

Diagnostic Accuracy of Artificial Intelligence Models for the Detection of Ankle and Foot Fractures: A Systematic Review and Meta-analysis

Danial Sharifi Razavi¹, Mohammadali Seiri², Mehrdad Farrokhi^{3, 4*}

1. Department of Orthopedics, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran.

2. Department of Orthopedics, School of Medicine, Iran University of Medical Sciences, Tehran, Iran.

3. ERIS Research Institute, Tehran, Iran.

4. Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Received: July 2026; Accepted: August 2026; Published online: October 2026

Abstract: **Introduction:** Foot and ankle fractures are common injuries that may be missed on radiographs, particularly when subtle or complex fracture patterns are present. Artificial intelligence (AI) has emerged as a potential adjunct to radiographic interpretation, but reported diagnostic performance varies across studies and AI models. Therefore, a comprehensive meta-analysis is warranted to provide a more robust understanding of the diagnostic performance and applicability of AI models in this field. **Methods:** In this systematic review and meta-analysis, we searched MEDLINE and PubMed Central (PMC) via PubMed, Scopus, and Web of Science from inception to September 2026. Diagnostic accuracy measures were pooled using a bivariate random-effects model, and subgroup analyses were performed to explore potential sources of heterogeneity. **Results:** The pooled sensitivity was 0.90 (95% confidence interval (CI): 0.86–0.92), while the pooled specificity was 0.88 (95% CI: 0.82–0.93). The pooled positive likelihood ratio (PLR) was 7.6 (95% CI: 4.8–12), and the pooled negative likelihood ratio (NLR) was 0.12 (95% CI: 0.09–0.16). The area under the summary receiver operating characteristic (SROC) curve was 0.94. **Conclusion:** AI models demonstrate strong diagnostic performance for detecting foot and ankle fractures on radiographs. Diagnostic performance differed by anatomical region, with higher sensitivity for foot fractures and higher specificity for ankle fractures. These findings support AI as a promising adjunct to radiographic interpretation, although prospective, multicenter studies with external validation are needed to confirm its clinical utility and real-world effectiveness.

Keywords: Ankle Fractures; Artificial Intelligence; Deep Learning; Diagnosis; Fractures; Bone; Foot Injuries; Meta-Analysis

Cite this article as: Razavi DS, Seiri M, Farrokhi M. Diagnostic Accuracy of Artificial Intelligence Models for the Detection of Ankle and Foot Fractures: A Systematic Review and Meta-analysis. Arch Acad Emerg Med. 2026;14(1):e57. <https://doi.org/10.22037/aaem.v14i1.3163>.

1. Introduction

The ankle is a complex biomechanical joint that, in concert with the 26 bones of the foot, is essential for maintaining lower-extremity stability, transmitting and supporting body weight, and enabling efficient locomotor functions such as walking (1, 2). Disruption of this intricate musculoskeletal system through fracture can substantially impair mobility and functional independence. The incidence of lower-extremity fractures, coupled with prolonged recovery and variable post-intervention outcomes, represents a substantial clinical and socioeconomic burden and remains an important challenge for healthcare systems (3-5). Ankle and foot injuries are a common reason for emergency department visits, and their early diagnosis is essential for ensuring appropriate intervention and preventing complications that may affect joint stability, mobility, and patients' quality of life (6). The complex anatomy of the ankle and foot, comprising a large number of bones, contributes to the difficulty of detecting fractures and may result in a higher rate of misdiagnosis, with an overall diagnostic error rate of 4.2% (7, 8). A wide variety of imaging modalities can be used to diagnose foot and ankle fractures, with X-ray radiography being the most common primary imaging modality. Although multiple radiographic views are routinely obtained, fracture lines may be subtle or obscured and can therefore be missed, necessitating further evaluation with other imaging

modalities, such as computed tomography (CT) or magnetic resonance imaging (MRI) (9). Missed fractures can occur for various reasons, including the subtle appearance of the fracture and the inherent limitations of the imaging modality. However, an important primary contributing factor is radiologist error, particularly incorrect interpretation of radiographic findings, which can be reduced through appropriate training and experience (10).

Artificial intelligence (AI) models have been introduced as a complementary approach to reducing the rate of radiological misinterpretation errors (11, 12). This has been facilitated by the increasing availability of large radiological datasets for model training, together with rapid advances in computer-assisted imaging technologies (11, 12).

While reducing radiologist interpretation errors is a potential benefit of AI technology, it should be recognized that the diagnostic performance of AI models varies considerably across different types of fractures and among different AI models (13, 14). Moreover, although several original studies have investigated the diagnostic performance of AI models for detecting foot and ankle fractures, their reported results have shown substantial variability in accuracy (15, 16). Thus, a comprehensive meta-analysis is warranted to provide a more robust understanding of the diagnostic performance and applicability of AI models in this field.

2. Methods

2.1. Study design and setting

*Corresponding Author: Mehrdad Farrokhi; Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: drfarrokhi2024@gmail.com; ORCID: 0000-0002-1559-2323; Tel: 00989384226664

We registered the protocol for this systematic review in PROSPERO (CRD420261500297). The review was conducted in accordance with the diagnostic test accuracy extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-DTA) guidelines (17).

2.2. Search strategy

We searched MEDLINE and PubMed Central (PMC) via PubMed, Scopus, and Web of Science from inception to September 2026. In PubMed, we combined Medical Subject Headings (MeSH) terms and synonymous keywords representing three main concepts: ankle and foot fractures, artificial intelligence (AI), and diagnosis. Terms within each concept were combined using the Boolean operator OR, and the three concepts were subsequently combined using AND. For Scopus and Web of Science, keyword-based search strategies were developed using the same three concepts and Boolean operators. The complete search strategies for all databases are provided in Supplementary Material 1.

We also manually screened the reference lists of eligible studies and examined studies that cited the included articles to identify additional potentially eligible studies. A gray literature search was conducted using similar search terms to identify potentially relevant records that may have been missed by the database searches. All retrieved records were exported to End-Note software, where automated duplicate detection and removal were performed.

2.3. Eligibility criteria

We included original diagnostic accuracy studies that evaluated the performance of artificial intelligence (AI) models for the detection of ankle or foot fractures in adults. Eligible studies were required to report sufficient data to construct a 2×2 contingency table, including true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), or to provide sufficient information to calculate these values. We excluded studies conducted exclusively in pediatric populations; studies that did not provide sufficient data to construct a 2×2 contingency table; editorials; narrative reviews; letters that did not report original research; case reports or case series involving fewer than 10 participants; and studies evaluating the risk or prediction of future fractures rather than the detection of existing fractures.

Two authors independently screened titles and abstracts of all retrieved records. Potentially eligible studies were subsequently assessed by the same authors through full-text review against the predefined eligibility criteria. Any disagreements at either stage were resolved through discussion and, when necessary, consultation with a third author.

2.4. Data extraction and quality assessment

The following data were extracted using a prespecified Excel data extraction form: first author's surname, participant age, sex distribution, anatomical region (ankle or foot), AI model, imaging modality used as input data, TP, FP, TN, and FN values, sensitivity, and specificity. Where available, sensitivity and specificity were extracted as reported; otherwise, they were calculated from the extracted 2×2 contingency table.

The methodological quality and risk of bias of the included studies were assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool. Data extraction and quality assessment were performed independently by two authors, with disagreements resolved through discussion and,

when necessary, adjudication by a third author.

2.5. Statistical analysis

Diagnostic accuracy measures were synthesized using the MIDAS package in Stata. Forest plots were generated to present study-level and pooled estimates of the main diagnostic parameters, including sensitivity and specificity. Summary receiver operating characteristic (SROC) curves were generated to assess the overall diagnostic performance of the AI models. The MIDAS package in Stata was used to obtain pooled estimates of diagnostic accuracy and to perform model checking and analyses of clinical applicability. Fagan nomograms and probability-modifying plots were generated to evaluate the potential clinical utility of AI models for the detection of foot and ankle fractures. Potential publication bias was assessed using Deeks' funnel plot asymmetry test. Sources of between-study heterogeneity were explored using subgroup analyses and meta-regression, where sufficient studies were available. The pooled accuracy estimate was calculated using the metaprop command in Stata with the exact method (cimethod(exact)).

3. Results

3.1. Overview of the included studies

A total of 1,276 studies were retrieved from the three investigated databases, with an additional six studies identified through a gray literature search. After automated deduplication, 314 duplicate records were removed, leaving 968 studies for title and abstract screening. Following title and abstract screening, 38 studies were eligible for full-text assessment. Of these, 10 studies, comprising 18 substudies, were ultimately included in the quantitative meta-analysis (Figure 1).

A total of 10 papers published between 2019 and 2025 were included in the systematic review, comprising 18 substudies reporting different AI models and/or anatomical regions. Participant ages varied considerably, ranging from approximately 18 to 98 years, while sex distribution varied across studies. The ankle was the most frequently investigated anatomical region, examined in 10 study entries, followed by the foot in 7 study entries, while one study investigated both the foot and ankle. Radiography was the predominant imaging modality across all included studies. A variety of AI approaches were evaluated, including convolutional neural networks, ensemble models, multitask deep neural networks, and object-detection/segmentation models, as well as commercially available AI systems. Reported diagnostic performance was generally high, with sensitivity ranging from 78% to 98.1% and specificity ranging from 62% to 100% (Table 1).

3.2. Publication bias and quality assessment

As shown in Figure 2, Deeks' funnel plot and asymmetry test indicated no evidence of publication bias among the included studies ($p = 0.29$).

Overall, most studies demonstrated a low risk of bias in index test, reference standard, and flow and timing domains. Some concerns were identified, mainly in the risk of bias of patient selection domain and its applicability concerns. These findings indicate that the included studies provided generally reliable evidence for the diagnostic accuracy assessment of AI models for detection of foot and ankle fractures (Table 2).

3.3. Meta-analyses

The pooled sensitivity was 0.90 (95% confidence interval (CI):

0.86–0.92), while the pooled specificity was 0.88 (95% CI: 0.82–0.93) (Figure 3). The pooled positive likelihood ratio (PLR) was 7.6 (95% CI: 4.8–12), and the pooled negative likelihood ratio (NLR) was 0.12 (95% CI: 0.09–0.16). The pooled diagnostic odds ratio (DOR) was 64 (95% CI: 33–123). The area under the summary receiver operating characteristic (SROC) curve was 0.94 (Figure 4). Pooled accuracy of AI models for detection of foot and ankle fractures was 0.86 (95% CI: 0.81–0.91) (Figure 5).

