ANNALS OF MEDICINE
2026, VOL. 58, NO. 1, 2610079
https://doi.org/10.1080/07853890.2025.2610079

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Enhanced fracture detection on radiographs with AI assistance for clinicians: a systematic review and meta-analysis

Han Qin, Yunxia Ding, Jiangyi Ju, Zhen Qu and Lihua Peng

Bishan Hospital of Chongqing Medical University (Bishan Hospital of Chongqing), Chongqing, China

ABSTRACT

Background: Emergency radiographic interpretation for fractures is prone to missed or misdiagnoses. Artificial intelligence (AI) is expected to become a powerful tool to assist clinicians in fracture detection.

Purpose: A systematic review and meta-analysis was performed to assess whether AI improves clinicians' ability to detect fractures on radiographs.

Materials and Methods: A literature search was conducted in PubMed, Web of Science, and Cochrane Library for studies published between January 1, 2010, and October 10, 2025. A meta-analysis of diagnostic accuracy studies was performed using a Summary Receiver Operating Characteristic (SROC) curve. The quality of included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool. Subgroup analysis and meta-regression were conducted to explore potential sources of heterogeneity.

Results: A total of 26 studies were included . The pooled sensitivity of clinicians increased from 77% (95% CI: 72–81) to 87% (95% CI: 83–90) with AI assistance, while the pooled specificity improved from 88% (95% CI: 85–90) to 92% (95% CI: 89–94). The corresponding AUC values were 0.90 (95% CI: 0.87–0.92) before and 0.95 (95% CI: 0.93–0.97) after AI assistance. Eight studies were rated as high risk of bias. Subgroup analysis and meta-regression identified potential sources of heterogeneity, including fracture location, AI model type, high risk of bias, and reference standards.

Conclusion: AI assistance significantly improves clinicians' diagnostic performance in detecting fractures on radiographs for extremity and trunk fractures.

ARTICLE HISTORY
Received 20 July 2025
Revised 16 December 2025
Accepted 22 December 2025

KEYWORDS
Radiographic interpretation; medical image analysis; deep learning; clinical decision support; assisted image interpretation; orthopedic imaging; machine learning models

Common diagnostic errors can be categorized into missed diagnoses, misdiagnoses, and delayed diagnoses, all of which may prevent patients from receiving timely and appropriate treatment [1]. Fractures are among the most frequently missed conditions in emergency departments, with Fernholm R et al. reporting a misdiagnosis rate of up to 24% for fractures [2]. Fracture detection primarily relies on the interpretation of radiographs, and missed fracture diagnosis is also one of the most common radiological errors [3]. Such diagnostic oversights most often occur between 5 pm and 3 am, possibly due to physicians' insufficient experience or fatigue [4]. Guly's study further revealed that in emergency settings, the majority (85.3%) of missed fracture diagnoses were made by junior doctors [5]. Artificial intelligence, however, can effectively reduce errors caused by fatigue. Moreover, AI's performance in interpreting fracture radiographs has been shown to approach that of human experts [6].

In recent years, artificial intelligence (AI) has been increasingly applied in the medical field, particularly in radiology [7,8]. With the advancement of AI, especially the emergence of deep learning, it has demonstrated tremendous potential in various medical applications [9]—for instance, predicting clinical disability in systemic sclerosis [10], assessing the severity of diabetic retinopathy [11], and identifying histopathological features [12]. The use of AI to interpret radiographs for fracture detection has also proven feasible. Alfred P. Yoon's study reported that AI achieved high sensitivity and specificity in

CONTACT Lihua Peng 140733@hospital.cqmu.edu.cn Bishan Hospital of Chongqing Medical University (Bishan Hospital of Chongqing), Chongqing, China.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07853890.2025.2610079.

detecting scaphoid fractures on wrist radiographs [13]. Similarly, Jeff Choi developed a hip fracture detection algorithm with an area under the receiver operating characteristic curve (AUC) of 0.98 [14]. Furthermore, AI can interpret images at a speed far beyond human capability. However, due to ethical constraints, AI cannot directly replace clinicians. Nevertheless, integrating AI as an assistive tool for radiograph interpretation represents a highly promising direction for future development [15].

To date, numerous studies have investigated the performance of clinicians assisted by artificial intelligence (AI) in fracture detection [16–18]. Several reviews and meta-analyses have also explored the relationship between AI and fracture detection. For instance, Kuo RYL et al. [15] conducted a meta-analysis of 42 studies and concluded that 'Artificial intelligence (AI) and clinicians had comparable reported diagnostic performance in fracture detection,' and further found through a meta-analysis of four studies that 'The addition of AI assistance improved clinician performance further.' Similarly, Julius Husarek et al. [19] performed a meta-analysis of 17 studies involving commercially available AI systems and reported that AI can provide valuable second opinions, thereby enhancing diagnostic confidence and accuracy. In addition, Mohammed Kutbi [20] highlighted in his review that AI has broad applications and significant benefits in fracture detection.

In our study, we conducted a meta-analysis of 26 studies, including both commercially available AI systems and researcher-developed models, to provide a more comprehensive evaluation of the feasibility of AI-assisted fracture detection based on clinician interpretation of radiographs. We also assessed potential directions for future development in this field.

Materials and methods

Registration

This systematic review was prospectively registered in PROSPERO (CRD 42025640034). Our study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and the PRISMA-Diagnostic Test Accuracy (PRISMA-DTA) guidelines [21,22].

Study selection

We included all studies that met the following criteria: (1) published in English; (2) compared clinicians' performance in interpreting fracture radiographs with and without AI assistance; and (3) used radiographs as the imaging modality. The exclusion criteria were as follows: reviews or meta-analyses, comments, studies involving animal fractures, and studies in which the imaging modality was not radiographs.

The search strategy used the terms ((ai) OR (artificial intelligence) OR (machine learning) OR (deep learning)) AND ((fracture) OR (fracture detection)) AND ((radiologist) OR (X-ray) OR (radiograph)) to retrieve relevant studies from PubMed, Web of Science, and the Cochrane Library between January 1, 2010, and October 10, 2025. EndNote X9 was used to remove duplicate records. Literature screening was conducted in two stages: first by reviewing titles and abstracts, followed by full-text assessment.

Data extraction

During data extraction, we primarily collected the numbers of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) from each study. Each individual radiograph, rather than each fracture site, was treated as a single sample. This approach was chosen because if a patient has multiple fractures and even one is missed, the patient would still require secondary treatment, potentially resulting in significant harm. Therefore, a radiograph was considered a true positive only when all fractures on the image were correctly identified. This definition also allowed for over-identification, as labeling non-fractured areas as fractured would have minimal impact on patient management. A case was defined as a false negative if any actual fracture was missed. A true negative was defined as a radiograph that was negative according to the reference standard and was also interpreted as negative by the reader. Conversely, a false positive was defined as a radiograph judged by the reference standard

to be negative but interpreted by the reader as positive for fracture. The reference standards used were those defined within each included study.

This process was independently performed by two reviewers, and the results were consolidated by a third reviewer. Any disagreements were resolved through discussion among all three reviewers.

Quality assessment

We used the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool to evaluate the risk of bias and applicability of the included studies [23]. In the 'Risk of Bias' domain, assessment was conducted across four key areas: Patient Selection, Index Test, Reference Standard, and Flow and Timing. In the 'Applicability Concerns' domain, evaluation focused on Patient Selection, Index Test, and Reference Standard.

For the second question under Patient Selection—'Was a case-control design avoided?'—we rated all studies as 'Yes.' This is because, in all included studies, patients were confirmed to have or not have fractures based on the reference standard, and none of the studies aimed to retrospectively compare fracture and non-fracture groups for risk factor exposure to determine causal relationships.

Regarding study design, the most appropriate experimental approach was the one used by Rikke Bachmann et al. [24]. In their study, patients were divided into two groups with equal numbers of fracture and non-fracture cases (Group 1 and Group 2). During the first reading session, Group 1 was interpreted without AI assistance, while Group 2 was interpreted with AI assistance. After a washout period, the same readers reinterpreted both groups, but the conditions were reversed—Group 1 with AI assistance and Group 2 without AI assistance. This design avoided bias due to differences in reader experience between two independent groups. If patients were not divided into two groups and the same readers simply reinterpreted the same radiographs after a washout period, the results could be affected by memory retention or improvement in reading ability over time. By dividing patients into two groups and alternating AI assistance before and after the washout period, this design ensured that both 'with AI' and 'without AI' conditions existed in each phase, thereby maximizing experimental validity. Therefore, in the Index Test domain under 'Concerns regarding applicability,' only studies employing this design were rated as Low concern, while others were rated as High concern or Unclear concern.

