DOI: 10.35191/medsoft_2026_1_38_Kvak

STUDY-LEVEL DIAGNOSTIC PERFORMANCE OF ARTIFICIAL INTELLIGENCE FOR FRACTURE DETECTION ON MUSCULOSKELETAL RADIOGRAPHS: A RETROSPECTIVE MULTI-READER STUDY

Daniel Kvak, Marek Biroš, Jindřiška Judasová

Abstract

Fracture detection on musculoskeletal radiographs remains challenging, particularly when findings are subtle or visible on only one projection. This retrospective two-centre multi-reader diagnostic performance study evaluated an artificial intelligence system at the level of the complete radiographic examination and compared its performance with seven anonymized readers. Radiographic examinations from AGEL Šternberk Hospital and AGEL Vítkovice Hospital, acquired between 15 December 2025 and 25 January 2026, were retrospectively analysed. The primary endpoint was study-level fracture positivity, defined as at least one positive image within an examination; the same aggregation rule was applied to AI and reader assessments. Examinations of the spine, ribs, and hemithorax were excluded because these anatomies were outside the supported scope. The cohort comprised 526 examinations, including 101 fracture-positive and 425 fracture-negative studies. Sensitivity was 0.911 (95% CI, 0.839–0.952), specificity 0.899 (0.866–0.924), PPV 0.681 (0.599–0.754), NPV 0.977 (0.957–0.988), and accuracy 0.901 (0.873–0.924). Mean reader sensitivity, specificity, PPV, NPV, and accuracy were 0.853, 0.853, 0.598, 0.961, and 0.853, respectively. The AI showed higher point estimates than the mean of the seven readers across all principal diagnostic metrics. These findings support prospective evaluation of AI as a second-reader or workflow-support tool for fracture detection, particularly in examinations containing subtle or easily overlooked findings.

Keywords

fracture detection, musculoskeletal radiography, artificial intelligence, diagnostic accuracy, multi-reader study

1 Introduction

Conventional radiography is widely used for fracture diagnosis, but interpretation remains challenging when fractures are nondisplaced, findings are subtle, or the abnormality is visible on only one projection. Overlapping structures and anatomical variants can further complicate assessment. These difficulties are clinically relevant because fracture recognition guides subsequent management, while missed findings can delay appropriate care. Artificial intelligence (AI) offers an additional means of identifying suspicious findings on radiographs and has therefore attracted increasing interest as diagnostic support [1, 2].

The evidence base suggests that AI can achieve diagnostic performance comparable to that of clinicians. In a systematic review and meta-analysis, Kuo et al. reported pooled sensitivity and specificity of 0.910 for AI in external validation datasets, with no statistically significant difference from clinician performance [1]. A later meta-analysis by Husarek et al. examined 17 studies of commercially available fracture-detection systems and reported pooled standalone AI sensitivity of 0.910 and specificity of 0.890 [2]. Performance varied across systems, anatomical regions, and study methods, emphasizing the value of evaluation in clearly defined clinical cohorts.

The unit of analysis is particularly important for interpreting fracture-detection results. A radiographic examination may contain multiple projections with different visibility of the same injury. Examination-level assessment addresses whether any image within the complete study is classified as fracture-positive. Applying the same aggregation rule to AI and readers enables a consistent comparison at this level. Previous multicentre research has demonstrated improved reader sensitivity and specificity with AI assistance [3].

The primary objective of this study was therefore to quantify examination-level AI sensitivity, specificity, predictive values, and accuracy against the final reference labels and to compare these estimates with the individual performance of seven anonymized readers, the arithmetic mean of their performance estimates, and a majority-vote assessment.