We found significant heterogeneity between included studies with respect to the sensitivity, specificity, DOR, NLR, and PLR ($I^2 > 50\%$ and $p < 0.05$). Bivariate boxplot showed five outliers in the plot indirectly indicating evidence of heterogeneity (Figure 6). The empirical Bayes prediction plot demonstrated close concordance between model-based and observed estimates of sensitivity and specificity across studies (Figure 7), supporting the adequacy of the fitted bivariate random-effects model.

The goodness-of-fit plot demonstrated a generally acceptable agreement between observed deviance residuals and the normal quantile, indicating an adequate model fit (Figure 8a). The bivariate normality assessment showed that the Mahalanobis distance values approximately followed the chi-square quantile, supporting the assumption of bivariate normality of the model (Figure 8b). Influence analysis identified study 17 as the most influential observation, with a Cook's distance exceeding the predefined threshold, while study 6 showed a moderate influence (Figure 8c). However, the outlier detection analysis indicated that most studies were located within the acceptable range of standardized residuals, with only study 17 showing a potential outlying effect (Figure 8d).

Fagan's nomogram demonstrated that a positive result from the AI models for ankle and foot fractures increased the pretest probability of fracture from 41% to 84%, whereas a negative result reduced the probability to 8% (Figure 9). The likelihood ratio plot demonstrated a similar diagnostic ability for both positive and negative test results, as the curves showed a relatively symmetric pattern. However, the negative test result curve was slightly closer to the (1,0) region, suggesting a marginally stronger ability to reduce disease probability compared with the ability of a positive result to confirm disease (Figure 10).

3.4. Subgroup and sensitivity analyses

Subgroup analyses showed that studies published in 2023 or later had significantly higher pooled sensitivity ($p < 0.001$). Studies with a sample size of ≥ 200 participants and those with a fracture prevalence of ≥ 0.50 also demonstrated significantly higher pooled sensitivity ($p < 0.001$). Studies evaluating clinically or commercially available AI models demonstrated significantly higher sensitivity than those evaluating deep learning models ($p < 0.001$). Furthermore, AI models demonstrated higher sensitivity for the detection of foot fractures ($p < 0.001$), whereas specificity was higher for the detection of ankle fractures ($p < 0.05$) (Figure 11).

The sensitivity analysis using one estimate per cohort, selecting the estimate with the highest area under the curve (AUC) among the available estimates, yielded results broadly similar to those of the primary analysis, with a sensitivity of 0.90 (also 0.90 in the primary analysis) and an increased specificity of 0.92 (compared with 0.88 in the primary analysis) (Figure 12).

A sensitivity analysis was performed by excluding studies with potentially different target conditions or imaging protocols,

including Wang et al. (fatigue fractures), Taseh et al. (fifth metatarsal fractures evaluated using smartphone photographs of radiographs), and Kim et al. (single-view radiographs). The pooled diagnostic performance remained comparable to the primary analysis, with a pooled sensitivity of 0.88 and specificity of 0.87 (Figure 13), suggesting that the overall findings were robust and not substantially influenced by these studies.

4. Discussion

In this systematic review and meta-analysis, we found that AI models demonstrated high overall diagnostic performance for the detection of foot and ankle fractures on radiographs. Across 10 studies comprising 18 substudies, the pooled sensitivity was 90% and specificity was 88%, with a SROC AUC of 0.94 and an overall pooled accuracy of 86%. The pooled PLR and NLR were 7.6 and 0.12, respectively, corresponding to a DOR of 64 and indicating the substantial ability of AI to discriminate between fracture and non-fracture cases. Fagan's nomogram further demonstrated clinically meaningful changes in post-test probability, with a positive AI result increasing the probability of fracture from 41% to 84%, while a negative result reduced it to 8%. Although significant heterogeneity was observed across the included studies, the fitted bivariate random-effects model showed adequate fit and generally consistent agreement between observed and model-based estimates. Subgroup analyses indicated that diagnostic sensitivity was significantly higher in more recent studies, studies with larger sample sizes or higher fracture prevalence, studies evaluating clinically or commercially available AI systems, and studies assessing foot fractures, whereas specificity was higher for ankle fractures. Collectively, these findings indicate that AI models exhibit robust diagnostic performance for the detection of foot and ankle fractures on radiographs and may serve as valuable adjunctive tools in clinical practice, with particular potential to enhance fracture detection and support diagnostic decision-making.

A recently published systematic review by Mangwani et al. (18) evaluated the diagnostic accuracy of machine learning and deep learning models for the detection of ankle fractures using a narrative rather than quantitative synthesis. Although their review included a larger number of studies (27 studies) than ours (10 papers comprising 18 substudies), this difference is primarily attributable to differences in eligibility criteria and synthesis methodology. We searched the databases from inception through September 2026, whereas Mangwani et al. searched through June 2025. Importantly, our review required sufficient data to construct 2×2 contingency tables, thereby enabling quantitative meta-analysis, whereas their review did not impose this requirement because its findings were synthesized narratively. Consequently, our inclusion criteria were more restrictive with respect to the availability of diagnostic accuracy data, which likely explains the smaller number of eligible papers. The distribution of AI models also differed between the two reviews. Mangwani et al. reported that ResNet and InceptionV3 were among the most frequently evaluated models, whereas our review identified a broader range of AI approaches including deep learning architectures, ensemble approaches, as well as clinically and commercially available AI systems. This distinction is particularly relevant because our subgroup analysis demonstrated significantly higher sensitivity among clinically or commercially available AI models compared with deep learning models. These findings suggest that diagnostic performance may vary not only

according to the underlying AI architecture but also according to the degree of clinical development and real-world implementation of the system. Mangwani et al. concluded that machine learning has substantial potential as a supplementary diagnostic approach for ankle fracture detection, based primarily on the direction and consistency of findings across individual studies. Our findings extend this evidence by providing a quantitative synthesis of diagnostic performance. Specifically, the pooled estimates demonstrated strong overall diagnostic performance, with meaningful effects on post-test probability following both positive and negative AI results. Taken together, these findings provide quantitative support for the potential role of AI as an adjunctive diagnostic tool for foot and ankle fractures and suggest that its clinical utility may extend beyond fracture detection alone to supporting the reliable exclusion of fractures when AI results are negative.

Another meta-analysis by Prasetyo et al. (19) evaluated the diagnostic performance of AI models for the detection of ankle and foot fractures. Although their database search extended through May 2026, only five studies were included in their quantitative synthesis, yielding a pooled sensitivity of 92.4% and specificity of 95.1%. These estimates were higher than those observed in our meta-analysis. The most plausible explanation for this discrepancy is the substantially larger number of studies included in our quantitative synthesis, which may have resulted in more comprehensive representation of the heterogeneity in AI models, study populations, imaging characteristics, and clinical settings. Accordingly, the higher estimates reported by Prasetyo et al. should be interpreted in the context of their smaller evidence base and potentially narrower study characteristics. In addition to providing a more extensive quantitative synthesis, our study incorporated clinically relevant subgroup analyses that further characterized sources of variation in diagnostic performance. Specifically, we separately evaluated the diagnostic performance of AI models for foot and ankle fractures and conducted a dedicated analysis of clinically or commercially available AI models via subgroup analyses. These analyses demonstrated important differences in performance across anatomical regions and AI model types, providing additional insights into the factors that may influence the diagnostic utility of AI in fracture detection. Thus, our findings not only provide pooled estimates based on a broader evidence base but also offer a more detailed assessment of the conditions under which AI models may provide the greatest diagnostic benefit.

Husarek et al. (20) recently conducted a comprehensive systematic review and meta-analysis of AI-based fracture detection on CT, with particular emphasis on the diagnostic performance of commercial versus non-commercial AI solutions. Their analysis demonstrated that commercial AI fracture detection solutions (CAAI-FDS) had a pooled sensitivity of 0.77 and specificity of 0.93, whereas non-CAAI-FDS achieved a higher pooled sensitivity of 0.88 with a comparable specificity of 0.92. Thus, commercially available solutions showed lower sensitivity but essentially similar specificity compared with non-commercial AI systems, suggesting that commercial availability does not necessarily translate into superior diagnostic performance. These findings differ from our subgroup analysis, in which clinically or commercially available AI models demonstrated significantly higher sensitivity than deep learning models. Several factors may account for this apparent discrepancy, including differences in imaging modality and anatomical region, as Husarek et al. focused on CT-based

fracture detection across multiple anatomical sites, whereas our analysis specifically evaluated foot and ankle fractures on radiographs. Nevertheless, the findings from both studies underscore the importance of evaluating AI performance according to the specific clinical context and model characteristics rather than assuming that commercial availability alone is indicative of superior diagnostic accuracy.

Despite these encouraging findings, AI-based fracture detection should currently be regarded as a supplementary or adjunct tool rather than a replacement for radiologist interpretation. Evidence from a large meta-analysis comparing AI algorithms with clinicians found that radiologists were superior to AI algorithms overall, mainly due to higher specificity across multiple fracture sites. The authors concluded that AI algorithms are on par with non-radiologists in fracture detection but remain inferior to radiologists in terms of overall detection accuracy (21).

In the subgroup analysis based on disease prevalence, studies with a prevalence ≥ 0.50 demonstrated higher sensitivity compared with those with lower prevalence. However, this finding should be interpreted cautiously. Studies with high disease prevalence may include case-control designs or enriched cohorts, in which patients with the target condition are more clearly represented and may differ substantially from those encountered in routine clinical practice. Such study designs can introduce spectrum bias by creating a greater separation between positive and negative cases, potentially leading to an overestimation of diagnostic performance. Therefore, the higher sensitivity observed in high-prevalence studies should not be interpreted as evidence of superior model performance, but rather as a potential consequence of differences in study population characteristics and disease spectrum.