Regarding the Reference Standard, if CT imaging data were available, fracture determination was much more reliable. Thus, studies using CT as part of the reference standard were rated as Low concern in applicability, while those without CT were rated as High concern or Unclear concern.

Finally, for the Flow and Timing item 'Was there an appropriate interval between index test and reference standard?', since all studies involved interpretation of the same radiographs and the images did not change during the study period, all studies were rated as 'Yes.'

This process was independently performed by two reviewers, and the results were consolidated by a third reviewer. Any disagreements were resolved through discussion among all three reviewers.

Statistical analysis

We performed a meta-analysis of diagnostic accuracy studies, generated forest plots, and used the Summary Receiver Operating Characteristic (SROC) curve [25] to evaluate the overall diagnostic performance. Regarding the area under the curve (AUC), values between 0.9 and 1 were considered excellent, those between 0.8 and 0.9 good, between 0.7 and 0.8 moderate, and below 0.7 poor.

Paired t-tests or Wilcoxon signed-rank tests were used to assess whether the differences between the two groups were statistically significant. The data were paired, representing clinicians' performance in interpreting radiographs independently and with AI assistance. Sensitivity and specificity data met or approximately met the assumption of normal distribution; therefore, paired t-tests were used. However, for accuracy, the histogram indicated that the normality assumption was not satisfied, so the Wilcoxon signed-rank test was applied instead to ensure the robustness of the results. A P-value of less than 0.05 was considered statistically significant.

Additionally, Wilcoxon rank-sum tests were used to compare the performance of standalone AI with that of clinicians without AI assistance, as well as to evaluate differences in reading time before and after AI assistance.

Heterogeneity analysis

We assessed inter-study heterogeneity using the I^2 statistic. Subgroup analyses were conducted based on algorithm type, patient age, sample size, fracture site, reference standard, number of clinician readers, and risk of bias. Meta-regression was performed to further explore potential sources of heterogeneity.

Publication bias

We assessed publication bias among the included studies by constructing a funnel plot.

Results

Study selection

A total of 3,323 studies were initially identified, including 1,720 from PubMed, 1,482 from Web of Science, and 121 from the Cochrane Library. After removing 688 duplicates using Endnote X9, the detailed screening process is illustrated in Figure 1. A total of 49 relevant studies were finally identified, of which 23 were excluded due to the inability to obtain detailed sample data or study results. Among these, two studies were in the preparatory phase and had not yet been implemented [26,27]; one study used median values for statistical analysis, making it impossible to extract detailed data for all samples [28]; and five studies did not define true positives (TP) based on identifying all fractures on a single radiograph. The remaining 15 studies were excluded for other reasons, mainly because the proportion of fracture and non-fracture cases in the radiographs was unclear. Additional reasons included results reported only as accuracy, lack of separate data for fracture detection, or insufficiently interpretable data. Ultimately, 26 studies were included in the meta-analysis. The characteristics of these included studies are summarized in Table 1. No restrictions were placed on patient age, sex, sample size, or study design during the selection process.

Characteristics

Among the included studies, ten used Gleamer's BoneView model [29,31,35,36,38–41,43,52], three used Radiobotics RBFracture [24,45,50], two used the Faster R-CNN algorithm [30,33], two used DenseNet [6,44], and two used the YOLO algorithm [32,51]. Eight studies involved patients aged 18 years or older [6,31,39,40,46,49,51,52], while another eight involved patients aged 22 years or younger [30,32,33,38,45,47,48,50]. Twenty studies focused on extremity fractures, five included multiple fracture sites [29,35,36,43,50], and one focused specifically on skull fractures [32].

The total sample size across studies was 9,218 radiographs. Four studies had sample sizes ≤100 [6,32,34,42], while the study by Andrea Dell'Aria included 101 radiographs [40]. Six studies had ≥500 samples [31,36,39,43,45,50]. Eleven studies maintained a 1:1 ratio between fracture and non-fracture cases, while seven studies included ≥25% of subtle (non-obvious) fractures [29,30,34,40,47,49,52]. In Loïc Duron's study, 'cases showing only obvious fractures (displaced, dislocated, or comminuted) according to the reference radiologist' were excluded [39].

All included studies established a reference standard. Four studies used consensus between two expert radiologists [29,36,37,49]; four used independent reviews by two physicians with disagreements resolved by a third [24,33,38,39]; ten studies determined the reference standard based on clinical information (e.g. CT, MRI, or follow-up data); and one study (John R. Zech) used the 'original radiology report' as the reference standard [30].

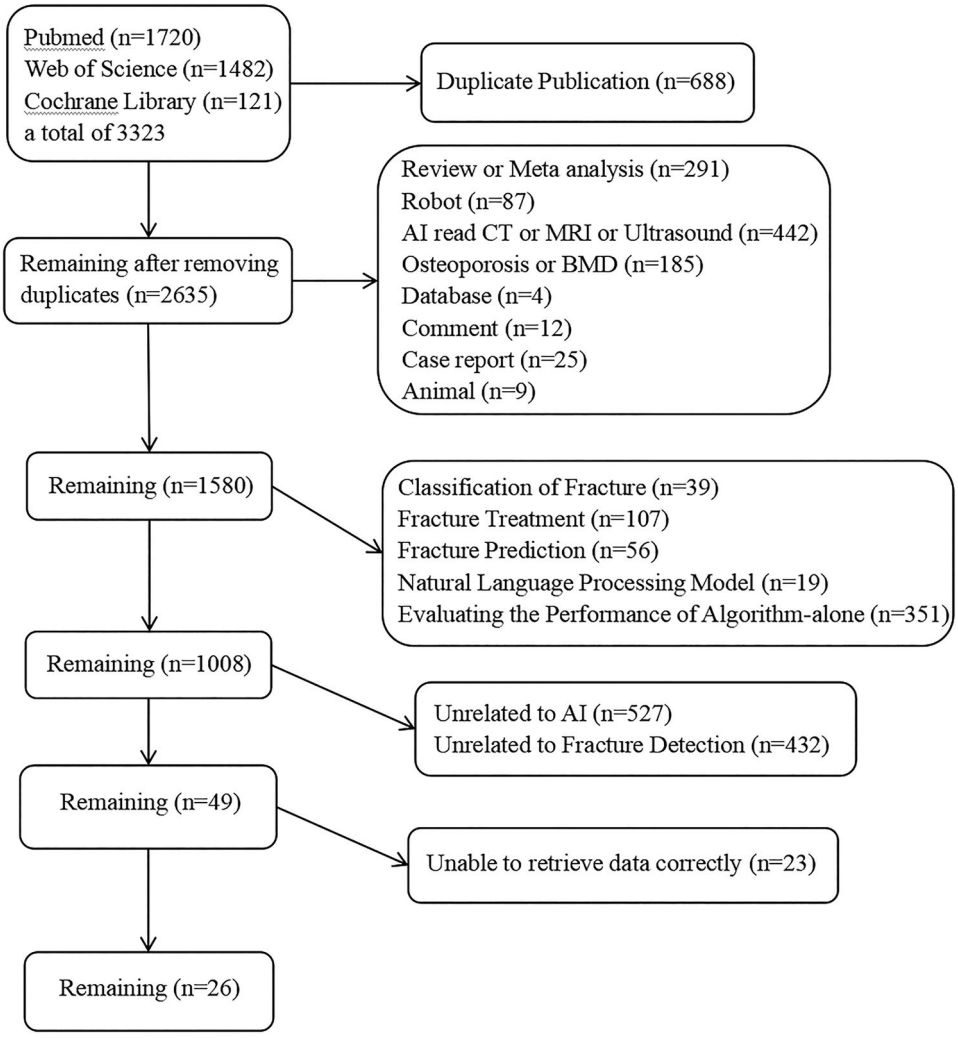

Figure 1. The literature screening process for inclusion.

The number of clinician readers interpreting the radiographs ranged from 2 to 120. Eleven studies involved ≥10 readers, while nine studies had ≤5 readers. In Mathieu Cohen's study, the initial radiology report was directly used as the clinician performance metric [31].

A total of 19 studies used multiple radiographic views, 4 studies used single radiographic view [6,30,44,49], 1 study included both single radiographic view and multiple radiographic views [29], and 2 studies did not specify the number of views [33,35].

Data extraction

The results of data extraction from the 26 studies included in the final analysis are presented in Table 2. The total sample size comprised 9,218 radiographs, with a cumulative total of 124,900 readings. Based on the data extracted from Table 2, we calculated sensitivity, specificity, and accuracy, and the final results are summarized in Table 3.

