Network architecture	Orthopedic surgeons and radiologistsphysicians
Kim et al (2021) [57]	Korea	Single-center retro	X-ray	Hallym University Sacred Heart Hospital, 1/2018–3/2020		42.1	1332:1277	Radio-ulnar	DenseNet-161	Dual radiological reporting
Kitamura et al (2019) [58]	USA	Single-center retro	X-ray	University of Pittsburgh Medical Center (UPMC), 200 Lothrop St., Pittsburgh, PA 15213		Unclear	Unclear	Ankle	Inception-v3+ Resnet+Resnet with drop/aux+ X-ception+ X-ception with drop/aux	A radiologist and radiology resident
						Unclear	Unclear		Inception-v3+ Resnet+ X-ception	
Krogue et al (2020) [59]	USA	Single-center retro	X-ray	University of California, San Francisco, 1998–2017		75.2	1162:1864	Hip	DenseNet	CT and MRI

Table 1 (continued)

Author year	Country	Study design	Medical images	Study cohort and ray sources		Population describe		Fracture part	Model	Gold standard
				Train set	Test set	Mean age	Male:female			
Langerhuizen et al (2020) [60]	Netherland	Single-center retro	X-ray	Amsterdam Movement Sciences (AMS) Amsterdam University		Unclear	Unclear	Scaphoid	CNN	CT and MRI
Li et al (2021) [61]	China	Single-center retro	DEXA	Taipei Veterans General Hospital, 2016–2018		76	Unclear	Lumber	ResNet34+ DenseNet121+ DenseNet201	CT and MRI
Ma et al (2021) [62]	China	Single-center retro	X-ray	Website Radiopaedia and Haikou People's Hospital		Unclear	Unclear	Five major parts of bones	Faster R-CNN +CrackNet	Unclear
MacKinnon et al (2018) [63]	UK	Single-center retro	X-ray	Royal Devon and Exeter Hospital, 1/2015–1/2016		Unclear	Unclear	Wrist	Inception v3	Radiological report
Mawatari et al (2020) [64]	Japan	Single-center retro	X-ray	Unclear		Train set: 81 Test set: 84	82:259	Hip	DCNN with the GoogLeNet	Three radiologists
Mehta et al (2020) [65]	USA	Single-center retro	DEXA	University of Pennsylvania, 1/2010–April 2018		Fracture: 70.79 Normal: 67.29	105:202	Lumber	SVM, linear SVM, radial basis function SVM, sigmoid SVM, cubic polynomial	CT and MRI
Monchka et al (2021) [66]	Canada	Single-center retro	DEXA	Manitoba Bone Mineral Density Registry, 2/2010–12/2017		75.8	498:8422	Lumber	Inception-Res-Net-v2 +DenseNet	Four expert physician readers
Mutasa et al (2020) [67]	USA	Single-center retro	X-ray	Columbia University, February 2000–2/2017		75	198:352	Femoral neck	CNN	A fellowship trained MSK radiologist
Ozkaya et al (2020) [68]	Turkey	Single-center retro	X-ray	Ataturk Training and Research Hospital, 2014–2020		42	Unclear	Scaphoid	Resnet-50	A radiologist
Rayan et al (2019) [69]	USA	Single-center retro	X-ray	A tertiary care children's center, 1/2014–12/2017		7.2	9630:8279	Pediatric elbow fractures	X-ception	Two senior radiology residents
Reichert et al (2021) [70]	Switzerland	Single-center retro	X-ray	Unclear	Louis Mourier ER, 3/2019	Unclear	Unclear	Foot, hand, wrist, ankle, femur, clavicle, shoulder	RetinaNet	Radiologists

Table 1 (continued)

Author year	Country	Study design	Medical images	Study cohort and ray sources		Population describe		Fracture part	Model	Gold standard
				Train set	Test set	Mean age	Male:female			
Ren et al (2021) [71]	USA	Single-center retro	X-ray	MURA dataset; the public domain, the LERA dataset	Johns Hopkins University	Unclear	Unclear	Triquetrum	DCNN	A member of the research team and a radiologist
								Segond	DCNN	
Sato et al (2021) [72]	Japan	Multicenter retro	X-ray	Gamagori City Hospital, Tsushima City Hospital, and Nagoya Daini Red Cross Hospital		81.1	1193:3658	Hip	EfficientNet-B4	Two orthopedic surgeons
Small et al (2021) [79]	USA	Single-center retro	CT	Lahey Hospital and Medical Center, 1/2015–12/2018		60.28	379:316	Cervical	Aidoc	Two fellowship-trained neuroradiologists
Urakawa et al (2019) [73]	Japan	Single-center retro	X-ray	Tsuruoka Municipal Shonai Hospital, 1/2006–7/2017		85	Unclear	Interchan-teric hip	VGG_16	A single board-certified orthopedic surgeon
Voter et al (2021) [80]	USA	Single-center retro	CT	University of Wisconsin, 1/2020–10/2020		60	958:946	Cervical	Aidoc	Neuroradiologist
Weikert et al (2020) [81]	Switzerland	Single-center retro	CT	University Hospital Basel, University of Basel, Basel, 2018		58.4	Unclear	Rib	ResNet+Fast Region-based CNN	Clinically approved written CT reports
Yoon et al (2021) [74]	China	Multicenter retro	X-ray	Chang Gung Memorial Hospital and Michigan Medicine, 1/2001–12/2019		Unclear	Unclear	Scaphoid	DCNN mod	Surgeon's interpretation
Yu et al (2020) [75]	USA	Single-center retro	X-ray	The Ohio State University, a 48-month period		Fracture: 69.4 Normal: 62.0	306:311	Hip	Inception-V3	A board-certified musculoskeletal radiologist

pro prospective, *retro* retrospective, *DEXA* dual-energy X-ray absorptiometry

Table 2 Diagnostic accuracy test results from studies included in the meta-analysis

Author year	No. of patients	TP	FP	FN	TN	Multicenter	Deep learning	Train set balance	Transfer learning	Data augmentation	Medical image type	Risk of bias	Localization of fractures	One vs more than one type of fracture
Al-Helo et al (2013) [76]	50	21	2	3	224	No	No	No	No	No	CT	Low	Yes	1
Bae et al (2021) [43]	2090	57	1	2	150	No	Yes	No	No	Yes	Ordinary plain X-ray	High	Yes	1
	3979	488	29	32	1550	Yes	Yes	No	No	Yes	Ordinary plain X-ray	High	Yes	1
	4189	108	4	3	305	Yes	Yes	No	No	Yes	Ordinary plain X-ray	High	Yes	1
Beyaz et al (2020) [44]	65	1111	207	230	558	No	Yes	No	No	Yes	Ordinary plain X-ray	Low	No	1
Blüthgen et al (2020) [45]	258	41	6	1	52	Yes	Yes	No	No	Yes	Ordinary plain X-ray	Low	Yes	1
	258	78	18	22	82	Yes	Yes	No	No	Yes	Ordinary plain X-ray	Low	Yes	1
Burns et al (2017) [77]	150	74	17	1	58	No	No	Yes	No	No	CT	Low	No	1
Cheng et al (2019) [47]	3605	49	8	1	42	No	Yes	Yes	Yes	Yes	Ordinary plain X-ray	High	Yes	1
Cheng et al (2020) [46]	3605	243	19	24	301	Yes	Yes	Yes	Yes	Yes	Ordinary plain X-ray	Low	Yes	1
Choi et al (2020) [49]	810	62	15	4	177	No	Yes	No	Yes	Yes	Ordinary plain X-ray	Low	Yes	1
	810	23	10	0	62	No	Yes	No	Yes	Yes	Ordinary plain X-ray	Low	Yes	1
Choi et al (2021) [48]	4735	115	6	10	119	Yes	Yes	Yes	No	Yes	Ordinary plain X-ray	High	Yes	1
	4735	126	7	14	103	Yes	Yes	Yes	No	Yes	Ordinary plain X-ray	High	Yes	1
Chung et al (2018) [50]	1891	131	1	1	49	Yes	Yes	No	Yes	No	Ordinary plain X-ray	Low	No	1
Derkatch et al (2019) [51]	12742	534	373	77	2838	No	Yes	No	No	No	DEXA	Low	Yes	1
Gan et al (2019) [52]	2340	135	6	15	144	Yes	Yes	No	No	Yes	Ordinary plain X-ray	Low	No	1
Guy et al (2021) [53]	623	238	213	153	443	No	Yes	No	No	No	Ordinary plain X-ray	High	Yes	1
	623	256	202	127	462	No	Yes	No	No	No	Ordinary plain X-ray	High	Yes	1
Hendrix et al (2021) [54]	2811	74	15	21	80	Yes	Yes	No	No	No	Ordinary plain X-ray	Low	Yes	1
Hu et al (2021) [78]	1697	80	36	8	128	No	Yes	No	No	No	CT	Low	Yes	1