4.1. Limitations

This study has several limitations that should be considered when interpreting its findings. First, despite a comprehensive search strategy, only 10 papers comprising 18 substudies met the predefined eligibility criteria and provided sufficient data for quantitative synthesis. The relatively limited evidence base may reduce the precision and generalizability of the pooled estimates and precludes more extensive subgroup analyses of potentially relevant sources of variation. Second, substantial heterogeneity was observed across several diagnostic performance measures, including sensitivity, specificity, PLR, NLR, and DOR. This heterogeneity likely reflects differences in patient populations, fracture prevalence, anatomical regions, AI architectures, model development and validation strategies, and clinical settings across the included studies. Although the bivariate random-effects model demonstrated adequate fit, the presence of heterogeneity warrants cautious interpretation of the pooled estimates. Third, the included studies evaluated a wide range of AI approaches, from individual deep learning architectures to ensemble and clinically or commercially available systems. Such methodological diversity makes direct comparison between individual AI models difficult and limits the extent to which the pooled estimates can be attributed to a specific model architecture. Moreover, differences in the availability and quality of training and test datasets, as well as differences in external versus internal validation, may have influenced reported diagnostic performance. Fourth, several studies had concerns regarding patient selection and applicability, indicating that the study populations may not fully represent the spectrum of patients encountered in routine clinical

practice. The relatively high fracture prevalence observed in some studies may also affect measures of diagnostic performance and their applicability to settings with lower disease prevalence. Finally, the evidence synthesized in this review primarily reflects diagnostic accuracy rather than clinical effectiveness. Although the pooled estimates and likelihood ratios suggest that AI can meaningfully alter the probability of fracture, the included studies do not establish whether the use of AI improves patient outcomes, reduces diagnostic errors, decreases reporting time, or changes clinical management.

5. Conclusions

AI models demonstrate strong diagnostic performance for detecting foot and ankle fractures on radiographs, with pooled sensitivity of 90%, specificity of 88%, and a SROC AUC of 0.94. Diagnostic performance differed by anatomical region, with higher sensitivity for foot fractures and higher specificity for ankle fractures. These findings support AI as a promising adjunct to radiographic interpretation, although prospective, multicenter studies with external validation are needed to confirm its clinical utility and real-world effectiveness.

6. Declarations

6.1. Acknowledgments

None.

6.2. Conflict of interests

None.

6.3. Funding and supports

None.

6.4. Authors' contributions

All authors contributed to study design, data collection, and writing the draft of the study. All authors read and approved the final version of the manuscript.

6.5. Ethical statement

Not applicable.

6.6. Informed consent

Not applicable.

6.7. Availability of supporting data

Not applicable.

6.8. Using artificial intelligence chatbots statement

We used ChatGPT to check grammar and enhance the academic quality and clarity of the text, which was primarily written by the authors. The final version of the text, after incorporating ChatGPT-assisted improvements, was thoroughly reviewed and approved by the authors.

7. References

1. Cheng CT, Hsu CP, Ooyang CH, Chou CY, Lin NY, Lin JY, et al. Evaluation of ensemble strategy on the development of multiple view ankle fracture detection algorithm. *Br J Radiol*. 2023;96(1145):20220924.
2. Lau H, Dreyer MA. Cuboid Stress Fractures. StatPearls. Treasure Island (FL): StatPearls Publishing, StatPearls Publishing LLC.;2026.

3. De Boer AS, Schepers T, Panneman MJ, Van Beeck EF, Van Lieshout EM. Health care consumption and costs due to foot and ankle injuries in the Netherlands, 1986-2010. *BMC Musculoskelet Disord*. 2014;15:128.
4. Davidovitch RI, Walsh M, Spitzer A, Egol KA. Functional outcome after operatively treated ankle fractures in the elderly. *Foot Ankle Int*. 2009;30(8):728-33.
5. Beckenkamp PR, Lin CW, Chaggar S, Herbert RD, van der Ploeg HP, Moseley AM. Prognosis of physical function following ankle fracture: a systematic review with meta-analysis. *J Orthop Sports Phys Ther*. 2014;44(11):841-51, b2.
6. Mehta SS, Rees K, Cutler L, Mangwani J. Understanding risks and complications in the management of ankle fractures. *Indian J Orthop*. 2014;48(5):445-52.
7. Brandser EA, Braksiek RJ, El-Khoury GY, Saltzman CL, Marsh JL, Clark WA, et al. Missed fractures on emergency room ankle radiographs: An analysis of 433 patients. *Emerg Radiol*. 1997;4(5):295-302.
8. Young KW, Park YU, Kim JS, Cho HK, Choo HS, Park JH. Misdiagnosis of Talar Body or Neck Fractures as Ankle Sprains in Low Energy Traumas. *Clin Orthop Surg*. 2016;8(3):303-9.
9. Judd DB, Kim DH. Foot fractures frequently misdiagnosed as ankle sprains. *Am Fam Physician*. 2002;66(5):785-94.
10. Wei CJ, Tsai WC, Tiu CM, Wu HT, Chiou HJ, Chang CY. Systematic analysis of missed extremity fractures in emergency radiology. *Acta Radiol*. 2006;47(7):710-7.
11. Gitto S, Serpi F, Albano D, Risoleo G, Fusco S, Messina C, et al. AI applications in musculoskeletal imaging: a narrative review. *Eur Radiol Exp*. 2024;8(1):22.
12. Sharma S. Artificial intelligence for fracture diagnosis in orthopedic X-rays: current developments and future potential. *SICOT J*. 2023;9:21.
13. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology*. 2022;304(1):50-62.
14. Ruitenbeek HC, Sahil S, Kumar A, Kushawaha RK, Tanamala S, Sathyamurthy S, et al. Cross-validation of an artificial intelligence tool for fracture classification and localization on conventional radiography in Dutch population. *Insights Imaging*. 2025;16(1):150.
15. Bousson V, Attané G, Benoist N, Perronne L, Diallo A, Hadid-Beurrier L, et al. Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms. *Acad Radiol*. 2023;30(10):2118-39.
16. Kim T, Goh TS, Lee JS, Lee JH, Kim H, Jung ID. Transfer learning-based ensemble convolutional neural network for accelerated diagnosis of foot fractures. *Phys Eng Sci Med*. 2023;46(1):265-77.
17. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, 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):388-96.
18. Mangwani J, Jithin SP, Akram N, Ravi L, Vaishya R. Machine learning and deep learning for ankle fracture detection in radiographs: A systematic review with narrative synthesis on diagnostic performance and future clinical integration. *Foot (Edinb)*. 2026;68:102263.
19. Prasetyo GT, Wijaya JH. The diagnostic accuracy of artificial intelligence in detecting foot and ankle fractures: a systematic review and meta-analysis of diagnostic test accuracy. *J Orthop Surg Res*. 2026.

20. Husarek J, Fuchss AMC, Ruder TD, Mougiakakou S, Exadaktylos A, Wahedi K, et al. Artificial intelligence for fracture detection on computed tomography: a comprehensive systematic review and meta-analysis of diagnostic test accuracy in non-commercial and commercial solutions. *Emerg Radiol*. 2026;33(2):351-75.
21. Nowroozi A, Salehi MA, Shobeiri P, Agahi S, Momtazmanesh S, Kaviani P, et al. Artificial intelligence diagnostic accuracy in fracture detection from plain radiographs and comparing it with clinicians: a systematic review and meta-analysis. *Clin Radiol*. 2024;79(8):579-88.
22. Guerhazi A, Tannoury C, Kompel AJ, Murakami AM, Ducarouge A, Gillibert A, et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. *Radiology*. 2022;302(3):627-36.
23. Kim J-H, Mo Y-C, Choi S-M, Hyun Y, Lee JW. Detecting Ankle Fractures in Plain Radiographs Using Deep Learning with Accurately Labeled Datasets Aided by Computed Tomography: A Retrospective Observational Study. *Appl Sci (Basel)*. 2021;11(19):8791.
24. Kitamura G, Chung CY, Moore BE, 2nd. Ankle Fracture Detection Utilizing a Convolutional Neural Network Ensemble Implemented with a Small Sample, De Novo Training, and Multiview Incorporation. *J Digit Imaging*. 2019;32(4):672-7.
25. Prijs J, Liao Z, To MS, Verjans J, Jutte PC, Stirler V, et al. Development and external validation of automated detection, classification, and localization of ankle fractures: inside the black box of a convolutional neural network (CNN). *Eur J Trauma Emerg Surg*. 2023;49(2):1057-69.
26. Taseh A, Shah A, Eftekhari M, Flaherty A, Ebrahimi A, Jones S, et al. Detecting Fifth Metatarsal Fractures on Radiographs Through the Lens of Smartphones: The FIXUS AI Algorithm. *Cureus*. 2025;17(8):e91284.
27. Wang Y, Li Y, Lin G, Zhang Q, Zhong J, Zhang Y, et al. Lower-extremity fatigue fracture detection and grading based on deep learning models of radiographs. *Eur Radiol*. 2023;33(1):555-65.