Quality assessment

The final results of the risk of bias and applicability assessment are shown in Figures 2 and 3, and this process was conducted using Review Manager 5.3.

Table 1. Characteristics of the included studies.

First Author	Year	Algorithm or Model	Target Condition	Sample size	Number of readers
Ali Guermazi [29]	2022	Boneview	Multi - center fractures	480	24
John R.Zech [30]	2023	Faster R-CNN	Wrist fracture	125	4
Mathieu Cohen [31]	2022	Boneview	Wrist fracture	647	41
Jae Won Choi [32]	2022	YOLOv 3	Skull fracture	95	5
John R.Zech [33]	2024	Faster R-CNN	Extremity fracture	240	7
Alfred P. Yoon [34]	2023	Unknown	Scaphoid fracture	15	120
Mathias Meetschen [35]	2024	Boneview	Multi - center fractures	200	4
Lisa Canoni-Meynet [36]	2022	Boneview	Multi - center fractures	500	3
Nils Hendrix [37]	2022	Unknown	Scaphoid fracture	219	5
Toan Nguyen [38]	2022	Boneview	Extremity fracture	300	8
Loïc Duron [39]	2021	Boneview	Extremity fracture	600	12
ANDREA DELL'ARIA [40]	2024	Boneview	Extremity fracture	101	2
Thibaut Jacques [41]	2024	Boneview	Extremity fracture	296	23
Marco Keller [42]	2024	Unknown	Distal radius fracture	20	22
Rikke Bachmann [24]	2024	RBFracture	Extremity fracture	334	15
Jonas Oppenheimer [43]	2023	Boneview	Multi - center fractures	1163	2
Justin D. Krogue [6]	2020	DenseNet169	Hip fracture	100	8
Chi-Tung Cheng [44]	2020	DenseNet-121	Hip fracture	200	30
Maria Ziegner [45]	2025	RBFracture	Extremity fracture	1672	3
R. Breu [46]	2024	U-Net	Distal Radius Fracture	200	11
Sean O'Rourke [47]	2025	Unknown	Extremity fracture	240	8
Cato Pauling [48]	2025	Milvue Suite-SmartUrgences	Extremity fracture	107	7
Pengyi Xing [49]	2024	Multi-module Collaborative Deep Learning Model	Femoral Neck Fracture	200	20
Praveen M Yogendra [50]	2024	RBFracture	Multi - center fractures	500	6
Yi Xie [51]	2025	YOLOv8n	Tibial Plateau Fracture	400	10
Ana Isabel Hernáiz Ferrer [52]	2025	BoneView, RBFracture	Scaphoid Fracture	264	2

Table 2. The results of data extraction from the 26 studies included in the research.

First Author	TP(Without AI)	TP(With AI)	FP (Without AI)	FP(With AI)	FN(Without AI)	FN(With AI)	TN(Without AI)	TN(With AI)
Ali Guermazi [29]	3732	4331	543	256	2028	1429	5217	5504
John R.Zech [30]	249	291	27	8	71	29	153	172
Mathieu Cohen [31]	188	217	16	32	59	30	384	368
Jae Won Choi [32]	65	72	60	35	30	23	320	345
John R.Zech [33]	644	721	175	98	196	119	665	742
Alfred P. Yoon [34]	864	1044	132	144	336	156	468	456
Mathias Meetschen [35]	116	154	46	42	84	46	154	158
Lisa Canoni-Meynet [36]	372	486	39	33	192	78	897	903
Nils Hendrix [37]	258	238	96	43	67	87	674	727
Toan Nguyen [38]	880	994	125	116	320	206	1075	1084
Loïc Duron [39]	1274	1429	189	115	526	371	1611	1685
ANDREA DELL'ARIA [40]	58	82	12	12	50	26	82	82
Thibaut Jacques [41]	2383	2600	312	296	1711	1494	2402	2418
Marco Keller [42]	96	120	7	1	4	0	93	119
Rikke Bachmann [24]	1771	1968	484	382	689	492	2066	2168
Jonas Oppenheimer [43]	622	670	46	42	112	64	1546	1550
Justin D. Krogue [6]	376	387	60	41	24	13	340	359
Chi-Tung Cheng [44]	1124	712	350	129	376	38	1150	621
Maria Ziegner [45]	1926	2009	252	206	375	292	2463	2509
R. Breu [46]	880	957	99	55	220	143	1001	1045
Sean O'Rourke [47]	280	315	148	148	200	165	332	332
Cato Pauling [48]	400	418	69	33	55	37	225	261
Pengyi Xing [49]	2750	2937	97	55	450	263	703	745
Praveen M Yogendra [50]	450	273	182	43	130	17	1238	667
Yi Xie [51]	1800	1908	104	30	200	92	1896	1970
Ana Isabel Hernáiz Ferrer [52]	190	432	46	92	72	92	220	440

TP: True Positive; FP: False Positive; FN: False Negative; TN: True Negative; AI: Artificial Intelligence.

In the 'Patient Selection' domain, 11 studies were rated as 'Low Risk of Bias', 6 studies as 'High Risk of Bias', and 2 studies as 'High Applicability Concerns'. In the 'Index Test' domain, all studies demonstrated a 'Low Risk of Bias', as readers were blinded to the reference standard when interpreting radiographs, and the thresholds were pre-specified. Only Rikke Bachmann's study [24] was rated as 'Low Applicability Concerns', while two studies were rated as 'High Applicability Concerns'. In the 'Reference Standard' domain, no study showed 'High Risk of Bias' or 'High Applicability Concerns.' In the 'Flow and Timing' domain, two studies were rated as 'High Risk of Bias.'

Table 3. The sensitivity, specificity, and accuracy in the included studies with and without AI assistance.

First Author	Se(Without AI)	Se(With AI)	Sp(Without AI)	Sp(With AI)	Ac(Without AI)	Ac(With AI)
Ali Guermazi [29]	0.648	0.752	0.906	0.956	0.777	0.854
John R.Zech [30]	0.778	0.909	0.850	0.956	0.804	0.926
Mathieu Cohen [31]	0.761	0.879	0.960	0.920	0.884	0.904
Jae Won Choi [32]	0.684	0.758	0.842	0.908	0.811	0.878
John R.Zech [33]	0.767	0.858	0.792	0.883	0.779	0.871
Alfred P. Yoon [34]	0.720	0.870	0.780	0.760	0.740	0.833
Mathias Meetschen [35]	0.580	0.770	0.770	0.790	0.675	0.780
Lisa Canoni-Meynet [36]	0.660	0.862	0.958	0.965	0.846	0.926
Nils Hendrix [37]	0.794	0.732	0.875	0.944	0.851	0.881
Toan Nguyen [38]	0.733	0.828	0.896	0.903	0.815	0.866
Loïc Duron [39]	0.708	0.794	0.895	0.936	0.801	0.865
ANDREA DELL'ARIA [40]	0.537	0.759	0.872	0.872	0.693	0.812
Thibaut Jacques [41]	0.582	0.635	0.885	0.891	0.703	0.737
Marco Keller [42]	0.960	1.000	0.930	0.992	0.945	0.996
Rikke Bachmann [24]	0.720	0.800	0.810	0.850	0.766	0.826
Jonas Oppenheimer [43]	0.847	0.913	0.971	0.974	0.932	0.954
Justin D. Krogue [6]	0.940	0.968	0.850	0.898	0.895	0.933
Chi-Tung Cheng [44]	0.749	0.949	0.767	0.828	0.758	0.889
Maria Ziegner [45]	0.837	0.873	0.907	0.924	0.875	0.901
R. Breu [46]	0.800	0.870	0.910	0.950	0.855	0.910
Sean O'Rourke [47]	0.583	0.656	0.692	0.692	0.638	0.674
Cato Pauling [48]	0.879	0.919	0.765	0.888	0.834	0.907
Pengyi Xing [49]	0.859	0.918	0.879	0.931	0.863	0.921
Praveen M Yogendra [50]	0.776	0.941	0.872	0.939	0.844	0.940
Yi Xie [51]	0.900	0.954	0.948	0.985	0.924	0.970
Ana Isabel Hernáiz Ferrer [52]	0.725	0.824	0.827	0.827	0.777	0.826

Se: Sensitivity; Sp: Specificity; Ac: Accuracy; AI: Artificial Intelligence.