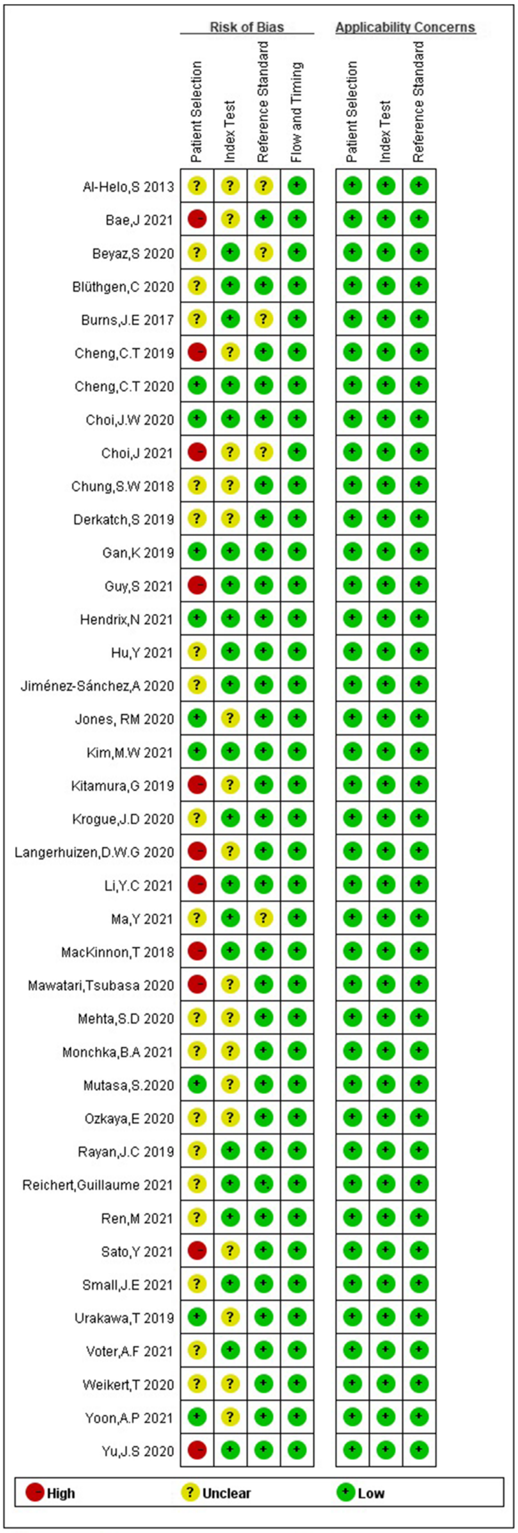
Table 2 (continued)

Author year	No. of patients	TP	FP	FN	TN	Multicenter	Deep learning	Train set balance	Transfer learning	Data augmentation	Medical image type	Risk of bias	Localization of fractures	One vs more than one type of fracture
Jiménez-Sánchez et al (2020) [55]	780	108	8	7	107	No	Yes	Yes	Yes	Yes	Ordinary plain X-ray	Low	Yes	1
Jones et al (2020) [56]	327612	2299	2544	116	11060	Yes	Yes	No	No	Yes	Ordinary plain X-ray	Low	Yes	> 1
Kim et al (2021) [57]	2609	271	66	29	624	No	Yes	Yes	Yes	Yes	Ordinary plain X-ray	Low	Yes	> 1
Kitamura et al (2019) [58]	596	32	7	8	33	No	Yes	Yes	No	Yes	Ordinary plain X-ray	High	No	1
	596	29	5	11	35	No	Yes	Yes	No	Yes	Ordinary plain X-ray	High	No	1
Krogue et al (2020) [59]	1118	203	13	15	207	No	Yes	Yes	Yes	Yes	Ordinary plain X-ray	Low	Yes	1
Langerhuizen et al (2020) [60]	300	42	20	8	30	No	Yes	Yes	Yes	No	Ordinary plain X-ray	High	No	1
Li et al (2021) [61]	941	129	45	12	644	No	Yes	No	Yes	No	DEXA	High	Yes	1
	941	75	70	4	567	No	Yes	No	Yes	No	DEXA	High	Yes	1
Ma et al (2021) [62]	3053	425	48	45	422	Yes	Yes	No	No	No	Ordinary plain X-ray	Low	Yes	> 1
	3053	49	6	7	50	Yes	Yes	No	No	No	Ordinary plain X-ray	Low	Yes	> 1
MacKinnon et al (2018) [63]	1389	45	6	5	44	No	Yes	Yes	Yes	Yes	Ordinary plain X-ray	High	No	> 1
Mawatari et al (2020) [64]	341	22	7	3	18	No	Yes	No	Yes	Yes	Ordinary plain X-ray	High	No	1
Mehta et al (2020) [65]	415	18	1	4	38	No	No	No	No	No	DEXA	Low	Yes	1
	415	18	0	4	39	No	No	No	No	No	DEXA	Low	Yes	1
	415	19	4	2	35	No	No	No	No	No	DEXA	Low	Yes	1
	415	13	0	9	39	No	No	No	No	No	DEXA	Low	Yes	1
Monchka et al (2021) [66]	12742	532	181	114	2995	No	Yes	No	No	Yes	DEXA	Low	Yes	1
	12742	568	400	78	2776	No	Yes	No	No	Yes	DEXA	Low	Yes	1
Mutasa et al (2020) [67]	550	63	2	7	33	No	Yes	No	No	Yes	Ordinary plain X-ray	Low	Yes	1
Ozkaya et al (2020) [68]	390	38	4	12	46	No	Yes	Yes	Yes	No	Ordinary plain X-ray	Low	No	1

Table 2 (continued)

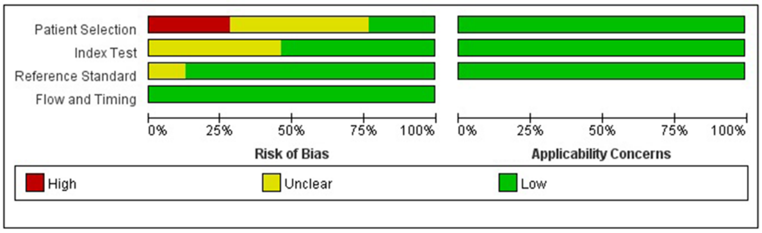
Author year	No. of patients	TP	FP	FN	TN	Multicenter	Deep learning	Train set balance	Transfer learning	Data augmentation	Medical image type	Risk of bias	Localization of fractures	One vs more than one type of fracture
Rayan et al (2019) [69]	21456	536	82	54	434	No	Yes	No	No	Yes	Ordinary plain X-ray	Low	No	1
Reichert et al (2021) [70]	125	24	14	1	86	No	Yes	No	No	Yes	Ordinary plain X-ray	Low	Yes	> 1
Ren et al (2021) [71]	684	24	3	1	22	Yes	Yes	Yes	Yes	Yes	Ordinary plain X-ray	Low	Yes	1
	684	11	1	1	11	Yes	Yes	Yes	Yes	Yes	Ordinary plain X-ray	Low	Yes	1
Sato et al (2021) [72]	4851	476	15	24	485	Yes	Yes	Yes	Yes	No	Ordinary plain X-ray	High	Yes	1
Small et al (2021) [79]	665	109	17	34	505	No	Yes	No	No	No	CT	Low	Yes	1
Urakawa et al (2019) [73]	1773	169	4	11	150	No	Yes	Yes	Yes	Yes	Ordinary plain X-ray	Low	No	1
Voter et al (2021) [80]	1904	67	106	55	1676	No	Yes	No	No	No	CT	Low	No	1
Weikert et al (2020) [81]	511	139	30	20	321	No	Yes	No	Yes	No	CT	Low	Yes	1
Yoon et al (2021) [74]	7729	806	108	119	1271	Yes	Yes	No	No	No	Ordinary plain X-ray	Low	Yes	1
Yu et al (2020) [75]	617	82	4	2	118	No	Yes	Yes	Yes	Yes	Ordinary plain X-ray	High	Yes	1

TP true positive, *FP* false positive, *FN* false negative, *TN* true negative, *DEXA* dual-energy X-ray absorptiometry