Table 1: Characteristics of the studies included in the meta-analysis

Author	Year	Age*	Gender	Anatomical Region	AI Model	Imaging Modality	Reference Standard	Sample Size	Sensitivity	Specificity
Bousson et al. (a) (15)	2023	41.3 ± 18.5	M: 61.3%; F: 38.7%	Ankle	SmartUrgence	Radiography	Final diagnosis by four musculoskeletal radiologists, based on radiographs, AI reports, available follow-up information and consensus when required	232	0.89	0.92
Bousson et al. (b) (15)	2023	41.3 ± 18.5	M: 61.3%; F: 38.7%	Ankle	Rayvolve	Radiography	Final diagnosis by four musculoskeletal radiologists, based on radiographs, AI reports, available follow-up information and consensus when required	232	0.92	0.72
Bousson et al. (c) (15)	2023	41.3 ± 18.5	M: 61.3%; F: 38.7%	Ankle	BoneView	Radiography	Final diagnosis by four musculoskeletal radiologists, based on radiographs, AI reports, available follow-up information and consensus when required	232	0.89	0.93
Bousson et al. (d) (15)	2023	41.3 ± 18.5	M: 61.3%; F: 38.7%	Foot	SmartUrgence	Radiography	Final diagnosis by four musculoskeletal radiologists, based on radiographs, AI reports, available follow-up information and consensus when required	186	0.85	0.90
Bousson et al. (e) (15)	2023	41.3 ± 18.5	M: 61.3%; F: 38.7%	Foot	Rayvolve	Radiography	Final diagnosis by four musculoskeletal radiologists, based on radiographs, AI reports, available follow-up information and consensus when required	186	0.90	0.62
Bousson et al. (f) (15)	2023	41.3 ± 18.5	M: 61.3%; F: 38.7%	Foot	BoneView	Radiography	Final diagnosis by four musculoskeletal radiologists, based on radiographs, AI reports, available follow-up information and consensus when required	186	0.98	0.79
Cheng et al. (1)	2023	36.25±17.34	M: 63%; F: 37%	Ankle	ImageNet and Xception	Radiography	Composite reference standard	100	0.84	0.9
Guermazi et al. (22)	2022	Fractures: 49±14; Non-Fractures: 60±15	Fractures: M: 33%, F: 67%; Non-Fractures: M: 27%, F: 73% NR	Foot and Ankle	Detectron2	Radiography	Independent interpretation by two experienced musculoskeletal radiologists; disagreements resolved by a third musculoskeletal radiologist	60	0.93	0.93
Kim et al. (a) (23)	2021	over 18 years	NR	Ankle	InceptionV3	AP Radiography	Simultaneous review of X-rays and CT images by three specialists	98	0.9	0.78
Kim et al. (b) (23)	2021	over 18 years	NR	Ankle	InceptionV3	Lateral Radiography	Simultaneous review of X-rays and CT images by three specialists	98	0.883	0.86
Kim et al. (c) (16)	2023	52.96 (20-89)	M: 81, F: 84	Foot	Ensemble	Radiography	Two radiologists	165	0.78	0.92
Kitamura et al. (24)	2019	NR	NR	Ankle	Ensemble_A	Radiography	Board-certified radiologist and a fourth-year radiology resident	80	0.8	0.83
Prijs et al. (a) (25)	2023	NR	NR	Ankle	Detectron2 Mask R-CNN, ResNet-50 backbone	Radiography	Consensus of 3 experienced fellowship-trained trauma surgeons	100	0.92	1

Prijs et al. (b) (25)	2023	NR	NR	Ankle	Detectron2 Mask R-CNN, ResNet-50 backbone	Radiography	Consensus of 3 experienced fellowship-trained trauma surgeons	100	0.88	1
Ruitenbeek et al. (a) (14)	2025	48 (18–98)	M: 7265, F: 7046	Ankle	Multitask deep neural network	Radiography	Final clinical radiology report, with expert MSK review/verification	2108	0.875	0.87
Ruitenbeek et al. (b) (14)	2025	48 (18–98)	M: 7265, F: 7046	Foot and toe	Multitask deep neural network	Radiography	Final clinical radiology report, with expert MSK review/verification	1214	0.80	0.80
Taseh et al. (26)	2025	Fracture: 56 (36–68); Controls: 62 (51–72)	Fracture: F: 75.5%, Control: F: 72.1%	Foot-Fifth metatarsal	ResNet 152-v2	Smartphone photographs of three-view plain foot radiographs	3 experienced orthopedic researchers using radiographs plus CT/MRI where available and clinical radiology notes	246	0.98	0.99
Wang et al. (27)	2023	Fracture: 20.2 ± 2.2; Controls 23.7 ± 4.7	Fracture: 32M/1F; Controls 26M/7F	Foot	ResNet-50	Radiography	Two radiologists + senior-radiologist consensus; clinical information and X-rays	131	0.90	0.66

*: age is presented as mean ± standard deviation or median (interquartile range). AI: artificial intelligence; M: Male; F: Female; NR: Not Reported; MSK: Musculoskeletal; CT: Computed Tomography; MRI: Magnetic Resonance Imaging; R-CNN: Region-based Convolutional Neural Networks; AP: anteroposterior.

Table 2: Quality appraisal of the included studies

Study	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Bousson et al. (a) (15)	😊	😊	😊	😊	😊	😊	😊
Bousson et al. (b) (15)	😊	😊	😊	😊	😊	😊	😊
Bousson et al. (c) (15)	😊	😊	😊	😊	😊	😊	😊
Bousson et al. (d) (15)	😊	😊	😊	😊	😊	😊	😊
Bousson et al. (e) (15)	😊	😊	😊	😊	😊	😊	😊
Bousson et al. (f) (15)	😊	😊	😊	😊	😊	😊	😊
Cheng et al. (1)	😊	😞	😊	😊	😊	😊	😊
Guermazi et al. (22)	😞	😊	😊	😊	😊	😊	😊
Kim et al. (a) (23)	😊	😊	😊	😊	😊	😊	😊
Kim et al. (b) (23)	😊	😊	😊	😊	😊	😊	😊
Kim et al. (c) (16)	😊	😊	😊	😊	😊	😊	😊
Kitamura et al. (24)	😊	😊	😊	😊	😊	😊	😊
Prijs et al. (a) (25)	😞	😊	😊	😊	😞	😊	😊
Prijs et al. (b) (25)	😞	😊	😊	😊	😞	😊	😊
Ruitenbeek et al. (a) (14)	😞	😊	😊	😊	😞	😊	😊
Ruitenbeek et al. (b) (14)	😞	😊	😊	😊	😞	😊	😊
Taseh et al. (26)	😞	😊	😊	😊	😞	😊	😊
Wang et al. (27)	😞	😊	😊	😊	😞	😊	😊

😊 Low Risk; 😞 High Risk; ? Unclear Risk.

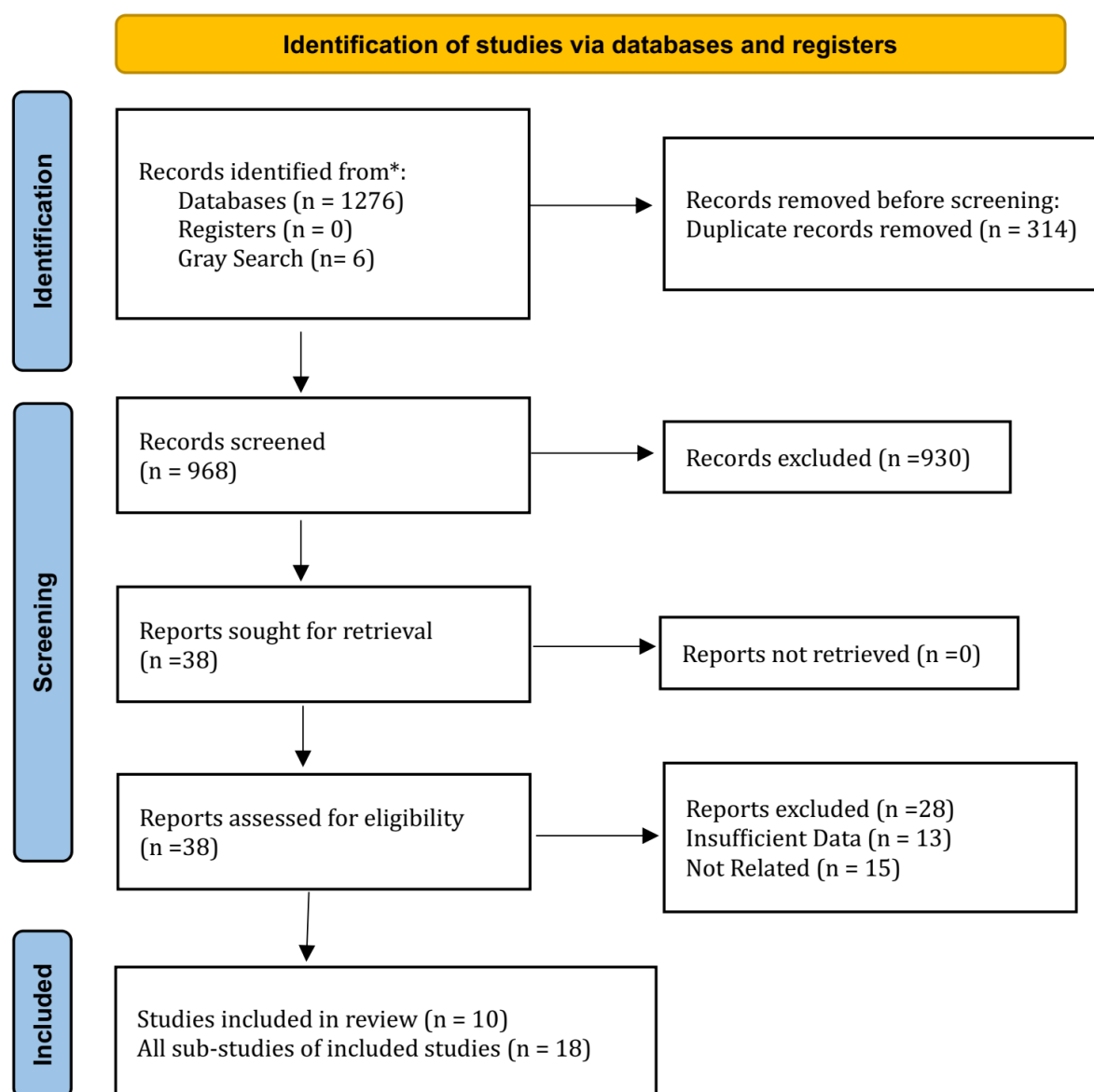


Figure 1: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart of the included studies.