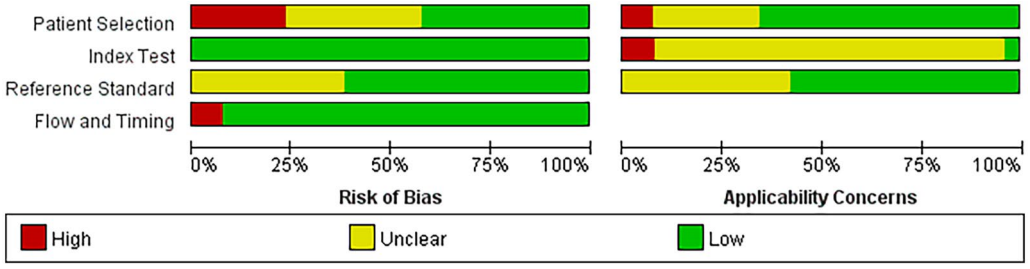

Figure 2. The results of risk of bias and applicability assessment for the 26 included studies.

Statistical analysis

The forest plots before and after AI assistance are shown in Figures 4 and 5, and the SROC curves evaluating clinicians' diagnostic performance before and after AI assistance are presented in Figures 6 and 7. The pooled sensitivity of clinicians increased from 77%(95%CI:72–81) to 87%(95%CI:83–90) with AI assistance, while pooled specificity improved from 88%(95%CI:85–90) to 92%(95%CI:89–94). The corresponding AUC values were 0.90(95%CI:0.87–0.92) before and 0.95(95%CI:0.93–0.97) after AI assistance. The analysis was performed using Stata 12.0.

Results of the paired t-test and Wilcoxon signed-rank test are summarized in Table 4. Sensitivity improved by an average of 9.5% (95% CI: 6.8–12.1) with AI assistance ($p < 0.001$), and specificity improved by an average of 3.7% (95% CI: 2.1–5.2) ($p < 0.001$), both showing statistically significant differences. For accuracy, the Z-value was −3.02 with $p < 0.05$, also indicating a significant difference. These findings demonstrate that clinicians with AI assistance achieved higher sensitivity, specificity, and accuracy than those without. This analysis was conducted using SPSS 27.0.1.

Among the included studies, three did not report the diagnostic performance (accuracy) of AI alone on the test dataset [39,42,49]. Comparing the independent AI performance from the remaining studies with that of clinicians without AI assistance using the Wilcoxon rank-sum test, we obtained a Z-value of −3.065 and a P-value of 0.002 (<0.05), indicating that the median accuracy of AI alone was significantly higher than that of unaided clinicians. This result is consistent with the findings reported by Pengran Liu et al. [53].

Figure 3. The results of risk of bias and applicability assessment for the 26 included studies.

In nine studies that reported reading times for clinicians with and without AI assistance, the mean reading time before AI assistance was 41.91 s, and after AI assistance was 34.73 s. The Wilcoxon signed-rank test yielded a Z-value of −0.839 with a P-value of 0.402 (>0.05), suggesting that there was no significant difference in reading time between clinicians with and without AI assistance when interpreting radiographs.

Heterogeneity analysis

The I^2 values of the forest plots before and after AI assistance were both greater than 50%, indicating substantial heterogeneity among the included studies. Subgroup analysis results showed the following:

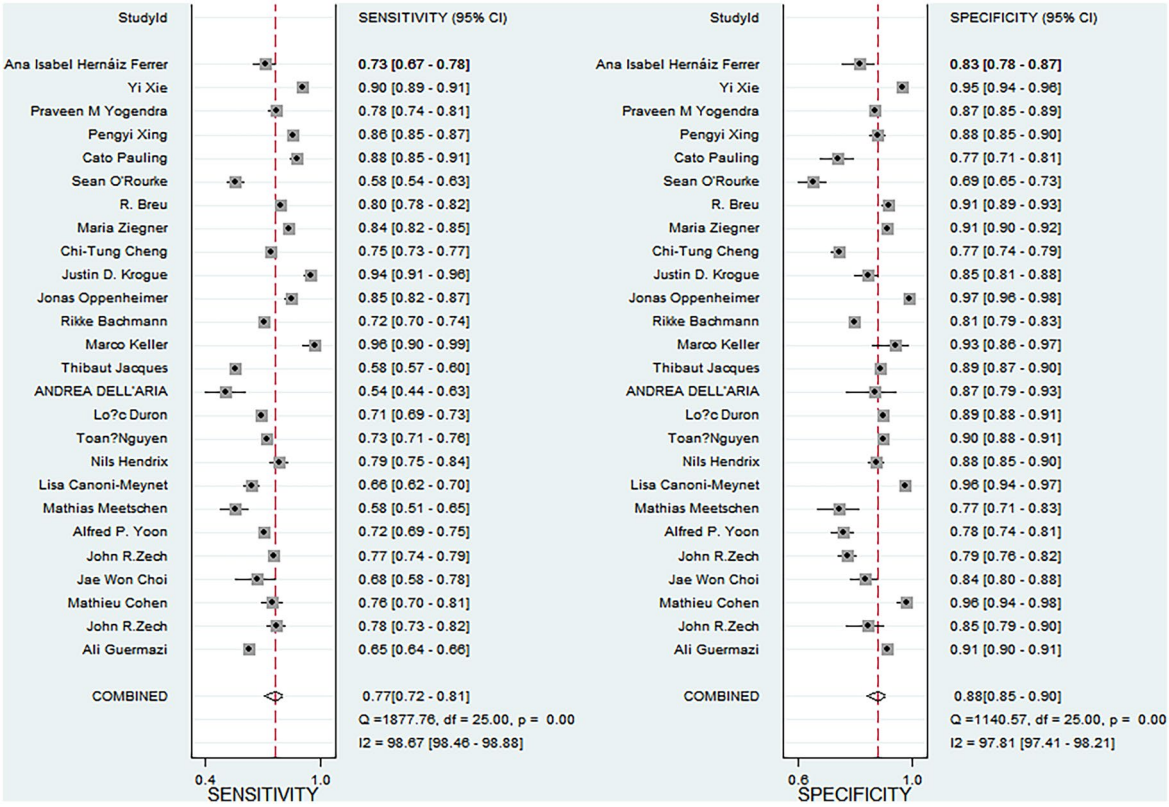

Figure 4. Results of the forest plot before AI assistance in included studies.

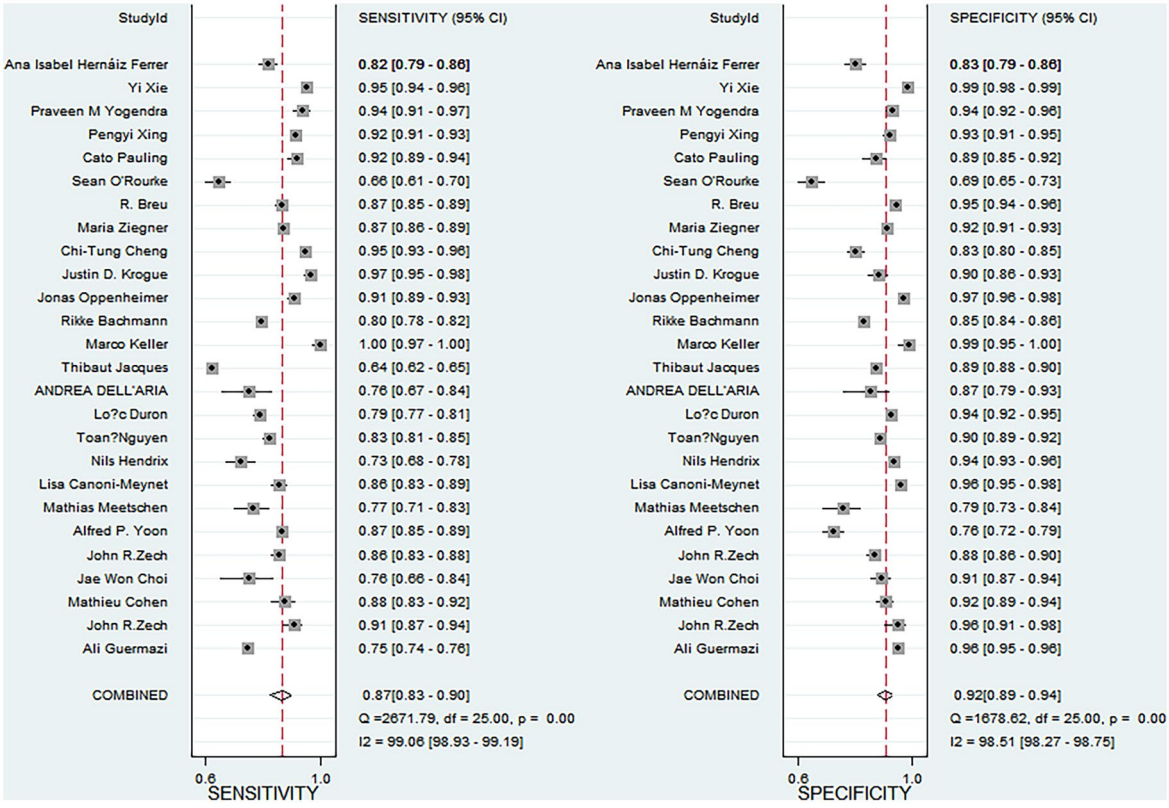

Figure 5. Results of the forest plot after AI assistance in included studies.