	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Al-Helo,S 2013	?	?	?	+	+	+	+
Bae,J 2021	−	?	+	+	+	+	+
Beyaz,S 2020	?	+	?	+	+	+	+
Blüthgen,C 2020	?	+	+	+	+	+	+
Burns,J.E 2017	?	+	?	+	+	+	+
Cheng,C.T 2019	−	?	+	+	+	+	+
Cheng,C.T 2020	+	+	+	+	+	+	+
Choi,J.W 2020	+	+	+	+	+	+	+
Choi,J 2021	−	?	?	+	+	+	+
Chung,S.W 2018	?	?	+	+	+	+	+
Derkatch,S 2019	?	?	+	+	+	+	+
Gan,K 2019	+	+	+	+	+	+	+
Guy,S 2021	−	+	+	+	+	+	+
Hendrix,N 2021	+	+	+	+	+	+	+
Hu,Y 2021	?	+	+	+	+	+	+
Jiménez-Sánchez,A 2020	?	+	+	+	+	+	+
Jones, RM 2020	+	?	+	+	+	+	+
Kim,M.W 2021	+	+	+	+	+	+	+
Kitamura,G 2019	−	?	+	+	+	+	+
Krogue,J.D 2020	?	+	+	+	+	+	+
Langerhuizen,D.W.G 2020	−	?	+	+	+	+	+
Li,Y.C 2021	−	+	+	+	+	+	+
Ma,Y 2021	?	+	?	+	+	+	+
MacKinnon,T 2018	−	+	+	+	+	+	+
Mawatari,Tsubasa 2020	−	?	+	+	+	+	+
Mehta,S.D 2020	?	?	+	+	+	+	+
Monchka,B.A 2021	?	?	+	+	+	+	+
Mutasa,S.2020	+	?	+	+	+	+	+
Ozkaya,E 2020	?	?	+	+	+	+	+
Rayan,J.C 2019	?	+	+	+	+	+	+
Reichert,Guillaume 2021	?	+	+	+	+	+	+
Ren,M 2021	?	+	+	+	+	+	+
Sato,Y 2021	−	?	+	+	+	+	+
Small,J.E 2021	?	+	+	+	+	+	+
Urakawa,T 2019	+	?	+	+	+	+	+
Voter,A.F 2021	?	+	+	+	+	+	+
Weikert,T 2020	?	?	+	+	+	+	+
Yoon,A.P 2021	+	?	+	+	+	+	+
Yu,J.S 2020	−	+	+	+	+	+	+

− High ? Unclear + Low



◀ **Fig. 2** Methodological quality assessment of the included studies using the QUADAS-2 tool. The methodological quality of the included studies was assessed according to the Quality Assessment of Diagnostic Accuracy Studies 2 tool for risk of bias and applicability concerns. Green represents low, yellow circle unclear, and red high risk of bias

The pooled sensitivity and specificity of the detectability of AI for diagnosing orthopedic fractures were 90% (95% CI 87–92%) and 92% (95% CI 90–94%), respectively (Fig. 3). The pooled positive likelihood ratio, negative likelihood ratio, and diagnostic odds ratio were 11.0 (95% CI 8.5–14.1), 0.11 (95% CI 0.09–0.14), and 100 (95% CI 66–150), respectively (Table 3). The overall pooled AUC was 0.96 (95% CI 94–98%), which indicated a high diagnostic performance (Fig. 4).

Cochran's Q test showed that heterogeneity was present (Q = 665.744, p < 0.001) across the studies, and the Higgins I^2 statistic demonstrated that heterogeneity was noticed in both sensitivity (I^2 = 96.52%, p < 0.001) and specificity (I^2 = 98.12%, p < 0.001) computations.

Deek's test was performed for the assessment of publication bias. The funnel plot for assessing publication bias was almost symmetrical, and the coefficient of bias demonstrated a p value of 0.21 (> 0.05), which further validated the presence of a low publication bias (Fig. 5).

Comparison of AI with human readers on orthopedic fracture diagnosis

In 16 of the included studies, the performance of AI was compared with human readers (n = 120) for the diagnosis of orthopedic fractures [45–47, 49, 50, 52, 54, 55, 59, 60, 64, 68, 72, 73, 75, 78]. The pooled sensitivity and specificity for all human readers on orthopedic fracture diagnosis was 90% (95% CI 85–93%) and 95% (95% CI 93–96%), respectively, with a corresponding AUC of 0.97 (95% CI 0.96–0.99). AI achieved comparable results to humans for diagnosing orthopedic fractures (AUC = 0.97, 95% CI 0.95–0.98) (Fig. 6a and Supplementary Fig. S1).

Among the 11 studies that included non-expert human readers (n = 68) [45–47, 49, 50, 52, 54, 59, 68, 72, 75], AI was superior than the non-expert human readers for diagnosing orthopedic fractures (AUC = 0.98 vs. AUC = 0.96,

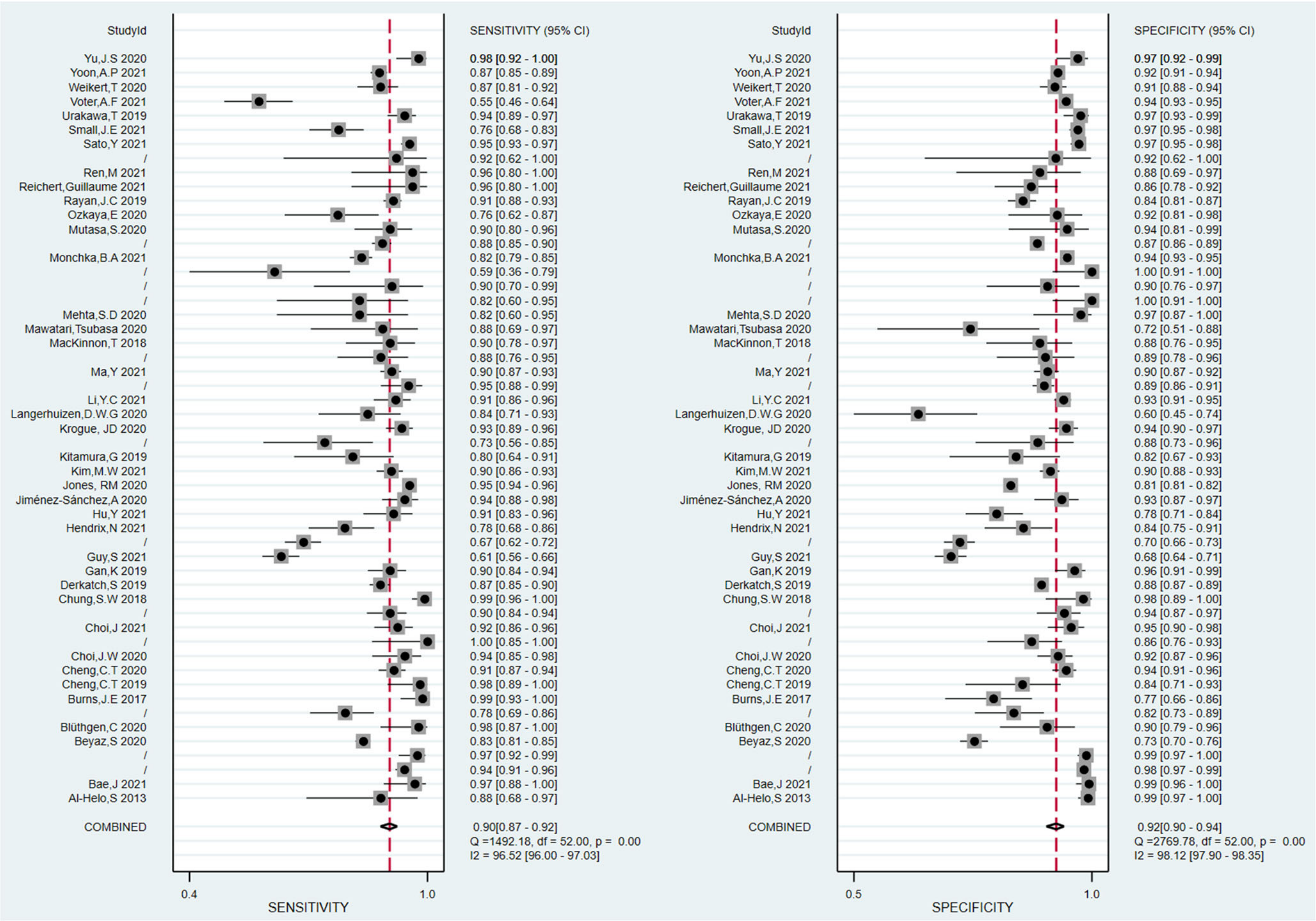


Fig. 3 Forest plots. Forest plots of the pooled sensitivity and specificity for the diagnostic performance of artificial intelligence for the diagnosis of orthopedic fractures. The numbers are pooled estimates with 95% CIs in parentheses; horizontal lines indicate 95% CIs