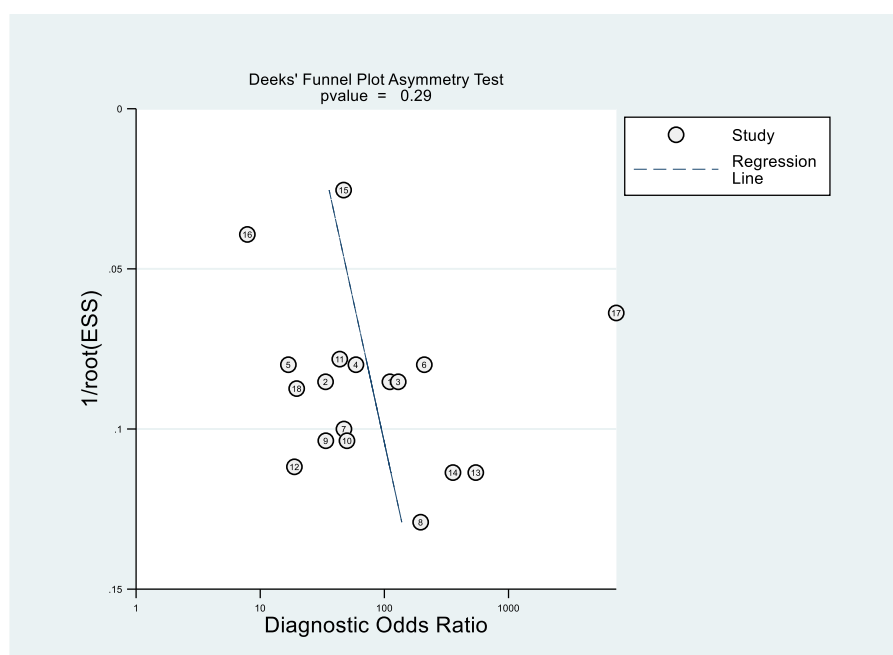


Figure 2: Deeks' funnel plot and Deeks' asymmetry test for the evaluation of publication bias.

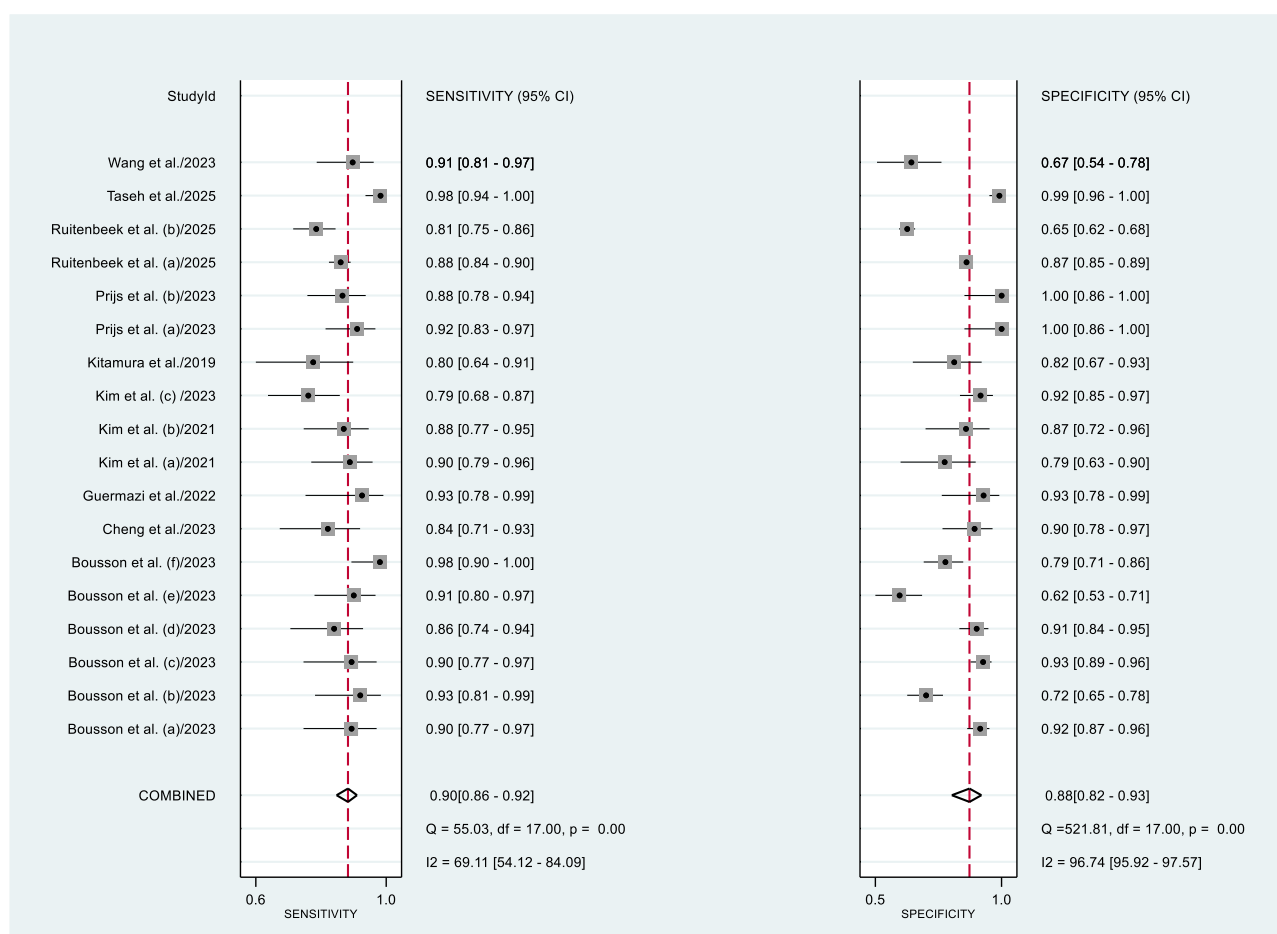


Figure 3: Pooled sensitivity and specificity of artificial intelligence (AI) models for detection of foot and ankle fractures. CI: confidence interval.

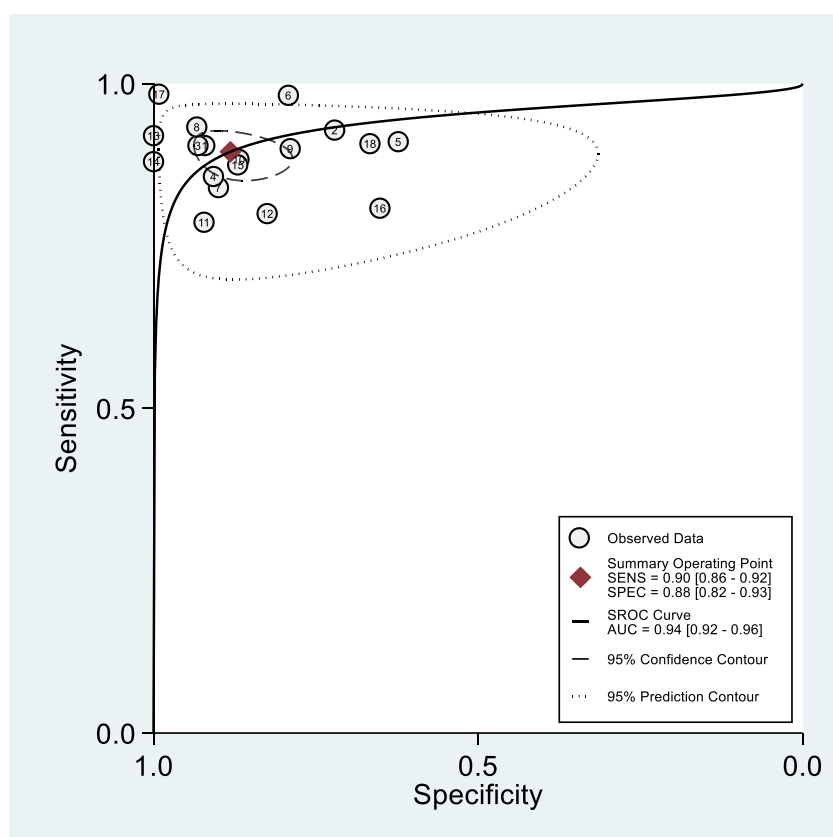


Figure 4: Area under the summary receiver operating characteristic (SROC) curve of artificial intelligence (AI) models for detection of foot and ankle fractures.

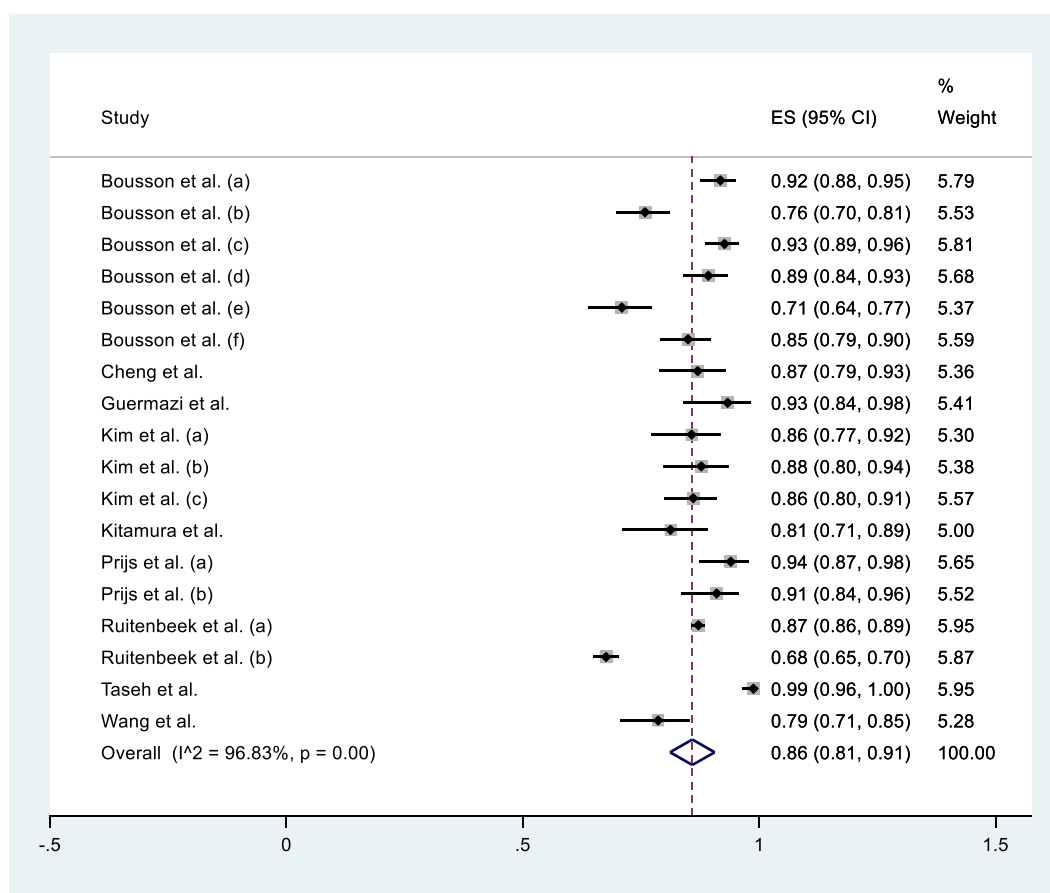


Figure 5: Pooled accuracy of artificial intelligence (AI) models for detection of foot and ankle fractures.