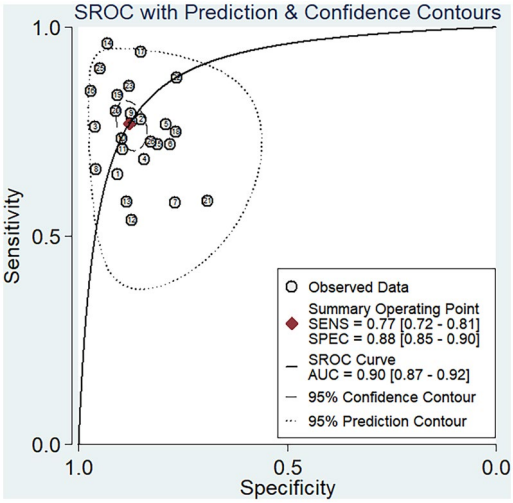

Figure 6. The performance of clinicians before AI assistance in the included studies.

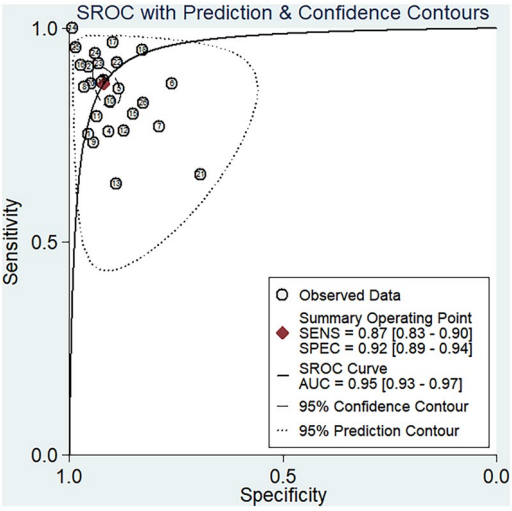

Figure 7. The performance of clinicians after AI assistance in the included studies.

Table 4. The results of paired t-test or wilcoxon signed-rank test for the data included in the study.

	Average difference	95%CI	t	P	d
Sensitivity	0.095	0.068;0.121	7.507	<0.001	1.472
Specificity	0.037	0.021;0.052	4.827	<0.001	0.947
	Z	p			
Accuracy	−3.020	0.003			
AI vs Clinicians	−3.065	0.002			
Reading time	−0.839	0.402			

95%CI: 95% confidence interval; AI: Artificial Intelligence.

For sensitivity, the I^2 values before and after AI assistance were: BoneView subgroup: 97.49 and 98.68; Patients aged ≥18 years: 98.24 and 97.90; Patients aged ≤22 years: 96.54 and 96.83; Extremity fracture subgroup: 98.79 and 99.23; Sample size between 100–500: 98.90 and 99.18; Reference standard involving clinical information (CT, MRI, etc.): 98.84 and 99.30; Reader number ≤5: 96.79 and 92.44; Reader number ≥10: 99.41 and 99.65; High risk of bias subgroup: 98.62 and 98.86; For specificity, the I^2 values before and after AI assistance were: BoneView subgroup: 95.92 and 97.77; Patients aged ≥18 years: 92.99 and 96.66; Patients aged ≤22 years: 96.94 and 97.51; Extremity fracture subgroup: 97.49 and 98.34; Sample size between 100–500: 97.80 and 98.62; Reference standard involving clinical information (CT, MRI, etc.):

96.86 and 96.48; Reader number ≤5: 96.39 and 96.18; Reader number ≥10: 98.82 and 99.33; High risk of bias subgroup: 98.07 and 98.76;

The meta-regression analysis indicated that, both before and after AI assistance, lower sensitivity might be associated with the use of the BoneView model (pre: 0.69 [0.61–0.76], post: 0.81 [0.74–0.88]) and with reference standards involving clinical information (CT, MRI, etc.) (pre: 0.75 [0.68–0.83], post: 0.87 [0.81–0.93]). Lower specificity was likely associated with extremity fractures (pre: 0.87 [0.84–0.90], post: 0.91 [0.88–0.94]), high risk of bias (pre: 0.84 [0.79–0.90], post: 0.90 [0.85–0.95]), and reference standards involving clinical information (pre: 0.85 [0.80–0.90], post: 0.88 [0.83–0.93]). higher specificity appeared to be related to the use of the BoneView model (pre: 0.91 [0.88–0.94], post: 0.92 [0.88–0.96]). Before AI assistance, lower sensitivity was also associated with high risk of bias (0.76 [0.68–0.84]); after AI assistance, higher sensitivity was associated with extremity fractures (0.88 [0.84–0.92]) and high risk of bias (0.87 [0.81–0.94]).

Publication bias

The results of the funnel plots are shown in Figures 8 and 9. Both before and after AI assistance, the P-values were greater than 0.05, indicating that the plots were symmetrical and that there was no evidence of publication bias.

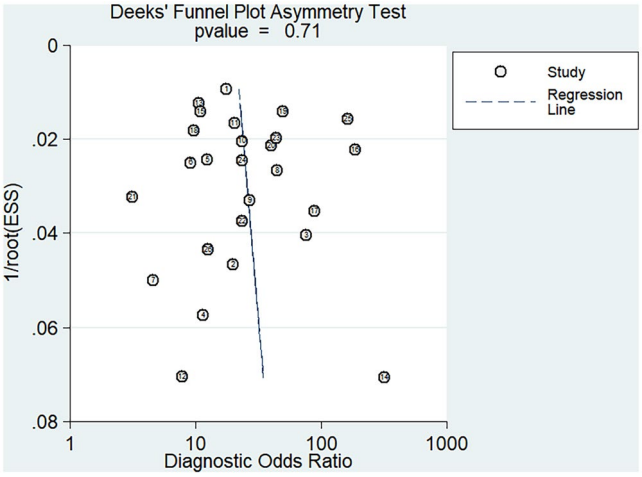

Figure 8. Funnel plot results before AI assistance in the included studies.

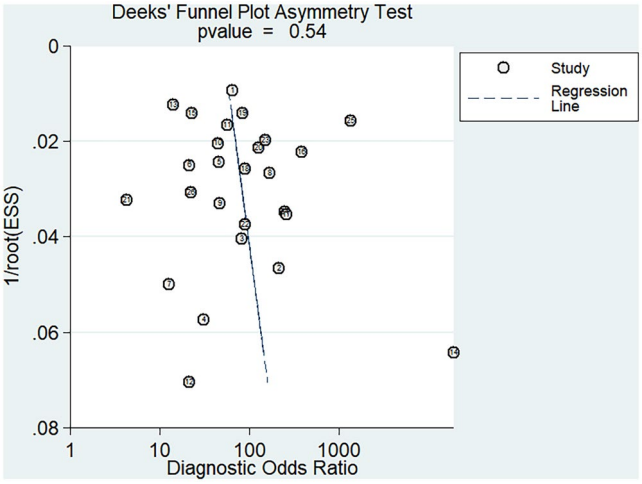

Figure 9. Funnel plot results after AI assistance in the included studies.

Discussion

Our study yielded three main findings. First, the introduction of artificial intelligence (AI) substantially improved the diagnostic performance of clinicians, with pooled sensitivity and specificity increasing to 87% (95% CI: 83–90) and 92% (95% CI: 89–94), respectively, and the AUC rising to 0.95 (95% CI: 0.93–0.97). Second, AI assistance did not significantly affect clinicians' reading speed. Finally, AI itself demonstrated high diagnostic accuracy, with an average standalone accuracy of 87.29%.