Table 3 Results of multiple subgroup analyses of artificial intelligence for diagnosis of orthopedic fractures

Analysis	No. of trials	No. of patients	Sensitivity	I^2 (%)	Specificity	I^2 (%)	PLR	NLR	Diagnostic odds ratio	AUC
Overall group	53	464478	0.90 [0.87, 0.92]	96.52	0.92 [0.90, 0.94]	98.12	11.0 [8.5, 14.1]	0.11 [0.09, 0.14]	100 [66, 150]	0.96 [0.94, 0.98]
Multicenter or Single-center										
Multicenter	17	376467	0.92 [0.89, 0.95]	94.39	0.94 [0.91, 0.96]	99.10	14.6 [9.6, 22.4]	0.08 [0.06, 0.12]	178 [89, 357]	0.97 [0.96, 0.99]
Single-center	36	88161	0.88 [0.85, 0.91]	95.55	0.91 [0.88, 0.93]	97.34	9.7 [7.1, 13.2]	0.13 [0.10, 0.17]	76 [47, 123]	0.95 [0.93, 0.97]
Algorithm										
Deep learning	47	462618	0.90 [0.87, 0.92]	96.72	0.91 [0.89, 0.93]	98.22	10.3 [7.9, 13.3]	0.11 [0.09, 0.14]	93 [60, 145]	0.96 [0.94, 0.97]
Classical learning	6	1860	0.88 [0.73, 0.95]	83.49	0.98 [0.90, 1.00]	90.80	43.3 [9.3, 200.3]	0.13 [0.06, 0.28]	345 [102, 1162]	0.98 [0.96, 0.99]
Train set balance										
Balance	18	33217	0.92 [0.89, 0.94]	75.18	0.91 [0.88, 0.94]	87.07	10.7 [7.4, 15.6]	0.09 [0.07, 0.13]	117 [64, 214]	0.97 [0.95, 0.98]
Imbalance	35	431261	0.89 [0.85, 0.92]	97.62	0.92 [0.89, 0.94]	98.66	11.1 [7.9, 15.7]	0.12 [0.09, 0.16]	92 [54, 157]	0.96 [0.94, 0.97]
Transfer learning										
With	20	28650	0.92 [0.90, 0.94]	66.52	0.92 [0.89, 0.94]	86.36	11.5 [8.1, 16.3]	0.08 [0.06, 0.11]	140 [78, 252]	0.97 [0.95, 0.98]
Without	33	435828	0.87 [0.84, 0.90]	97.47	0.92 [0.89, 0.94]	98.53	10.8 [7.6, 15.5]	0.14 [0.10, 0.18]	79 [46, 134]	0.95 [0.93, 0.97]
Data augmentation										
With	30	417893	0.92 [0.90, 0.93]	94.93	0.92 [0.89, 0.94]	98.38	11.5 [8.4, 15.8]	0.09 [0.07, 0.12]	127 [78, 206]	0.97 [0.95, 0.98]
Without	23	46585	0.87 [0.81, 0.91]	96.31	0.92 [0.87, 0.95]	97.85	10.4 [6.7, 16.1]	0.15 [0.10, 0.21]	71 [36, 137]	0.95 [0.93, 0.97]
Medical image										
Ordinary plain X-ray	38	417733	0.91 [0.88, 0.93]	97.23	0.91 [0.88, 0.93]	98.18	10.3 [7.5, 14.0]	0.10 [0.08, 0.13]	102 [59, 175]	0.96 [0.94, 0.98]
DEXA	9	41768	0.84 [0.80, 0.88]	78.51	0.93 [0.89, 0.96]	94.19	12.8 [7.9, 20.7]	0.17 [0.13, 0.21]	76 [55, 106]	0.94 [0.91, 0.96]
CT	6	4977	0.86 [0.71, 0.94]	94.21	0.93 [0.84, 0.97]	96.41	11.8 [5.6, 25.1]	0.15 [0.07, 0.32]	80 [33, 193]	0.96 [0.94, 0.97]
Risk of bias										
High	17	35151	0.90 [0.85, 0.94]	97.52	0.92 [0.86, 0.95]	98.64	10.9 [6.0, 19.6]	0.11 [0.07, 0.17]	100 [36, 278]	0.96 [0.94, 0.97]
Low	36	429327	0.89 [0.86, 0.92]	95.01	0.92 [0.89, 0.93]	97.76	10.7 [8.4, 13.6]	0.12 [0.09, 0.15]	92 [65, 131]	0.96 [0.94, 0.97]
Localization of fractures										
Yes	40	431287	0.90 [0.88, 0.92]	96.83	0.93 [0.90, 0.94]	98.59	12.2 [9.2, 16.2]	0.10 [0.08, 0.13]	117 [75, 184]	0.97 [0.95, 0.98]
No	13	33191	0.88 [0.81, 0.93]	95.32	0.88 [0.81, 0.93]	96.03	7.6 [4.6, 12.5]	0.13 [0.08, 0.23]	58 [24, 138]	0.94 [0.92, 0.96]
Type of fracture more										
More than one	6	337841	0.92 [0.89, 0.94]	86.65	0.88 [0.84, 0.91]	96.76	7.6 [5.9, 9.6]	0.09 [0.07, 0.12]	83 [67, 103]	0.96 [0.94, 0.97]
One	47	126637	0.90 [0.87, 0.92]	95.59	0.92 [0.90, 0.94]	97.61	11.6 [8.7, 15.6]	0.11 [0.09, 0.14]	104 [65, 165]	0.96 [0.94, 0.98]

DEXA dual-energy X-ray absorptiometry, *PLR* positive likelihood ratio, *NLR* negative likelihood ratio

sensitivity = 95% vs. sensitivity = 88%) but had comparable specificity with the non-expert human readers (93% vs. 93%) (Fig. 6b and Supplementary Fig. S2). (Expert-level human readers were defined as radiologists, orthopedic surgeons, etc. with at least 5 years of experience in the field of orthopedic fracture diagnosis.)

Two studies of three trials compared the performance of human-algorithm integration systems (AI and a radiologist together), AI, and human readers on fracture recognition [46, 59]. Human-algorithm integration systems achieved non-inferiority results compared with AI, and both achieved better results than human readers (AUC = 0.99 vs. 0.99 vs. 0.97, sensitivity = 98% vs. 99% vs. 94%, specificity = 91% vs. 89% vs. 87%) (Supplementary Fig. S3 and Fig. S4).

Subgroup analysis

Table 3 shows the detailed results of subgroup analyses for exploring the potential source of heterogeneity. After grouping according to whether transfer learning was applied, a significant drop in I^2 was observed in sensitivity (from 96.52 to 66.52%) and specificity (from 98.12 to 86.36%). After grouping according to whether the train set was balanced, a significant drop in I^2 was observed in sensitivity (from 96.52 to 75.18%) and specificity (from 98.12 to 87.07%). Both suggested the use of transfer learning and a balanced train set were the main sources of heterogeneity. Studies with multicenter study design yielded higher sensitivity (92% vs. 88%) and specificity (94% vs. 91%) than single-center study design. Further, utilizing plain X-rays as input images for AI to diagnose fractures achieved results comparable to CT (AUC 0.96 vs. 0.96). Moreover, studies with transfer learning achieved higher sensitivity (92% vs. 87%) and diagnostic odds ratio (140 vs. 79) than studies without transfer learning, and studies with data augmentation demonstrated higher sensitivity (92% vs. 87%) and diagnostic odds ratio (127 vs. 71) than studies without data augmentation.