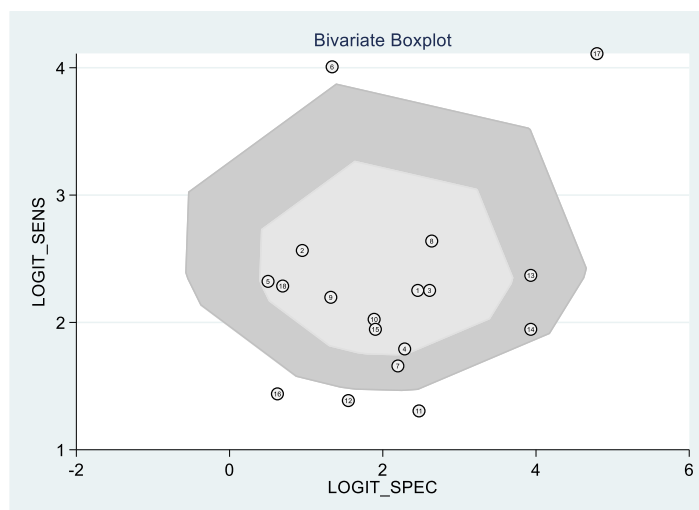


Figure 6: Boxplot for evaluation of heterogeneity and identifying outliers. SENS: sensitivity; SPEC: specificity.

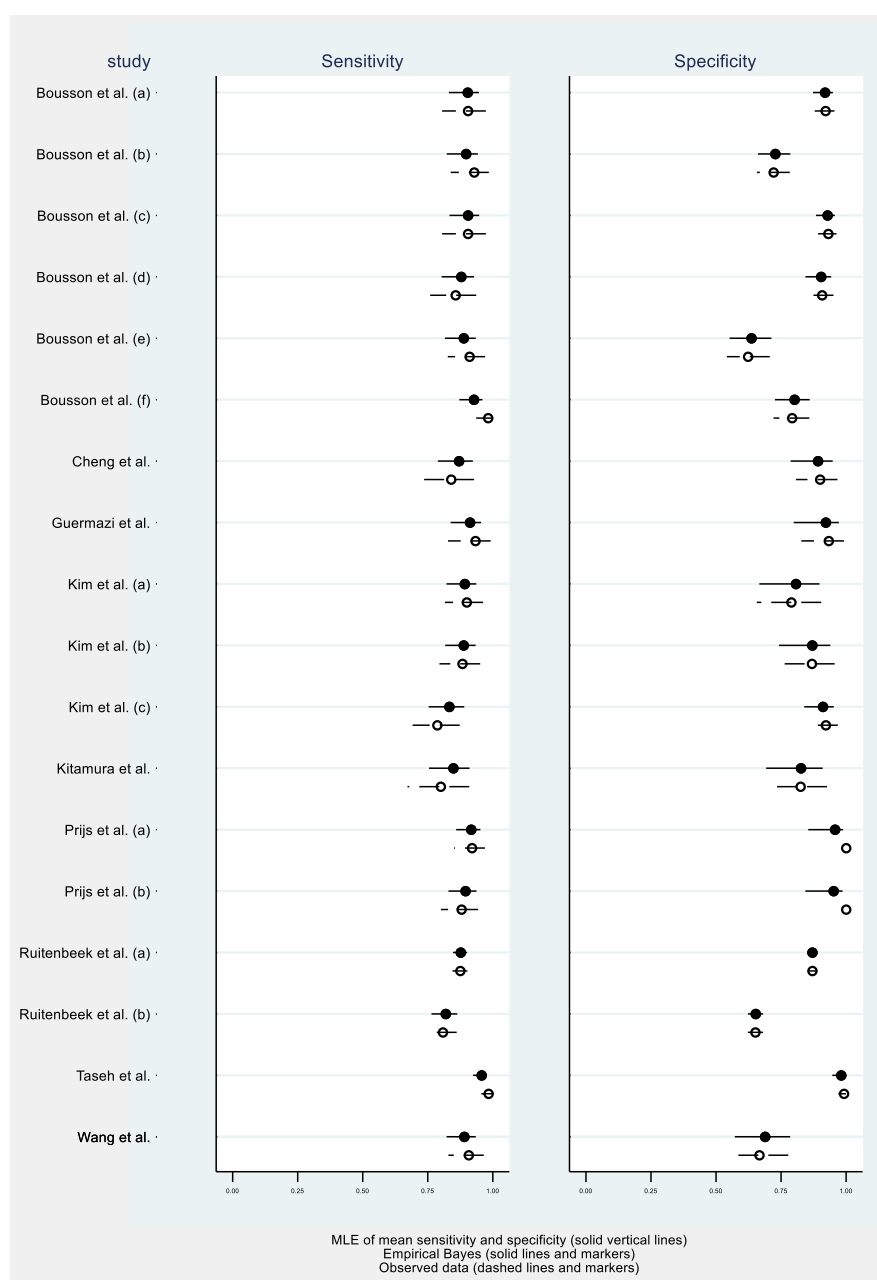


Figure 7: Forest plot of observed and empirical Bayes estimates.

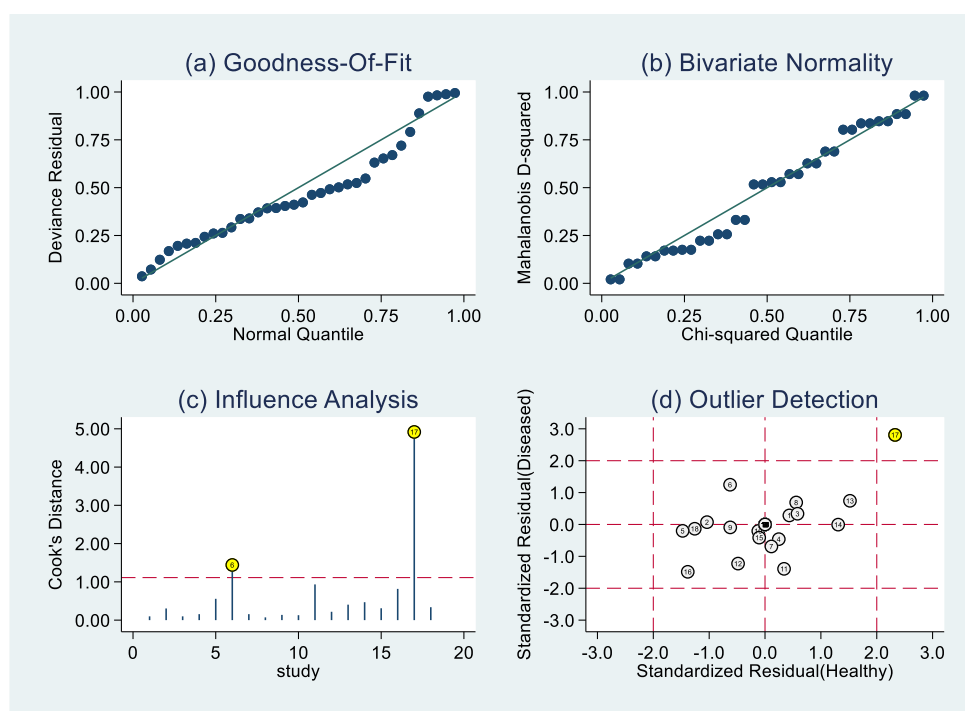


Figure 8: Goodness-of-fit and bivariate normality plots for model evaluation, with influence and outlier diagnostic plots for detecting influential and outlier studies.

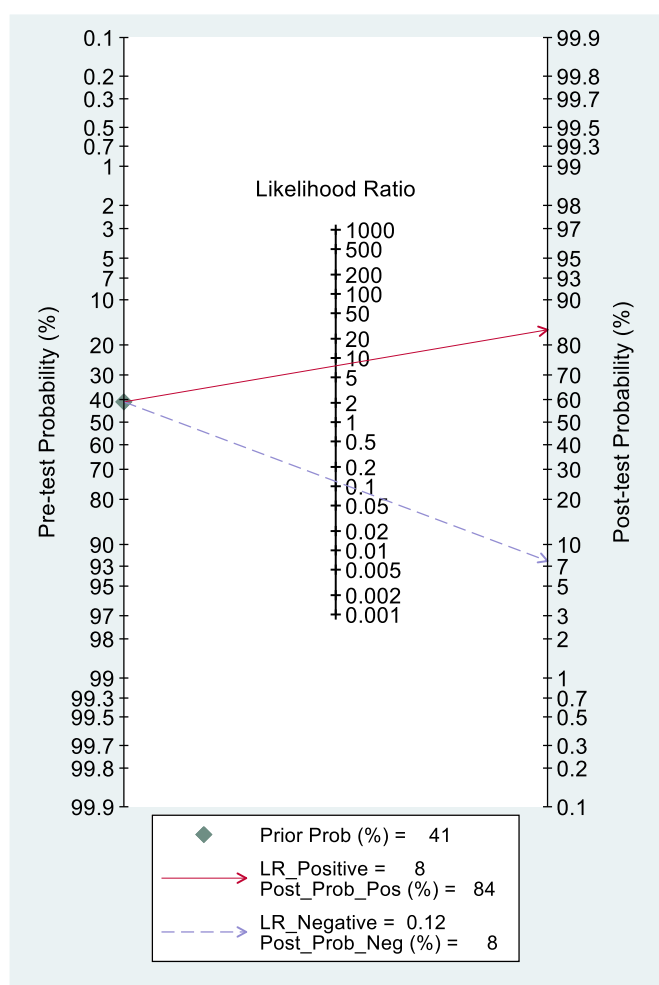


Figure 9: Fagan's nomogram illustrating the effect of the artificial intelligence (AI) models' results on the post-test probability of foot and ankle fractures. LR: likelihood ratio.