In our study, the definition of true positive (TP) was set as 'all fractures detected on a radiograph', which differs from the conventional criterion adopted in many studies, where detecting at least one fracture is considered a true positive. Our definition emphasizes the 'completeness of clinical diagnosis', as missing any fracture—particularly in emergency fracture screening—may delay treatment and lead to long-term functional impairment. Therefore, this definition better aligns with clinical practice needs and the safety goals of orthopedic and emergency departments. However, it may result in a lower apparent sensitivity compared with standard definitions, since cases with multiple fractures require identification of all lesions to be classified as true positives, thus raising the diagnostic threshold. This 'lower sensitivity' does not reflect a real decline in diagnostic ability, but rather a prioritization of diagnostic comprehensiveness under our definition. Consequently, direct comparisons between our results and studies using the standard TP definition may underestimate the true efficacy of AI-assisted diagnosis. Future research should aim to establish a standardized TP definition framework to ensure consistency and comparability across studies.

In the Quality Assessment, eight studies were rated as high risk of bias, and four studies were rated as high applicability concerns. Specifically, six studies were rated as high risk of bias in Patient Selection [35,37,47,49–51], and two studies were rated as high risk of bias in Flow and Timing [24,44]. These biases may have introduced multiple potential effects on the pooled results. In Patient Selection, many studies did not include consecutive or randomly selected cases, or included too few complex cases, which may have resulted in an overly simplified set of fracture radiographs. Consequently, both the pooled sensitivity and specificity before AI assistance may have been overestimated, leading to an underestimation of AI's true assisting effect. In Flow and Timing, cases with inconsistent interpretations or loss of readers were sometimes excluded, potentially omitting rare cases and introducing response bias (e.g. more self-disciplined or AI-interested participants were more likely to complete the experiment). This may have caused the pooled sensitivity and specificity after AI assistance to be inflated, thus overestimating the effectiveness of AI support.

Future research should aim to minimize bias interference by adopting rigorous random sampling, a unified reference standard, and comprehensive experimental design.

Emergency departments face the dual challenge of a continuously increasing number of trauma patients and the urgent need to rapidly identify high-risk injuries [54]. Mis-triage of fracture cases may lead to severe complications such as neurovascular injury or life-threatening hemorrhage. AI-assisted fracture detection provides a transformative solution by enabling rapid and accurate risk stratification, thereby optimizing the workflow of emergency triage. The study by John R. Zech et al. [33] demonstrated that, with AI assistance, clinicians' accuracy in reading fracture radiographs improved from 77.9% to 87.1%, while the average reading time decreased from 52.1 s to 38.9 s. This improvement can significantly enhance patient throughput in emergency departments. In particular, for junior radiologists, AI can serve as a 'second pair of eyes', helping to avoid under-triage of subtle but clinically significant fractures, and improving the overall triage efficiency without compromising diagnostic accuracy.

Fracture misdiagnosis is one of the major types of medical disputes, and the application of artificial intelligence (AI) in this field brings both opportunities and new challenges in forensic risk management [55]. By providing objective and traceable diagnostic results, AI enhances the accountability of medical diagnosis. For instance, if AI correctly identifies a fracture that is subsequently missed by a radiologist, it could strengthen the plaintiff's legal claim. Conversely, if AI also fails to detect an extremely subtle and difficult-to-recognize fracture, it could support the defense's argument regarding the inherent diagnostic difficulty. Moreover, AI assistance reduces the number of false negatives (FN), thereby directly decreasing the risk of malpractice claims arising from missed diagnoses. However, AI also raises new medico-legal questions regarding liability attribution. Globally, there is still no consensus on whether diagnostic errors

related to AI should be attributed to the radiologist, the AI developer, or the medical institution [56]. Existing legal frameworks, which were developed based on traditional medical models, are often insufficient to address the complexities introduced by AI. To mitigate these challenges, medical institutions should establish clear guidelines for the use of AI, defining the radiologist's ultimate responsibility for the final diagnostic interpretation, while requiring AI developers to disclose the core principles and validation datasets of their algorithms. Meanwhile, regulatory authorities should accelerate the development of comprehensive legal frameworks that delineate shared responsibilities among clinicians, developers, and healthcare institutions.

Although AI-based fracture detection systems using radiographs have demonstrated remarkable clinical value, they still face limitations in identifying complex fractures such as occult spinal fractures. Integrating AI with other imaging modalities (e.g. CT, MRI, and ultrasound) could overcome the diagnostic bottleneck of single-modality systems and further enhance clinical applicability [57]. For example, AI can enable cross-modal fusion analysis between radiographs and CT images. In cases where fractures are suspected on radiographs but cannot be clearly confirmed, the AI system can automatically link the corresponding CT images and accurately localize fracture lines through multi-plane reconstruction techniques.

With the continuous advancement of diagnostic technology, the amount of information contained in medical images has increased dramatically. This creates a challenge for radiologists and clinicians, who must balance the goal of improving diagnostic accuracy and patient satisfaction with maintaining workflow efficiency [58]. The rapid progress of deep learning has brought about a qualitative leap in image recognition. For example, the U-Net network, based on deep learning, has become one of the most popular architectures for medical image segmentation [59]. VGGNet is widely applied in image classification and object detection tasks, effectively identifying both object categories and locations in images [60]. ResNet has achieved remarkable results in various computer vision tasks such as image classification and semantic segmentation, driving the evolution of deep learning models toward deeper and more complex architectures [61]. These advances have made significant contributions to improving the level of medical diagnosis [62], demonstrating the potential of deep learning–based medical image analysis to be applied in Computer-Aided Diagnosis (CAD) systems. Such systems provide valuable decision support for clinicians, enhancing both the accuracy and efficiency of diagnostic and therapeutic processes [58].

To date, numerous AI applications for fracture detection have been developed [20,63,64]. One of the most prominent examples is BoneView, developed by Loïc Duron et al. [39] at Gleamer. They collected a dataset of 60,170 radiographs from trauma patients across 22 public hospitals and private radiology centers in France between January 2011 and May 2019. The dataset was randomly divided into 70% training, 10% validation, and 20% internal testing subsets. The model is a deep convolutional neural network built on the Detectron2 framework [65]. The AI system is integrated into radiology software as a diagnostic assistance tool that highlights each region of interest (ROI) with bounding boxes and provides a confidence score for the presence of a fracture within each ROI. With the continuous development of artificial intelligence, people are no longer solely focused on model development but have gradually gained a new understanding and demand for interpretability. In the field of medical imaging, researchers are increasingly using Explainable Artificial Intelligence (XAI) to interpret the results of their algorithms. If a method can provide deeper insights into how a neural network makes decisions and/or make those decisions more understandable, it can be considered a good explanation [66]. Therefore, it is evident that AI-assisted technologies are gradually maturing, and with the rapid advancement of science and technology, these technologies will be further optimized.

Among the included studies, only one focused on skull fractures [32]. This is primarily because current algorithmic models capable of detecting fractures in the craniofacial region remain underdeveloped. And due to the structural complexity of the craniofacial bones, CT imaging rather than radiographs is typically preferred for the assessment of head and facial injuries [67]. Among the common anatomical sites for AI-assisted fracture detection, the hand and foot have shown the highest detection accuracy [38,39]. These findings suggest that although AI-based fracture detection on radiographs has achieved considerable progress, it is still far from perfect and has substantial room for improvement. For instance, the Grad-CAM function can be used to generate heatmaps based on input images, highlighting regions of potential fractures to enhance model interpretability and diagnostic precision [68].

Our study has several limitations. First, we included only English-language studies, which may have led to the exclusion of relevant research published in other languages. Second, although our database searches initially identified a larger number of studies, some were excluded due to insufficient or unextractable data, increasing the likelihood of missing potentially relevant studies. Third, among all included papers, only the study conducted by Rikke Bachmann et al. [24] employed the most rigorous methodology, while the others introduced potential sources of bias that may have affected the accuracy of the meta-analytic estimates. Additionally, when searching for relevant studies, we conducted relevant searches only in PubMed, Web of Science, and the Cochrane Library, excluding EMBASE, IEEE Xplore, Scopus, and other databases, which may have led to the omission of relevant engineering or computer science AI studies. Finally, among the included studies, only seven explicitly reported that the proportion of occult (subtle) fractures was ≥25% [29,30,34,40,47,49,52]. In real-world clinical practice, such subtle fractures are the most likely to be missed in diagnosis. Furthermore, as noted earlier, only one study focused on skull fractures [32]. These limitations collectively constrain the generalizability and practical applicability of our conclusions.