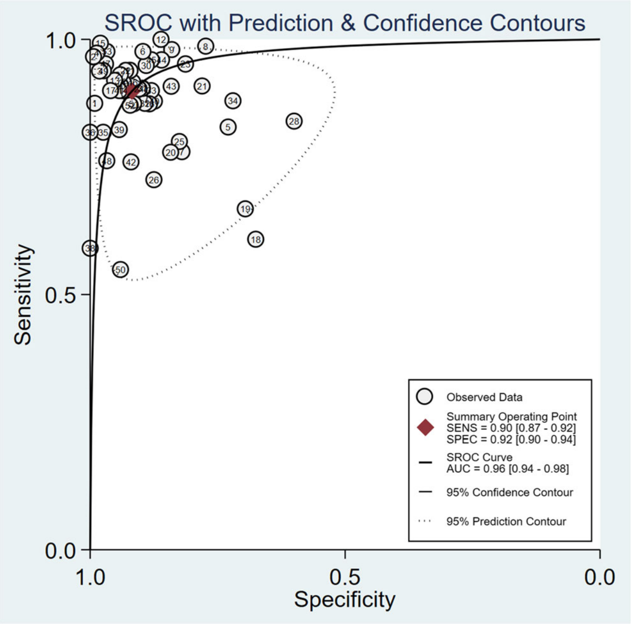


Fig. 4 The SROC curve. The SROC curve for the diagnostic performance of artificial intelligence for the diagnosis of orthopedic fractures

Sensitivity analysis

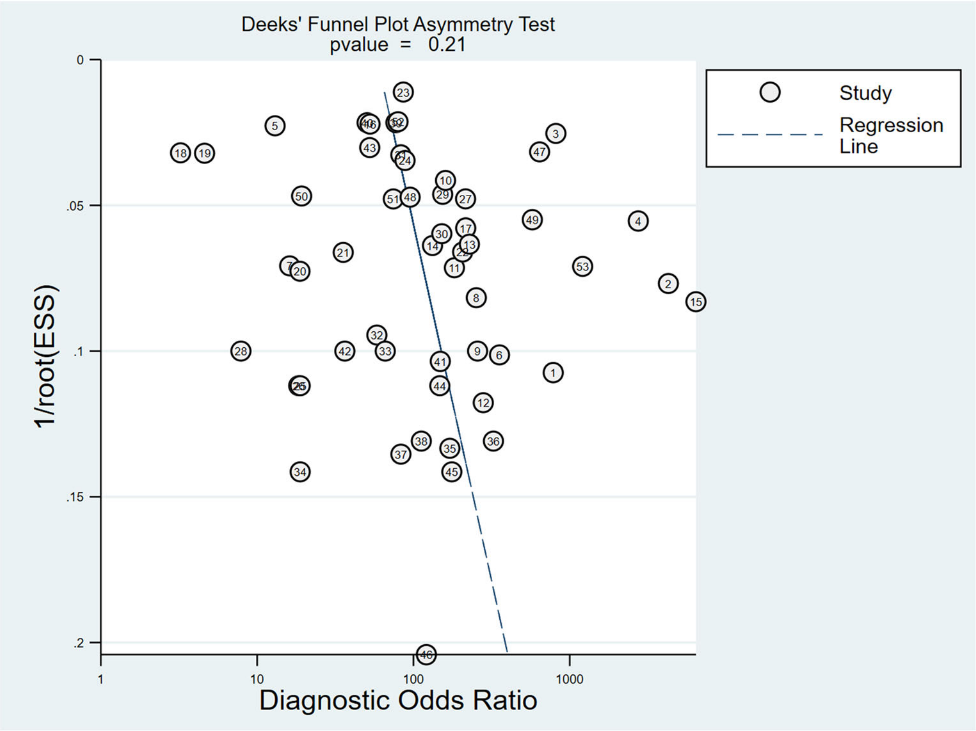
The sensitivity analysis results of AI performance in terms of sensitivity and specificity are shown in Table 4. The results showed that omitting any study had a relatively low influence on the overall combined estimates.

Discussion

There existed one published meta-analysis reporting the diagnostic utility of AI in orthopedic fracture diagnosis [89]. However, obvious differences between our meta-analysis and the study above should be considered. First, this is the first systematic review and meta-analysis exploring up to nine factors affecting model performance (where the multicenter study is adopted, whether the fracture is localized, medical image type, etc.) and comparing AI to experts and non-experts. Second, we conducted a comprehensive literature search from inception to September 29, 2021, and quantitatively analyzed 39 studies (464,478 patients) in total. Third, not only did we have a subgroup analysis to explore heterogeneity, but we also used a sensitivity analysis. Finally, we compared the effect of human-algorithm integration systems, AI, and human readers on fracture recognition.

In subgroup analysis, the use of AI demonstrated better diagnostic performance in multicenter designed studies than those with single-center design. This may be attributed to the greater number of images with different imaging formats and a larger amount of data [72]. In addition to increasing the number of images in the training set, the use of images from different healthcare facilities increased the diversity of the dataset and thus, increased the generalization ability of the model, and demonstrated more reliable results [44]. Single-center AI studies lacked large enough cases and diversity in imaging sources and were more prone to selection bias than multi-institutional datasets [46, 65].

Fig. 5 Deek's funnel plot. Funnel plot of the included studies. (p = 0.21 > 0.05, suggesting a low publication bias)



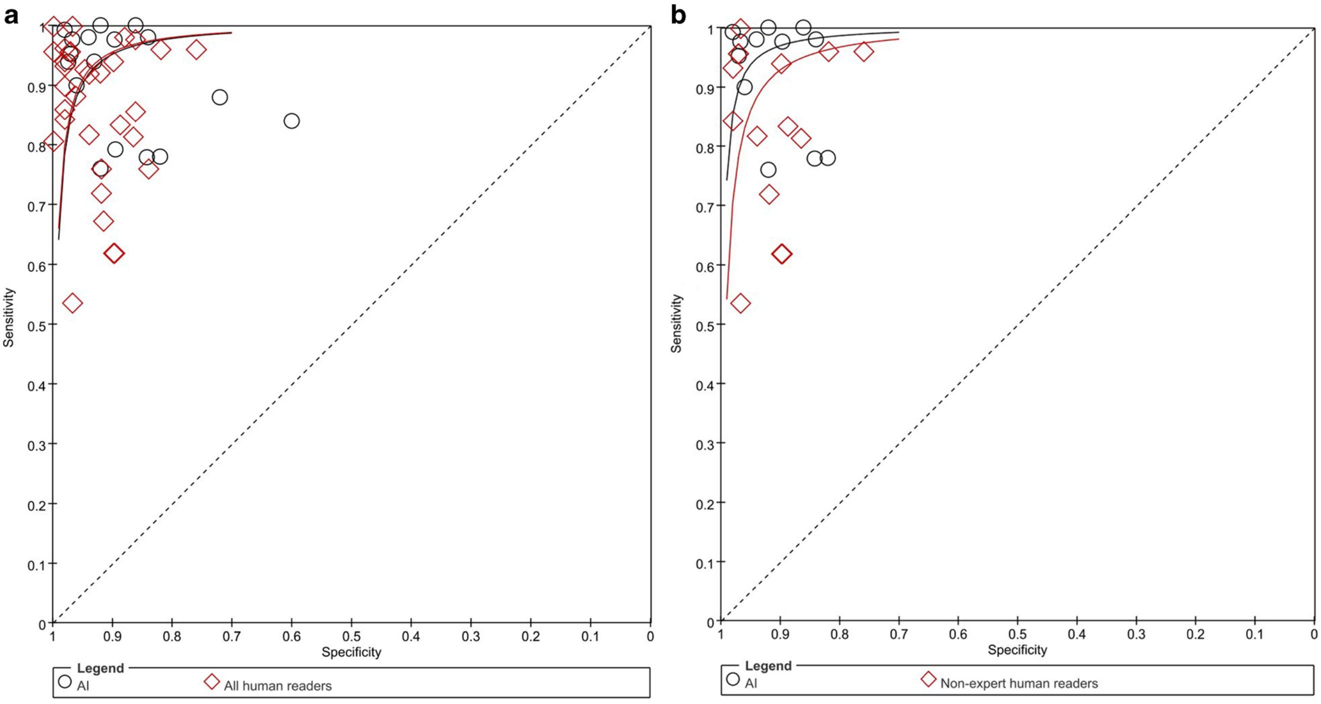


Fig. 6 The SROC curve compares AI with all human readers and AI with non-expert human readers. **a** The SROC curves of the diagnostic performance of artificial intelligence (AI) and all human readers; **b** the SROC curves of the diagnostic performance of artificial intelligence (AI) and non-expert human readers

The subgroup analysis showed that the use of transfer learning and a balanced train set were the main sources of heterogeneity. Transfer learning was closely related to improving model performance, which was concordant with the findings of previous studies [82–84]. Transfer learning was adopted to train the AI model or data for many iterations based on well-known pre-trained models [85]. The improved model performance of transfer learning may also be attributed to the parameters and weights obtained in advance through training on large sample datasets. From our subgroup analysis, studies with a balanced train set achieved higher sensitivity than studies with an unbalanced train set. A balanced dataset means that the number of fractured images is close to that of non-fractured images, which allows the model to learn both types more evenly. In a real-world setting, imbalanced datasets tend to have fewer fracture images, which results in insufficient fracture images for training the model. This may lead to a reduction in the ability for detecting a fracture.

Further, studies with data augmentation achieved better results than studies without data augmentation. Using data augmentation to artificially enlarge a dataset could mitigate the limitations of small datasets and thereby improve the generalizability. Further, despite the risk of imprecisely copying or converting an image, data augmentation provides more training examples by integrating the distinctive features in multiple directions [63]. Considering that any researcher may face the issues pertained in smaller training datasets, it is particularly important to overcome such shortcomings by correctly implementing data augmentation.