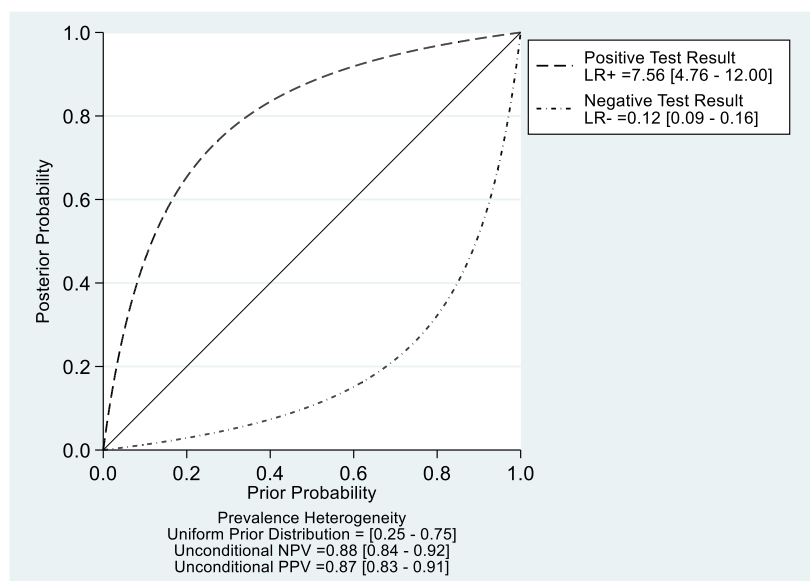


Figure 10: Probability-modifying plot for artificial intelligence (AI) models. NPV: Negative Predictive Value; PPV: Positive Predictive Value; LR+: Positive Likelihood Ratio; LR-: Negative Likelihood Ratio.

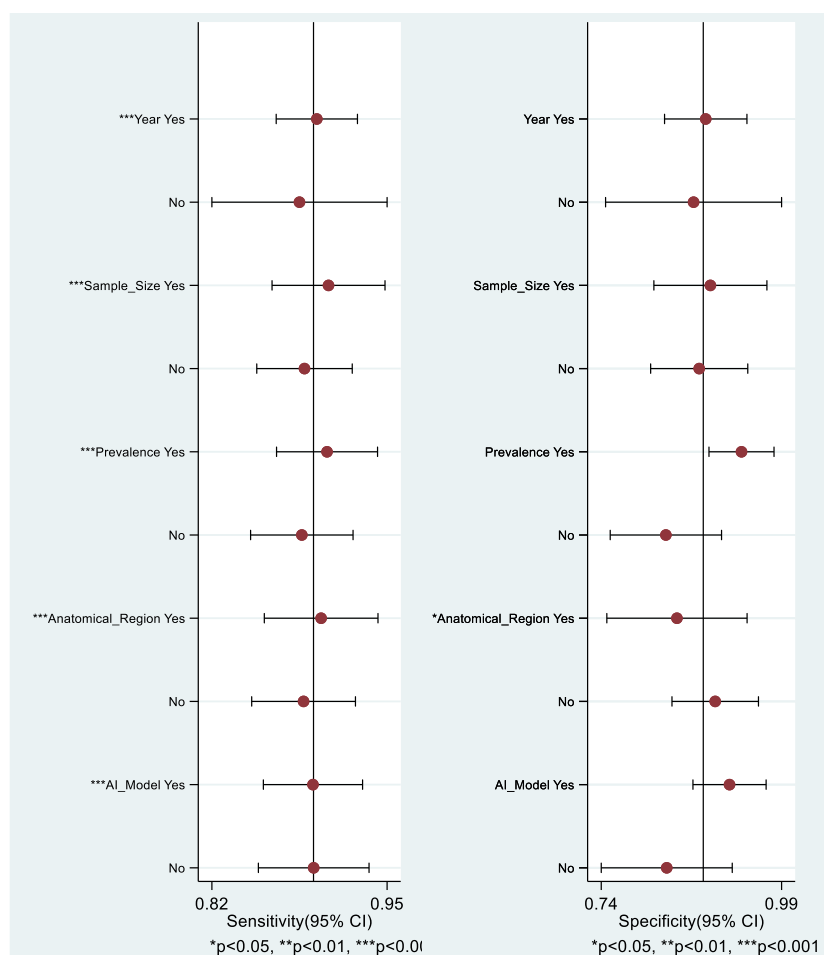


Figure 11: Forest plot of meta-regression and subgroup analyses. For the year variable, “Yes” indicates studies published in 2023 or later (14 studies vs 4 studies). For sample size, “Yes” indicates studies with a sample size of ≥ 200 participants (6 studies vs 12 studies). For prevalence, “Yes” indicates a fracture prevalence of ≥ 0.50 (8 studies vs 10 studies). For anatomical region, “Yes” indicates studies involving the foot, whereas “No” indicates studies involving the ankle (7 studies vs 10 studies). For artificial intelligence (AI) model type, “Yes” indicates a deep learning model, whereas “No” indicates a clinical or commercially available AI model (10 studies vs 8 studies). CI: confidence interval.

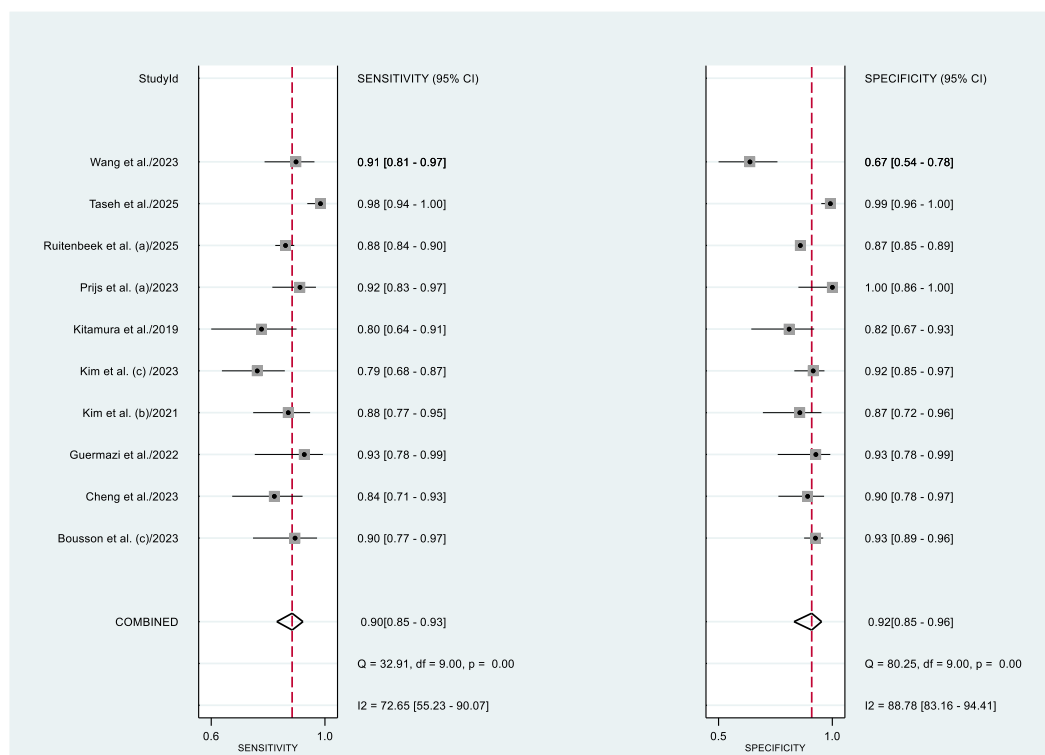


Figure 12: Sensitivity analysis of studies using one estimate per cohort, selecting the estimate with the highest area under the curve (AUC) among the available estimates.

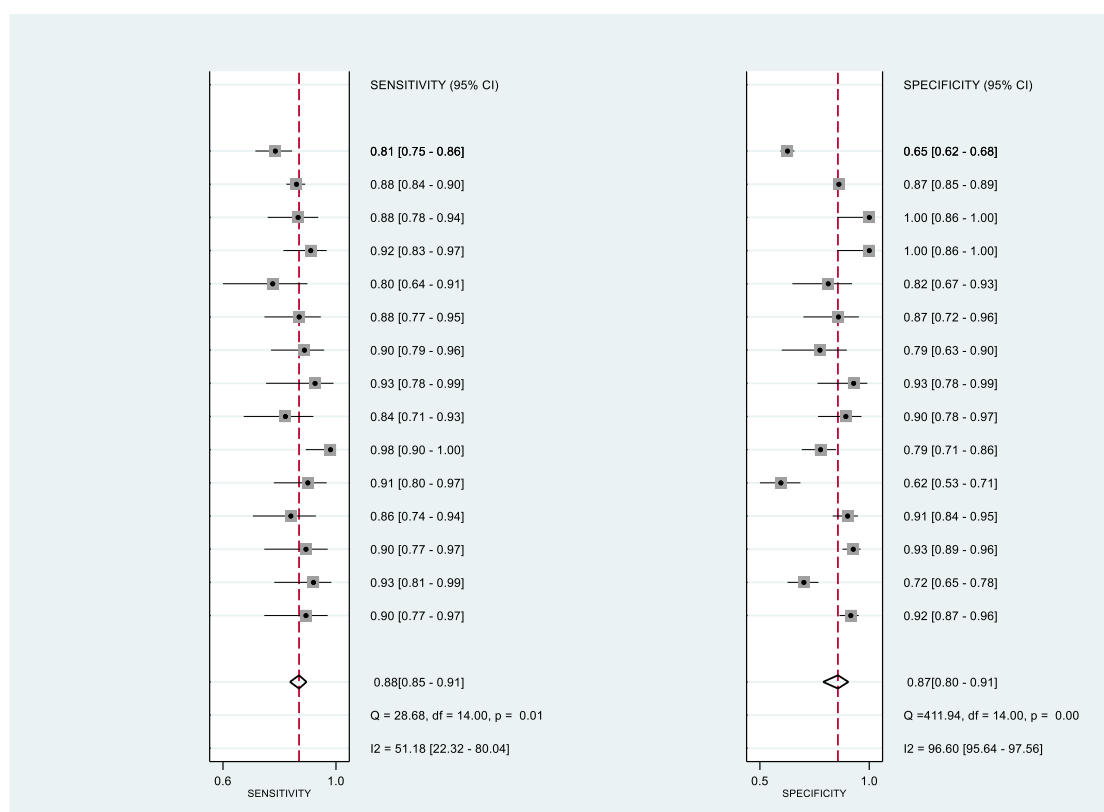


Figure 13: Sensitivity analysis excluding three potentially heterogeneous studies: Wang et al. (fatigue fractures), Taseh et al. (fifth metatarsal fractures assessed using smartphone photographs of radiographs), and Kim et al. (single-view radiographs).