In conclusion, AI assistance significantly improves the diagnostic accuracy of clinicians in interpreting radiographs, with an average accuracy of 87.6%. In future research, to obtain more reliable and generalizable conclusions, studies should include a sufficient number of patients and adopt more rigorous and standardized methodologies. For example, during radiograph reading sessions, clinicians could be restricted to a fixed time limit to better simulate the urgency and time constraints encountered in real-world clinical practice. Additionally, providing clinicians with more detailed patient information would help to recreate realistic diagnostic contexts. Establishing a more refined reference standard is also essential, as it would allow for the possibility that AI systems could even surpass human experts in specific diagnostic tasks. However, AI can never function independently of clinical practice—it is designed to serve and assist clinicians, not to replace them. Therefore, future algorithmic development and AI-based applications should prioritize user–clinician interaction, focusing on simplifying operational workflows and ensuring seamless integration into clinical routines, thereby enhancing both diagnostic efficiency and clinical usability.

Ethical approval

Institutional Review Board approval was not required because it is a systematic review and meta-analysis.

Author contributions

CRediT: **Han Qin**: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – original draft; **Yunxia Ding**: Conceptualization, Methodology, Project administration, Software, Validation, Writing – review & editing; **Jiangyi Ju**: Conceptualization, Data curation, Formal analysis, Methodology, Software; **Zhen Qu**: Conceptualization, Data curation, Methodology, Project administration, Visualization; **Lihua Peng**: Formal analysis, Project administration, Resources, Software, Supervision, Validation, Writing – review & editing.

Informed consent

Written informed consent was not required for this study because it is a systematic review and meta-analysis.

Disclosure statement

The authors declare no financial or non-financial competing interests related to this work. All authors confirm this disclosure is accurate.

Funding

The authors state that this work has not received any funding.

ORCID

Han Qin http://orcid.org/0009-0002-1032-4028
Yunxia Ding http://orcid.org/0009-0001-3980-3324
Jiangyi Ju http://orcid.org/0009-0001-7061-3318
Zhen Qu http://orcid.org/0000-0003-1879-8666
Lihua Peng http://orcid.org/0009-0004-6518-3458

Data availability statement

The data supporting the findings of this study are available upon reasonable request from the corresponding author. All relevant data generated or analyzed during this study are included in the supplementary materials of this article, which can be accessed as part of the manuscript submission. For any additional information or data inquiries, please contact [Lihua Peng] at [140733@hospital.cqmu.edu.cn].

References

[1] Singh H, Schiff GD, Graber ML, et al. The global burden of diagnostic errors in primary care. BMJ Qual Saf. 2017;26(6):484–494. doi: 10.1136/bmjqs-2016-005401.
[2] Fernholm R, Pukk Härenstam K, Wachtler C, et al. Diagnostic errors reported in primary healthcare and emergency departments: a retrospective and descriptive cohort study of 4830 reported cases of preventable harm in Sweden. Eur J Gen Pract. 2019;25(3):128–135. doi: 10.1080/13814788.2019.1625886.
[3] Gergenti L, Olympia RP. Etiology and disposition associated with radiology discrepancies on emergency department patients. Am J Emerg Med. 2019;37(11):2015–2019. doi: 10.1016/j.ajem.2019.02.027.
[4] Mattijssen-Horstink L, Langeraar JJ, Mauritz GJ, et al. Radiologic discrepancies in diagnosis of fractures in a Dutch teaching emergency department: a retrospective analysis[J]. Scand J Trauma Resusc Emerg Med. 2020;28(1):38. doi: 10.1186/s13049-020-00727-8.
[5] Guly HR. Diagnostic errors in an accident and emergency department. Emerg Med J. 2001;18(4):263–269. doi: 10.1136/emj.18.4.263.
[6] Krogue JD, Cheng KV, Hwang KM, et al. Automatic Hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell. 2020;2(2):e190023. doi: 10.1148/ryai.2020190023.
[7] Sakai K, Yamada K. Machine learning studies on major brain diseases: 5-year trends of 2014-2018. Jpn J Radiol. 2019;37(1):34–72. doi: 10.1007/s11604-018-0794-4.
[8] Nakata N. Recent technical development of artificial intelligence for diagnostic medical imaging. Jpn J Radiol. 2019;37(2):103–108. doi: 10.1007/s11604-018-0804-6.
[9] Schmauch B, Herent P, Jehanno P, et al. Diagnosis of focal liver lesions from ultrasound using deep learning. Diagn Interv Imaging. 2019;100(4):227–233. doi: 10.1016/j.diii.2019.02.009.
[10] Roca P, Attye A, Colas L, et al. Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. Diagn Interv Imaging. 2020;101(12):795–802. doi: 10.1016/j.diii.2020.05.009.
[11] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–2410. doi: 10.1001/jama.2016.17216.
[12] Sirinukunwattana K, Ahmed Raza SE, Tsang Y-W, et al. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging. 2016;35(5):1196–1206. doi: 10.1109/TMI.2016.2525803.
[13] Yoon AP, Lee YL, Kane RL, et al. Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. JAMA Netw Open. 2021;4(5):e216096. doi: 10.1001/jamanetworkopen.2021.6096.
[14] Choi J, Hui JZ, Spain D, et al. Practical computer vision application to detect hip fractures on pelvic X-rays: a bi-institutional study. Trauma Surg Acute Care Open. 2021;6(1):e000705. doi: 10.1136/tsaco-2021-000705.
[15] Kuo RYL, Harrison C, Curran TA, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology. 2022;304(1):50–62. doi: 10.1148/radiol.211785.
[16] Sun H, Wang X, Li Z, et al. Automated rib fracture detection on chest X-ray using contrastive learning. J Digit Imaging. 2023;36(5):2138–2147. doi: 10.1007/s10278-023-00868-z.
[17] Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115(45):11591–11596. doi: 10.1073/pnas.1806905115.
[18] Guo L, Zhou C, Xu J, et al. Deep learning for chest X-ray diagnosis: competition between radiologists with or without artificial intelligence assistance. J Imaging Inform Med. 2024;37(3):922–934. doi: 10.1007/s10278-024-00990-6.
[19] Husarek J, Hess S, Razaeian S, et al. Artificial intelligence in commercial fracture detection products: a systematic review and meta-analysis of diagnostic test accuracy. Sci Rep. 2024;14(1):23053. doi: 10.1038/s41598-024-73058-8.
[20] Kutbi M. Artificial intelligence-based applications for bone fracture detection using medical images: a systematic review. Diagnostics (Basel)). 2024;14(17):1879. doi: 10.3390/diagnostics14171879.