Another important finding was that diagnosis using AI achieved comparable results to humans and was superior to non-expert human readers. Additionally, our results showed that human-algorithm integration systems achieved non-inferiority results compared with AI, and both achieved better results than human readers. It indicates that human-algorithm integration systems have the potential to improve the delivery of efficient and high-quality care in massive clinical practice while allowing physicians to focus on more conceptually demanding tasks by offloading their more mundane duties. Additionally, Krogue et al [59] showed that AI and experts together achieved better results compared with experts alone (accuracy 95.5% vs. 93.5%). They also showed that when using the model as an aid, residents and attending physicians improved their performance, with aided residents approximating the performance of fellowship-trained experts. It revealed that AI could be a valuable tool in training human readers to better evaluate radiographs for a fracture.

In addition, plain X-rays remained the most commonly used AI training medical images and achieved comparable results with CT. Plain X-rays were usually the initial diagnostic modality of orthopedic fracture because they were cheap, and readily available [76]. Six studies used CT images as the training set of AI models, focusing on rib [78, 81], thoracolumbar [76, 77], and cervical fracture [79, 80] for

Table 4 Sensitivity analysis for the whole group excluding one study at a time

Study	Sensitivity			Specificity			Diagnostic odds ratio		
	Value	I^2(%)	*p* value	Value	I^2(%)	*p* value	Value	I^2(%)	*p* value
Al-Helo et al (2013) [76]	0.90 [0.87, 0.92]	96.37	< 0.01	0.91 [0.89, 0.93]	98.00	< 0.01	95 [63, 144]	96.20	< 0.01
Bae et al (2021) [43]	0.89 [0.87, 0.91]	96.03	< 0.01	0.91 [0.88, 0.93]	97.55	< 0.01	83 [56, 121]	95.80	< 0.01
Beyaz et al (2020) [44]	0.90 [0.88, 0.92]	96.15	< 0.01	0.92 [0.90, 0.94]	98.06	< 0.01	104 [69, 157]	95.80	< 0.01
Blüthgen et al (2020) [45]	0.90 [0.87, 0.92]	96.51	< 0.01	0.92 [0.90, 0.94]	98.16	< 0.01	102 [67, 155]	96.20	< 0.01
Burns et al (2017) [77]	0.90 [0.87, 0.92]	96.51	< 0.01	0.92 [0.90, 0.94]	98.14	< 0.01	99 [65, 150]	96.20	< 0.01
Cheng et al (2019) [47]	0.90 [0.87, 0.92]	96.50	< 0.01	0.92 [0.90, 0.94]	98.12	< 0.01	99 [65, 150]	96.20	< 0.01
Cheng et al (2020) [46]	0.90 [0.87, 0.92]	96.51	< 0.01	0.92 [0.89, 0.94]	98.10	< 0.01	99 [65, 150]	96.10	< 0.01
Choi et al (2020) [49]	0.90 [0.87, 0.92]	96.48	< 0.01	0.92 [0.90, 0.94]	98.11	< 0.01	98 [64, 150]	96.20	< 0.01
Choi et al (2021) [48]	0.90 [0.87, 0.92]	96.49	< 0.01	0.92 [0.89, 0.94]	98.08	< 0.01	98 [64, 150]	96.20	< 0.01
Chung et al (2018) [50]	0.89 [0.87, 0.92]	96.32	< 0.01	0.92 [0.89, 0.93]	98.01	< 0.01	93 [62, 138]	96.10	< 0.01
Derkatch et al (2019) [51]	0.90 [0.87, 0.92]	96.56	< 0.01	0.92 [0.90, 0.94]	98.11	< 0.01	101 [67, 154]	96.20	< 0.01
Gan et al (2019) [52]	0.90 [0.87, 0.92]	96.49	< 0.01	0.92 [0.89, 0.94]	98.09	< 0.01	98 [65, 149]	96.20	< 0.01
Guy et al (2021) [53]	0.91 [0.88, 0.92]	93.83	< 0.01	0.92 [0.90, 0.94]	97.68	< 0.01	114 [78, 166]	91.80	< 0.01
Hendrix et al (2021) [54]	0.90 [0.88, 0.92]	96.56	< 0.01	0.92 [0.90, 0.94]	98.17	< 0.01	103 [68, 156]	96.20	< 0.01
Hu et al (2021) [78]	0.90 [0.87, 0.92]	96.57	< 0.01	0.92 [0.90, 0.94]	98.16	< 0.01	102 [67, 155]	96.20	< 0.01
Jiménez-Sánchez et al (2020) [55]	0.90 [0.87, 0.92]	96.48	< 0.01	0.92 [0.89, 0.94]	98.10	< 0.01	98 [65, 149]	96.20	< 0.01
Jones et al (2020) [56]	0.90 [0.87, 0.92]	94.97	< 0.01	0.92 [0.90, 0.94]	97.20	< 0.01	99 [65, 151]	96.00	< 0.01
Kim et al (2021) [57]	0.90 [0.87, 0.92]	96.52	< 0.01	0.92 [0.89, 0.94]	98.12	< 0.01	100 [66, 153]	96.20	< 0.01
Kitamura et al (2019) [58]	0.90 [0.88, 0.92]	96.62	< 0.01	0.92 [0.90, 0.94]	98.22	< 0.01	106 [70, 162]	96.20	< 0.01
Krogue et al (2020) [59]	0.90 [0.87, 0.92]	96.47	< 0.01	0.92 [0.89, 0.94]	98.09	< 0.01	98 [65, 149]	96.10	< 0.01
Langerhuizen et al (2020) [60]	0.90 [0.88, 0.92]	96.60	< 0.01	0.92 [0.90, 0.94]	98.18	< 0.01	104 [69, 157]	96.20	< 0.01
Liet al (2021) [61]	0.90 [0.87, 0.92]	96.47	< 0.01	0.92 [0.89, 0.94]	98.10	< 0.01	98 [64, 151]	96.20	< 0.01
Ma et al (2021) [62]	0.90 [0.87, 0.92]	96.55	< 0.01	0.92 [0.89, 0.94]	98.14	< 0.01	102 [66, 156]	96.20	< 0.01
MacKinnon et al (2018) [63]	0.90 [0.87, 0.92]	96.52	< 0.01	0.92 [0.89, 0.94]	98.14	< 0.01	100 [66, 153]	96.20	< 0.01
Mawatari et al (2020) [64]	0.90 [0.88, 0.92]	96.57	< 0.01	0.92 [0.90, 0.94]	98.17	< 0.01	102 [68, 155]	96.20	< 0.01
Mehta et al (2020) [65]	0.90 [0.88, 0.92]	96.46	< 0.01	0.91 [0.89, 0.93]	98.10	< 0.01	98 [64, 152]	96.40	< 0.01
Monchka et al (2021) [66]	0.90 [0.88, 0.92]	96.62	< 0.01	0.92 [0.89, 0.94]	97.94	< 0.01	102 [67, 157]	96.20	< 0.01
Mutasa et al (2020) [67]	0.90 [0.87, 0.92]	96.50	< 0.01	0.92 [0.89, 0.94]	98.12	< 0.01	99 [65, 150]	96.20	< 0.01
Ozkaya et al (2020) [68]	0.90 [0.88, 0.92]	96.54	< 0.01	0.92 [0.89, 0.94]	98.16	< 0.01	102 [67, 154]	96.20	< 0.01
Rayanet al (2019) [69]	0.90 [0.87, 0.92]	96.49	< 0.01	0.92 [0.90, 0.94]	98.15	< 0.01	101 [66, 154]	96.20	< 0.01
Reichert et al (2021) [70]	0.90 [0.87, 0.92]	96.52	< 0.01	0.92 [0.90, 0.94]	98.14	< 0.01	100 [66, 152]	96.20	< 0.01
Ren et al (2021) [71]	0.90 [0.87, 0.92]	96.47	< 0.01	0.92 [0.89, 0.94]	98.11	< 0.01	98 [65, 150]	96.30	< 0.01
Sato et al (2021) [72]	0.90 [0.87, 0.92]	96.40	< 0.01	0.92 [0.89, 0.93]	98.03	< 0.01	96 [63, 144]	96.00	< 0.01
Small et al (2021) [79]	0.90 [0.88, 0.92]	96.47	< 0.01	0.92 [0.89, 0.93]	98.06	< 0.01	99 [65, 151]	96.20	< 0.01
Urakawa et al (2019) [73]	0.90 [0.87, 0.92]	96.43	< 0.01	0.92 [0.89, 0.93]	98.06	< 0.01	96 [64, 145]	96.10	< 0.01
Voter et al (2021) [80]	0.90 [0.88, 0.92]	96.11	< 0.01	0.92 [0.89, 0.94]	98.03	< 0.01	102 [67, 154]	96.10	< 0.01
Weikert et al (2020) [81]	0.90 [0.88, 0.92]	96.52	< 0.01	0.92 [0.89, 0.94]	98.12	< 0.01	101 [66, 153]	96.20	< 0.01
Yoon et al (2021) [74]	0.90 [0.88, 0.92]	96.59	< 0.01	0.92 [0.89, 0.94]	98.07	< 0.01	100 [66, 153]	96.10	< 0.01
Yuet al (2020) [75]	0.90 [0.87, 0.92]	96.37	< 0.01	0.92 [0.89, 0.93]	98.04	< 0.01	95 [63, 143]	96.10	< 0.01

suspected fractures or more detailed fracture information. Meanwhile, DEXA played an important role in the included X-ray studies, all of which were used to identify patients at risk of vertebral fractures associated with low bone mineral density. Because plain X-rays were easier and widely used in daily clinical work, it might be suitable to use plain X-rays as input images when developing computer-assisted screening systems.