Appendix

Supplementary Table 1: Search strategy in different databases

PubMed: 2026-09-07: 568
#1 "Ankle Fractures"[MeSH Terms] OR "Ankle Injuries"[MeSH Terms] OR "Foot Injuries"[MeSH Terms] OR "ankle fracture*"[Title/Abstract] OR "ankle injur*"[Title/Abstract] OR "ankle sprain*"[Title/Abstract] OR "ankle trauma"[Title/Abstract] OR "malleolar fracture*"[Title/Abstract] OR "syndesmoti* fracture*"[Title/Abstract] OR "lateral malleolus fracture*"[Title/Abstract] OR "medial malleolus fracture*"[Title/Abstract] OR "bimalleolar fracture*"[Title/Abstract] OR "posterior malleolus fracture*"[Title/Abstract] OR "trimalleolar fracture*"[Title/Abstract] OR "bimalleolar equivalent fracture*"[Title/Abstract] OR "foot fracture*"[Title/Abstract] OR "foot injur*"[Title/Abstract] OR "foot sprain*"[Title/Abstract] OR "foot trauma"[Title/Abstract] OR "fracture* of the foot"[Title/Abstract] OR "foot bone fracture*"[Title/Abstract] OR "fractured foot"[Title/Abstract] OR "tars* fracture*"[Title/Abstract] OR "metatars* fracture*"[Title/Abstract] OR "toe fracture*"[Title/Abstract] OR "phalan* fracture*"[Title/Abstract] OR "calcane* fracture*"[Title/Abstract] OR "talus fracture*"[Title/Abstract] OR "talar fracture*"[Title/Abstract] OR "navicular fracture*"[Title/Abstract] OR "cuboid fracture*"[Title/Abstract] OR "cuneiform fracture*"[Title/Abstract] OR "sesamoid fracture*"[Title/Abstract] #2 "Artificial Intelligence"[MeSH Terms] OR "Machine Learning"[MeSH Terms] OR "Deep Learning"[MeSH Terms] OR "Convolutional Neural Networks"[MeSH Terms] OR "Recurrent Neural Networks"[MeSH Terms] OR "Image Processing, Computer-Assisted"[MeSH Terms] OR "Diagnosis, Computer-Assisted"[MeSH Terms] OR "Autoencoder"[MeSH Terms] OR "Graph Neural Networks"[MeSH Terms] OR "Long Short Term Memory"[MeSH Terms] OR "Ensemble Learning"[MeSH Terms] OR "Random Forest"[MeSH Terms] OR "Support Vector Machine"[MeSH Terms] OR "Decision Trees"[MeSH Terms] OR "Fuzzy Logic"[MeSH Terms] OR "Boosting Machine Learning Algorithms"[MeSH Terms] OR "Transfer Machine Learning"[MeSH Terms] OR "artificial intelligen*"[Title/Abstract] OR "machine learn*"[Title/Abstract] OR "deep learn*"[Title/Abstract] OR "neural network*"[Title/Abstract] OR "computer vision"[Title/Abstract] OR "image process*"[Title/Abstract] OR "image recognition"[Title/Abstract] OR "computer-aided detection"[Title/Abstract] OR "computer assisted detection"[Title/Abstract] OR "computer-aided diagnosis"[Title/Abstract] OR "computer assisted diagnosis"[Title/Abstract] OR "classifier*"[Title/Abstract] OR "autoencoder*"[Title/Abstract] OR "long short term memory"[Title/Abstract] OR "ensemble learning"[Title/Abstract] OR "random forest"[Title/Abstract] OR "random-forest"[Title/Abstract] OR "support vector machine"[Title/Abstract] OR "decision tree"[Title/Abstract] OR "fuzzy logic"[Title/Abstract] OR "eXtreme gradient boosting"[Title/Abstract] OR "transfer learn*"[Title/Abstract] OR "attention mechanism"[Title/Abstract] OR "self-attention"[Title/Abstract] OR "residual network*"[Title/Abstract] OR "fully convolutional network*"[Title/Abstract] OR "generative adversarial network*"[Title/Abstract] OR "vision transformer*"[Title/Abstract] OR "transformer network*"[Title/Abstract] OR "U-Net"[Title/Abstract] OR "UNet"[Title/Abstract] OR "ResNet"[Title/Abstract] OR "DenseNet"[Title/Abstract] OR "EfficientNet"[Title/Abstract] OR "Inception"[Title/Abstract] OR "YOLO"[Title/Abstract] OR "Mask R-CNN"[Title/Abstract] OR "Faster R-CNN"[Title/Abstract] OR "multilayer perceptron*"[Title/Abstract] OR "representation learning"[Title/Abstract] OR "end-to-end learning"[Title/Abstract] OR "self-supervise*"[Title/Abstract] OR "semi-supervised learning"[Title/Abstract] OR "weakly supervised learning"[Title/Abstract] #3 "diagnosis"[MeSH Terms] OR "Early Diagnosis"[MeSH Terms] OR "Sensitivity and Specificity"[MeSH Terms] OR "diagnos*"[Title/Abstract] OR "Sensitivity"[Title/Abstract] OR "Specificity"[Title/Abstract] OR "Accuracy"[Title/Abstract] OR "identif*"[Title/Abstract] OR "detect*"[Title/Abstract] #4 #1 AND #2 AND #3 Scopus: 2026-09-07: 511 Documents #1 TITLE-ABS-KEY ("Ankle Fracture*" OR "ankle injur*" OR "Ankle Sprain*" OR "ankle trauma" OR "malleolar fracture*" OR "Syndesmoti* Injur*" OR "Lateral Malleolus Fracture*" OR "Medial Malleolus Fracture*" OR "Bimalleolar Fracture*" OR "Posterior Malleolus Fracture*" OR "Trimalleolar Fracture*" OR "Bimalleolar Equivalent Fracture*" OR "foot Fracture*" OR "foot injur*" OR "foot Sprain*" OR "foot trauma" OR "fracture* of the foot" OR "foot bone fracture*" OR "fractured foot" OR "tars* fracture*" OR "metatars* fracture*" OR "toe fracture*" OR "phalan* fracture*" OR "calcane* fracture*" OR "talus fracture*" OR "talar fracture*" OR "navicular fracture*" OR "cuboid fracture*" OR "cuneiform fracture*" OR "sesamoid fracture*") #2 TITLE-ABS-KEY ("artificial intellig*" OR "machine learn*" OR "deep learn*" OR "neural network*" OR "computer vision" OR "image process*" OR "image recognition" OR "computer-aided detection" OR "computer assisted detection" OR "computer-aided diagnosis" OR "computer assisted diagnosis" OR "classifier*" OR "autoencoder*" OR "long short term memory" OR "ensemble learning" OR "random forest" OR "random-forest" OR "support vector machine" OR "decision tree" OR "fuzzy logic" OR "eXtreme gradient boosting" OR "transfer learn*" OR "attention mechanism" OR "self-attention" OR "residual network*" OR "fully convolutional network*" OR "generative adversarial network*" OR "vision transformer*" OR "transformer network*" OR "U-Net" OR "UNet" OR "ResNet" OR "DenseNet" OR "EfficientNet" OR "Inception" OR "YOLO" OR "Mask R-CNN" OR "Faster R-CNN" OR "multilayer perceptron*" OR "representation learning" OR "end-to-end learning" OR "self-supervise*" OR "semi-supervised learning" OR "weakly supervised learning") #3 TITLE-ABS-KEY ("diagnos*" OR "Sensitivity" OR "Specificity" OR "Accuracy" OR "identif*" OR "detect*") #4 #1 AND #2 AND #3 Web of Science: 2026-09-07: 197 #1 TS=("Ankle Fracture*" OR "ankle injur*" OR "Ankle Sprain*" OR "ankle trauma" OR "malleolar fracture*" OR "Syndesmoti* Injur*" OR "Lateral Malleolus Fracture*" OR "Medial Malleolus Fracture*" OR "Bimalleolar Fracture*" OR "Posterior Malleolus Fracture*" OR "Trimalleolar Fracture*" OR "Bimalleolar Equivalent Fracture*" OR "foot Fracture*" OR "foot injur*" OR "foot Sprain*" OR "foot trauma" OR "fracture* of the foot" OR "foot bone fracture*" OR "fractured foot" OR "tars* fracture*" OR "metatars* fracture*" OR "toe fracture*" OR "phalan* fracture*" OR "calcane* fracture*" OR "talus fracture*" OR "talar fracture*" OR "navicular fracture*" OR "cuboid fracture*" OR "cuneiform fracture*" OR "sesamoid fracture*")

#2

TS=("artificial intellig*" OR "machine learn*" OR "deep learn*" OR "neural network*" OR "computer vision" OR "image process*" OR "image recognition" OR "computer-aided detection" OR "computer assisted detection" OR "computer-aided diagnosis" OR "computer assisted diagnosis" OR "classifier*" OR "autoencoder*" OR "long short term memory" OR "ensemble learning" OR "random forest" OR "random-forest" OR "support vector machine" OR "decision tree" OR "fuzzy logic" OR "eXtreme gradient boosting" OR "transfer learn*" OR "attention mechanism" OR "self-attention" OR "residual network*" OR "fully convolutional network*" OR "generative adversarial network*" OR "vision transformer*" OR "transformer network*" OR "U-Net" OR "UNet" OR "ResNet" OR "DenseNet" OR "EfficientNet" OR "Inception" OR "YOLO" OR "Mask R-CNN" OR "Faster R-CNN" OR "multilayer perceptron*" OR "representation learning" OR "end-to-end learning" OR "self-supervise*" OR "semi-supervised learning" OR "weakly supervised learning")

#3

TS=("diagnos*" OR "Sensitivity" OR "Specificity" OR "Accuracy" OR "identif*" OR "detect*")

#4

#1 AND #2 AND #3