[21] Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097. doi: 10.1371/journal.pmed.1000097.
[22] McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA. 2018;319(4):388–396. Erratum in: JAMA. 2019 Nov 26;322(20):2026. doi: 10.1001/jama.2019.18307. doi: 10.1001/jama.2017.19163.
[23] Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–536. doi: 10.7326/0003-4819-155-8-201110180-00009.
[24] Bachmann R, Gunes G, Hangaard S, et al. Improving traumatic fracture detection on radiographs with artificial intelligence support: a multi-reader study. BJR Open. 2024;6(1):tzae011. doi: 10.1093/bjro/tzae011.
[25] Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982–990. doi: 10.1016/j.jclinepi.2005.02.022.
[26] Novak A, Hollowday M, Espinosa Morgado AT, et al. Evaluating the impact of artificial intelligence-assisted image analysis on the diagnostic accuracy of front-line clinicians in detecting fractures on plain X-rays (FRACT-AI): protocol for a prospective observational study. BMJ Open. 2024;14(9):e086061. doi: 10.1136/bmjopen-2024-086061.
[27] Shelmerdine SC, Pauling C, Allan E, et al. Artificial intelligence (AI) for paediatric fracture detection: a multireader multicase (MRMC) study protocol. BMJ Open. 2024;14(12):e084448. doi: 10.1136/bmjopen-2024-084448.
[28] Cheng CT, Chen CC, Cheng FJ, et al. A human-algorithm integration system for hip fracture detection on plain radiography: system development and validation study. JMIR Med Inform. 2020;8(11):e19416. doi: 10.2196/19416.
[29] Guermazi A, Tannoury C, Kompel AJ, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. Radiology. 2022;302(3):627–636. Epub 2021 Dec 21. doi: 10.1148/radiol.210937.
[30] Zech JR, Carotenuto G, Igbinoba Z, et al. Detecting pediatric wrist fractures using deep-learning-based object detection. Pediatr Radiol. 2023;53(6):1125–1134. doi: 10.1007/s00247-023-05588-8.
[31] Cohen M, Puntonet J, Sanchez J, et al. Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs. Eur Radiol. 2023;33(6):3974–3983. Epub 2022 Dec 14. doi: 10.1007/s00330-022-09349-3.
[32] Choi JW, Cho YJ, Ha JY, et al. Deep learning-assisted diagnosis of pediatric skull fractures on plain radiographs. Korean J Radiol. 2022;23(3):343–354. doi: 10.3348/kjr.2021.0449.
[33] Zech JR, Ezuma CO, Patel S, et al. Artificial intelligence improves resident detection of pediatric and young adult upper extremity fractures. Skeletal Radiol. 2024;53(12):2643–2651. doi: 10.1007/s00256-024-04698-0.
[34] Yoon AP, Chung WT, Wang CW, et al. Can a deep learning algorithm improve detection of occult scaphoid fractures in plain radiographs? A clinical validation study. Clin Orthop Relat Res. 2023;481(9):1828–1835. doi: 10.1097/CORR.0000000000002612.
[35] Meetschen M, Salhöfer L, Beck N, et al. AI-assisted X-ray fracture detection in residency training: evaluation in pediatric and adult trauma patients. Diagnostics (Basel)). 2024;14(6):596. doi: 10.3390/diagnostics14060596.
[36] Canoni-Meynet L, Verdot P, Danner A, et al. Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. Diagn Interv Imaging. 2022;103(12):594–600. Epub 2022 Jun 29. doi: 10.1016/j.diii.2022.06.004.
[37] Hendrix N, Hendrix W, van Dijke K, et al. Musculoskeletal radiologist-level performance by using deep learning for detection of scaphoid fractures on conventional multi-view radiographs of hand and wrist. Eur Radiol. 2023;33(3):1575–1588. doi: 10.1007/s00330-022-09205-4.
[38] Nguyen T, Maarek R, Hermann AL, et al. Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. Pediatr Radiol. 2022;52(11):2215–2226. Epub 2022 Sep 28. doi: 10.1007/s00247-022-05496-3.
[39] Duron L, Ducarouge A, Gillibert A, et al. Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study. Radiology. 2021;300(1):120–129. Epub 2021 May 4. doi: 10.1148/radiol.2021203886.
[40] Dell'Aria A, Tack D, Saddiki N, et al. Radiographic detection of post-traumatic bone fractures: contribution of artificial intelligence software to the analysis of senior and junior radiologists. J Belg Soc Radiol. 2024;108(1):44. doi: 10.5334/jbsr.3574.
[41] Jacques T, Cardot N, Ventre J, et al. Commercially-available AI algorithm improves radiologists' sensitivity for wrist and hand fracture detection on X-ray, compared to a CT-based ground truth. Eur Radiol. 2024;34(5):2885–2894. doi: 10.1007/s00330-023-10380-1.
[42] Keller M, Rohner M, Honigmann P. The potential benefit of artificial intelligence regarding clinical decision-making in the treatment of wrist trauma patients. J Orthop Surg Res. 2024;19(1):579. doi: 10.1186/s13018-024-05063-6.
[43] Oppenheimer J, Lüken S, Hamm B, et al. A prospective approach to integration of AI fracture detection software in radiographs into clinical workflow. Life (Basel)). 2023;13(1):223. doi: 10.3390/life13010223.
[44] Cheng CT, Chen CC, Fu CY, et al. Artificial intelligence-based education assists medical students' interpretation of hip fracture. Insights Imaging. 2020;11(1):119. doi: 10.1186/s13244-020-00932-0.
[45] Ziegner M, Pape J, Lacher M, et al. Real-life benefit of artificial intelligence-based fracture detection in a pediatric emergency department. Eur Radiol. 2025;35(10):5881–5890. doi: 10.1007/s00330-025-11554-9.
[46] Breu R, Avelar C, Bertalan Z, et al. Artificial intelligence in traumatology. Bone Joint Res. 2024;13(10):588–595. doi: 10.1302/2046-3758.1310.BJR-2023-0275.R3.

[47] O'Rourke S, Xu S, Carrero S, et al. AI as teacher: effectiveness of an AI-based training module to improve trainee pediatric fracture detection. Skeletal Radiol. 2025;54(9):1949–1957. doi: 10.1007/s00256-025-04927-0.
[48] Pauling C, Laidlow-Singh H, Evans E, et al. External validation of an artificial intelligence tool for fracture detection in children with osteogenesis imperfecta: a multireader study. Eur Radiol. 2026;36(1):515–525. doi: 10.1007/s00330-025-11790-z.
[49] Xing P, Zhang L, Wang T, et al. A deep learning algorithm that aids visualization of femoral neck fractures and improves physician training. Injury. 2024;55(12):111997. doi: 10.1016/j.injury.2024.111997.
[50] M Yogendra P, Goh AGW, Yee SY, et al. Accuracy of radiologists and radiology residents in detection of paediatric appendicular fractures with and without artificial intelligence. BMJ Health Care Inform. 2024;31(1):e101091. doi: 10.1136/bmjhci-2024-101091.
[51] Xie Y, Chen X, Yang H, et al. Integrating blockchain technology with artificial intelligence for the diagnosis of tibial plateau fractures. Eur J Trauma Emerg Surg. 2025;51(1):119. doi: 10.1007/s00068-025-02793-y.
[52] Hernáiz Ferrer AI, Bortolotto C, Carone L, et al. Application of artificial intelligence in the diagnosis of scaphoid fractures: impact of automated detection of scaphoid fractures in a real-life study. Radiol Med. 2025;130(10):1633–1641. doi: 10.1007/s11547-025-02028-5.
[53] Liu P, Lu L, Chen Y, et al. Artificial intelligence to detect the femoral intertrochanteric fracture: the arrival of the intelligent-medicine era. Front Bioeng Biotechnol. 2022;10:927926. doi: 10.3389/fbioe.2022.927926.
[54] Raita Y, Goto T, Faridi MK, et al. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. 2019;23(1):64. doi: 10.1186/s13054-019-2351-7.
[55] Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci. 2024;19(1):27. doi: 10.1186/s13012-024-01357-9.
[56] Pai SN, Jeyaraman M, Jeyaraman N, et al. Doctor, bot, or both: questioning the medicolegal liability of artificial intelligence in indian healthcare. Cureus. 2024;16(9):e69230. doi: 10.7759/cureus.69230.
[57] Li Y, Wang J, Pan X, et al. MRI-mediated intelligent multimodal imaging system: from artificial intelligence to clinical imaging diagnosis. Drug Discov Today. 2025;30(7):104399. doi: 10.1016/j.drudis.2025.104399.
[58] Chan HP, Samala RK, Hadjiiski LM, et al. Deep learning in medical image analysis. Adv Exp Med Biol. 2020;1213:3–21. doi: 10.1007/978-3-030-33128-3_1.
[59] Kundu S, Karale V, Ghorai G, et al. Nested U-net for segmentation of red lesions in retinal fundus images and sub-image classification for removal of false positives. J Digit Imaging. 2022;35(5):1111–1119. doi: 10.1007/s10278-022-00629-4.
[60] Lin K, Zhou T, Gao X, et al. Deep convolutional neural networks for construction and demolition waste classification: VGGNet structures, cyclical learning rate, and knowledge transfer. J Environ Manage. 2022;318:115501. doi: 10.1016/j.jenvman.2022.115501.
[61] Xu W, Fu YL, Zhu D. ResNet and its application to medical image processing: research progress and challenges. Comput Methods Programs Biomed. 2023;240:107660. doi: 10.1016/j.cmpb.2023.107660.
[62] Zhou T, Cheng Q, Lu H, et al. Deep learning methods for medical image fusion: a review. Comput Biol Med. 2023;160:106959. doi: 10.1016/j.compbiomed.2023.106959.
[63] AlGhaithi A, Al Maskari S. Artificial intelligence application in bone fracture detection. J Musculoskelet Surg Res. 2021;5(1):4. doi: 10.4103/jmsr.jmsr_132_20.
[64] Lo Mastro A, Grassi E, Berritto D, et al. Artificial intelligence in fracture detection on radiographs: a literature review. Jpn J Radiol. 2025;43(4):551–585. Epub ahead of print. doi: 10.1007/s11604-024-01702-4.
[65] Wu Y, Kirillov A, Massa F, et al. Detectron2. 2019. Available from: https://github.com/facebookresearch/detectron2
[66] van der Velden BHM, Kuijf HJ, Gilhuijs KGA, et al. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470. doi: 10.1016/j.media.2022.102470.
[67] Pham TD, Holmes SB, Coulthard P. A review on artificial intelligence for the diagnosis of fractures in facial trauma imaging. Front Artif Intell. 2023;6:1278529. doi: 10.3389/frai.2023.1278529.
[68] Kraus M, Anteby R, Konen E, et al. Artificial intelligence for X-ray scaphoid fracture detection: a systematic review and diagnostic test accuracy meta-analysis. Eur Radiol. 2024;34(7):4341–4351. doi: 10.1007/s00330-023-10473-x.