However, there is still no wide acceptance and implementation of such technology in clinical practice. One of the underlying reasons was the so-called inscrutable "black box" conundrum of deep learning [86] referring to the inability of the interpreters to clearly understand all the features displayed for making proper clinical decisions. Hence, the method for visual interpretation, such as gradient-weighted class activation mapping (Grad-CAM) [87], has been proposed. Grad-CAM generates a heatmap that visualizes the class-discriminative regions and helps the physician identify the pathologic region. Our results showed that AI with detecting and localizing fractures achieved promising results (sensitivity = 90%, specificity = 93%, AUC = 0.97). However, its actual value for localizing fracture lines could be reduced as it could show the fracture as a rough area, but cannot show the fracture line itself. Although the handcrafted features selected by experts in cML seem to be effective, such observations using small sample size data could limit reproducibility. Thus, more complex network architectures combined with larger training data may enable DL models to discover previously unknown cues.

Another roadblock is the coherence of the datasets used with real-world data in terms of the clinical aspect. We observed that nine studies excluded images that contained fractures in any other parts. Two included studies were considered to have non-consecutive patient selection owing to the risk of obscuring the disease spectrum in the dataset [86, 88] and up to 26 studies were single-center studies. Many reviewers of AI studies recommend consecutive and multicenter study design or external validation methods to enhance the clinical impact and generalizability of the obtained results [18–20, 72].

Our study had several limitations. First, studies with a high or unclear risk of bias in the domain of patient selection were observed in the majority of the included studies, representing a possibility of combined sensitivity and specificity overestimation related to patient selection bias. Second, high heterogeneity was observed in both sensitivity and specificity analysis. Therefore, subgroup analyses were performed, which showed that multicenter study design, application of transfer learning, or data augmentation were associated with the diagnostic performance of AI. Lastly, the majority of the included studies (38/39) built AI models without integrating important clinical information of orthopedic fractures, such as injury details, and symptoms, which conflicts with the considerations of clinical practice.

Conclusion

Our findings showed promising results for quantitative AI-based diagnosis of orthopedic fractures. Diagnosis using AI achieved comparable results to humans and was superior to non-expert human readers. Multicenter study design and application of transfer learning or data augmentation were associated with the improvement of AI performance. Further randomized, large-scale, prospective studies are required to validate our findings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-08956-4.

Acknowledgements The authors would like to acknowledge Professor Hao Liu, whose expertise was invaluable in formulating the research questions and methodology.

Funding This study was supported by a grant from the Popularization and Application Project of the Sichuan Provincial Health Commission (21PJ037)

Declarations

Guarantor The scientific guarantor of this publication is Professor Hao Liu (MD, PhD) of West China Hospital, China.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors (Yi Yang) has significant statistical expertise (6 years of experience in a systematic review and meta-analysis). Also, multiple authors have significant statistical expertise.

Informed consent No informed consent was needed for the conducting of this review.

Ethical approval Institutional Review Board approval was not required because of the nature of the study (meta-analysis), which did not include specimens or involve any treatments or interventions.

Study subjects or cohorts overlap All of the included studies have been previously reported, either as an original research paper.

Methodology
- Systematic review
- Meta-analysis
- Performed at one institution

References

1. Buhr AJ, Cooke AM (1959) Fracture patterns. Lancet 273:531–536
2. Court-Brown CM, Caesar B (2006) Epidemiology of adult fractures: a review. Injury 37:691–697
3. Sahlin Y (1990) Occurrence of fractures in a defined population: a 1-year study. Injury 21:158–160
4. Çolak I, Bekler HI, Bulut G, Eceviz E, Gülabi D, Çeçen GS (2018) Lack of experience is a significant factor in the missed diagnosis of perilunate fracture dislocation or isolated dislocation. Acta Orthop Traumatol Turc 52:32–36
5. Moonen PJ, Mercelina L, Boer W, T Fret (2017) Diagnostic error in the Emergency Department: follow up of patients with minor trauma in the outpatient clinic. Scand J Trauma Resusc Emerg Med 25:13

6. Wei CJ, Tsai WC, Tiu CM, Wu HT, Chiou HJ, Chang CY (2006) Systematic analysis of missed extremity fractures in emergency radiology. Acta Radiol 47:710–717
7. Bottle A, Aylin P (2006) Mortality associated with delay in operation after hip fracture: observational study. BMJ 332:947–951
8. Leer-Salvesen S, Engesæter LB, Dybvik E, Furnes O, Kristensen TB, Gjertsen JE (2019) Does time from fracture to surgery affect mortality and intraoperative medical complications for hip fracture patients? An observational study of 73 557 patients reported to the Norwegian Hip Fracture Register. Bone Joint J 101-b:1129-1137
9. McKinney SM, Sieniek M, Godbole V et al (2020) International evaluation of an AI system for breast cancer screening. Nature 577: 89–94
10. Rodríguez-Ruiz A, Krupinski E, Mordang JJ et al (2019) Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 290:305–314
11. Rodriguez-Ruiz A, Lång K, Gubern-Merida A et al (2019) Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 111: 916–922
12. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H (2018) Artificial intelligence in retina. Prog Retin Eye Res 67:1–29
13. Vujosevic S, Aldington SJ, Silva P et al (2020) Screening for diabetic retinopathy: new perspectives and challenges. Lancet Diabetes Endocrinol 8:337–347
14. Kikinis R, Wells WM 3rd (2020) Detection of brain metastases with deep learning single-shot detector algorithms. Radiology 295:416–417
15. Xue J, Wang B, Ming Y et al (2020) Deep learning-based detection and segmentation-assisted management of brain metastases. Neuro Oncol 22:505–514
16. Abbasi J (2020) Artificial intelligence-based skin cancer phone apps unreliable. JAMA 323:1336
17. Esteva A, Kuprel B, Novoa RA et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542: 115–118
18. Gregory J, Welliver S, Chong J (2020) Top 10 reviewer critiques of radiology artificial intelligence (AI) articles: qualitative thematic analysis of reviewer critiques of machine learning/ deep learning manuscripts submitted to JMRI. J Magn Reson Imaging 52:248–254
19. Park SH, Han K (2018) Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 286:800–809
20. Park SH, Kressel HY (2018) Connecting technological innovation in artificial intelligence to real-world medical practice through rigorous clinical validation: what peer-reviewed medical journals could do. J Korean Med Sci 33:e152
21. Duron L, Ducarouge A, Gillibert A et al (2021) Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study. Radiology 300:120–129
22. Kirienko M, Sollini M, Ninatti G et al (2021) Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI. Eur J Nucl Med Mol Imaging 48:3791–3804
23. Lee AY, Yanagihara RT, Lee CS et al (2021) Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 44:1168–1175
24. Novakovsky G, Saraswat M, Fornes O, Mostafavi S, Wasserman WW (2021) Biologically relevant transfer learning improves transcription factor binding prediction. Genome Biol 22:280
25. Shi H, Li J, Mao, Hwang KS (2021) Lateral transfer learning for multiagent reinforcement learning. IEEE Trans Cybern 1–13
26. Xiao Y, Liang F, Liu B (2022) A transfer learning-based multi-instance learning method with weak labels. IEEE Trans Cybern 52:287–300
27. Zhen L, Hu P, Peng X, Goh RSM, Zhou JT (2022) Deep multimodal transfer learning for cross-modal retrieval. IEEE Trans Neural Netw Learn Syst 33:798–810
28. Chaitanya K, Karani N, Baumgartner CF et al (2021) Semi-supervised task-driven data augmentation for medical image segmentation. Med Image Anal 68:101934
29. Gao J, Hua Y, Hu G, Wang C, Robertson NM (2021) Discrepancy-guided domain-adaptive data augmentation. IEEE Trans Neural Netw Learn Syst 1–12
30. Tran NT, Tran VH, Nguyen NB, Nguyen TK, Cheung NM (2021) On data augmentation for GAN training. IEEE Trans Image Process 30:1882–1897
31. Jonsdottir KY, Østergaard L, Mouridsen K (2009) Predicting tissue outcome from acute stroke magnetic resonance imaging: improving model performance by optimal sampling of training data. Stroke 40: 3006–3011
32. Rank N, Pfahringer B, Kempfert J et al (2020) Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. NPJ Digit Med 3:139
33. Sanders WS, Johnston CI, Bridges SM, Burgess SC, Willeford KO (2011) Prediction of cell penetrating peptides by support vector machines. PLoS Comput Biol 7:e1002101
34. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G (2020) Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology 132:379–394
35. Kumar A, Pirogova E, Mahmoud SS, Fang Q (2021) Classification of error-related potentials evoked during stroke rehabilitation training. J Neural Eng 18
36. Reichstein M, Camps-Valls G, Stevens B et al (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566:195–204
37. Schwendicke F, Golla T, Dreher M, Krois J (2019) Convolutional neural networks for dental image diagnostics: A scoping review. J Dent 91:103226
38. Jin C, Chen W, Cao Y et al (2020) Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 11:5088
39. Kalmet PHS, Sanduleanu S, Primakov S et al (2020) Deep learning in fracture detection: a narrative review. Acta Orthop 91:215–220
40. Liberati A, Altman DG, Tetzlaff J et al (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 6:e1000100
41. Whiting PF, Rutjes AW, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155:529–536
42. Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560
43. Bae J, Yu S, Oh J et al (2021) External validation of deep learning algorithm for detecting and visualizing femoral neck fracture including displaced and non-displaced fracture on plain X-ray. J Digit Imaging 34:1099–1109
44. Beyaz S, Açıcı K, Sümer E (2020) Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg 31:175–183
45. Blüthgen C, Becker AS, Vittoria DMI, Meier A, Martini K, Frauenfelder T (2020) Detection and localization of distal radius fractures: deep learning system versus radiologists. Eur J Radiol 126:108925
46. Cheng CT, Chen CC, Cheng FJ et al (2020) A human-algorithm integration system for hip fracture detection on plain radiography: system development and validation study. JMIR Med Inform 8: e19416

47. Cheng CT, Ho TY, Lee TY et al (2019) Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 29:5469–5477
48. Choi J, Hui JZ, Spain D, Su YS, Cheng CT, Liao CH(2021) Practical computer vision application to detect hip fractures on pelvic X-rays: a bi-institutional study. Trauma Surg Acute Care Open 6:e000705
49. Choi JW, Cho YJ, Lee S et al (2020) Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. Invest Radiol 55:101–110
50. Chung SW, Han SS, Lee JW et al (2018) Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 89:468–473
51. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD (2019) Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: a registry-based cohort study of dual X-ray absorptiometry. Radiology 293:405–411
52. Gan K, Xu D, Lin Y et al (2019) Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthop 90: 394–400
53. Guy S, Jacquet C, Tsenkoff D, Argenson JN, Ollivier M (2021) Deep learning for the radiographic diagnosis of proximal femur fractures: limitations and programming issues. Orthop Traumatol Surg Res 107:102837
54. Hendrix N, Scholten E, Vernhout B et al (2021) Development and validation of a convolutional neural network for automated detection of scaphoid fractures on conventional radiographs. Radiol Artif Intell 3:e200260
55. Jiménez-Sánchez A, Kazi A, Albarqouni S et al (2020) Precise proximal femur fracture classification for interactive training and surgical planning. Int J Comput Assist Radiol Surg 15:847–857
56. Jones RM, Sharma A, Hotchkiss R et al (2020) Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digit Med 3:144
57. Kim MW, Jung J, Park SJ et al (2021) Application of convolutional neural networks for distal radio-ulnar fracture detection on plain radiographs in the emergency room. Clin Exp Emerg Med 8:120–127
58. Kitamura G, Chung CY, Moore BEN (2019) Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. J Digit Imaging 32:672–677
59. Krogue JD, Cheng KV, Hwang KM et al (2020) Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell 2:e190023
60. Langerhuizen DWG, Bulstra AEJ, Janssen SJ et al (2020) Is deep learning on par with human observers for detection of radiographically visible and occult fractures of the scaphoid? Clin Orthop Relat Res 478:2653–2659
61. Li YC, Chen HH, Horng-Shing LH, Wu HTH, Chang MC, Chou PH (2021) Can a deep-learning model for the automated detection of vertebral fractures approach the performance level of human subspecialists? Clin Orthop Relat Res 479:1598–1612
62. Ma Y, Luo Y (2021) Bone fracture detection through the two-stage system of Crack-Sensitive Convolutional Neural Network. Inform Med Unlocked 22
63. MacKinnon T (2018) Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 73:439–445
64. Mawatari T, Hayashida Y, Katsuragawa S et al (2020) The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs. Eur J Radiol 130:109188
65. Mehta SD, Sebro R (2020) Computer-aided detection of incidental lumbar spine fractures from routine dual-energy X-ray absorptiometry (DEXA) studies using a support vector machine (SVM) classifier. J Digit Imaging 33:204–210
66. Monchka BA, Kimelman D, Lix LM, Leslie WD (2021) Feasibility of a generalized convolutional neural network for automated identification of vertebral compression fractures: the Manitoba Bone Mineral Density Registry. Bone 150:116017
67. Mutasa S, Varada S, Goel A, Wong TT, Rasiej MJ (2020) Advanced deep learning techniques applied to automated femoral neck fracture detection and classification. J Digit Imaging 33: 1209–1217
68. Ozkaya E, Topal FE, Bulut T, Gursoy M, Ozuysal M, Karakaya Z (2022) Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. Eur J Trauma Emerg Surg 48:585–592
69. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A (2019) Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiol Artif Intell 1:e180015
70. Reichert G, Bellamine A, Fontaine M et al (2021) How can a deep learning algorithm improve fracture detection on X-rays in the emergency room? J Imaging 7
71. Ren M, Yi PH (2022) Deep learning detection of subtle fractures using staged algorithms to mimic radiologist search pattern. Skeletal Radiol 51:345–353
72. Sato Y, Takegami Y, Asamoto T et al (2021) Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: a multicenter study. BMC Musculoskelet Disord 22:407
73. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N (2019) Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol 48:239–244
74. Yoon AP, Lee YL, Kane RL, Kuo CF, Lin C, Chung KC (2021) Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. JAMA Netw Open 4:e216096
75. Yu JS, Yu SM, Erdal BS et al (2020) Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. Clin Radiol 75:237.e231-237.e239
76. Al-Helo S, Alomari RS, Ghosh S et al (2013) Compression fracture diagnosis in lumbar: a clinical CAD system. Int J Comput Assist Radiol Surg 8:461–469
77. Burns JE, Yao J, Summers RM (2017) Vertebral body compression fractures and bone density: automated detection and classification on CT images. Radiology 284:788–797
78. Hu Y, He X, Zhang R, Guo L, Gao L, Wang J (2021) Slice grouping and aggregation network for auxiliary diagnosis of rib fractures. Biomed Signal Process Control 67
79. Small JE, Osler P, Paul AB, Kunst M (2021) CT cervical spine fracture detection using a convolutional neural network. AJNR Am J Neuroradiol 42:1341–1347
80. Voter AF, Larson ME, Garrett JW, Yu JPJ (2021) Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of cervical spine fractures. AJNR Am J Neuroradiol 42:1550–1556
81. Weikert T, Noordtzij LA, Bremerich J et al (2020) Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography. Korean J Radiol 21:891–899
82. Caravagna G, Giarratano Y, Ramazzotti D et al (2018) Detecting repeated cancer evolution from multi-region tumor sequencing data. Nat Methods 15:707–714
83. Schwessinger R, Gosden M, Downes D et al (2020) DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nat Methods 17:1118–1124

84. Wang J, Agarwal D, Huang M et al (2019) Data denoising with transfer learning in single-cell transcriptomics. Nat Methods 16: 875–878
85. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359
86. Thrall JH, Li X, Li Q et al (2018) Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol 15:504–508
87. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization 2017 IEEE International Conference on Computer Vision (ICCV), pp 618-626
88. Sica GT (2006) Bias in research studies. Radiology 238:780–789
89. Kuo RYL, Harrison C, Curran TA et al (2022) Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology. https://doi.org/10.1148/radiol.211785:211785