2 Methods

2.1 Study design and eligible examinations

This retrospective multi-reader diagnostic performance study analysed musculoskeletal radiographic examinations acquired at AGEL Šternberk Hospital and AGEL Vítkovice Hospital between 15 December 2025 and 25 January 2026. The unit of analysis was the complete examination. Examinations of the spine, ribs, or hemithorax were excluded as these anatomies were outside the scope of the evaluated AI software. The cohort comprised selected radiographic examinations representing either the initial diagnosis of a fracture or confirmed fracture-negative findings according to the original hospital radiology reports. Follow-up examinations of previously diagnosed fractures were excluded. The final cohort contained 526 examinations, including 101 fracture-positive and 425 fracture-negative examinations (prevalence, 19.2%). AGEL Šternberk contributed 309 examinations and AGEL Vítkovice contributed 217. There was a total of 1,258 images evaluable by AI and radiologists in multi-reader study.

2.2 AI and reader assessment

The index test was an AI system for fracture detection on radiographs of the appendicular skeleton and pelvis (Carebot AI Bones 1.12; Carebot s. r. o.). The system processes DICOM images using a deep-learning object detector. The AI model was developed using a multicentre dataset of 125,070 anonymized musculoskeletal radiographs from European hospitals. The fracture-training data included 37,563 radiographs containing one or more fractures, annotated with 42,313 bounding boxes. Reference annotations were created by physicians with varying levels of radiology experience. For this analysis, binary AI fracture outputs were grouped; an examination was AI-positive if at least one AI-evaluable image was positive. The same image-to-examination rule was applied to each of seven anonymized readers. Reader-majority assessment was positive when at least four of the seven examination-level reader assessments were positive.

2.3 Reference standard

The reference standard was the original radiology report issued by the hospital radiology team as part of routine clinical care. The radiologists had access to clinical information, follow-up findings, and other imaging studies where available. Fracture presence or absence was determined from the findings documented in these reports.

DOI: 10.35191/medsoft_2026_1_38_Kvak



Original report (Czech)

RTG pravého zápěstí: Vícefragmentová intraartikulární fraktura dist. meta-epifýzy radia s osovou disl. dorzálně. Abrupce proc. styl. ulnae.

English excerpt

Right wrist radiograph: Comminuted intra-articular fracture of the distal radius with dorsal angulation. Avulsion fracture of the ulnar styloid process.

Reference standard: fracture present.

Figure 1 – right wrist radiographs (A–C) with the original Czech hospital radiology report and its English translation. The report describes a comminuted intra-articular distal radius fracture and an ulnar styloid avulsion fracture, establishing the examination as fracture-positive according to the reference standard.

2.4 Analysis

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated against the final reference labels for the AI, each reader, and the reader-majority assessment. Mean reader performance was the arithmetic mean of the seven individually calculated metrics. Two-sided 95% confidence intervals for the AI and reader-majority proportions were calculated using the Wilson score method. Comparisons between AI and readers were descriptive; no formal paired tests of their differences were performed.

3 Results

3.1 Diagnostic performance

AI identified 92 true positives (TP) and 382 true negatives (TN), with 43 false positives (FP) and 9 false negatives (FN). Its sensitivity was 0.911 (95% CI 0.839–0.952), specificity 0.899 (0.866–0.924), PPV 0.681 (0.599–0.754), NPV 0.977 (0.957–0.988), and accuracy 0.901 (0.873–0.924). The F1 score was 0.780, balanced accuracy was 0.905, and MCC was 0.730.

The AI point estimates were higher than the mean of the seven individual readers for all five principal metrics. The reader majority had higher specificity, PPV, and accuracy than the AI, whereas the AI had higher sensitivity and NPV (Table 1). The majority-vote confusion counts were 89 true positives, 396 true negatives, 29 false positives, and 12 false negatives.

Assessment	Sensitivity	Specificity	PPV	NPV	Accuracy
AI	0.911	0.899	0.681	0.977	0.901
Mean of seven readers	0.853	0.853	0.598	0.961	0.853
Reader majority	0.881	0.932	0.754	0.971	0.922

Table 1 – study-level diagnostic performance against the final reference labels.

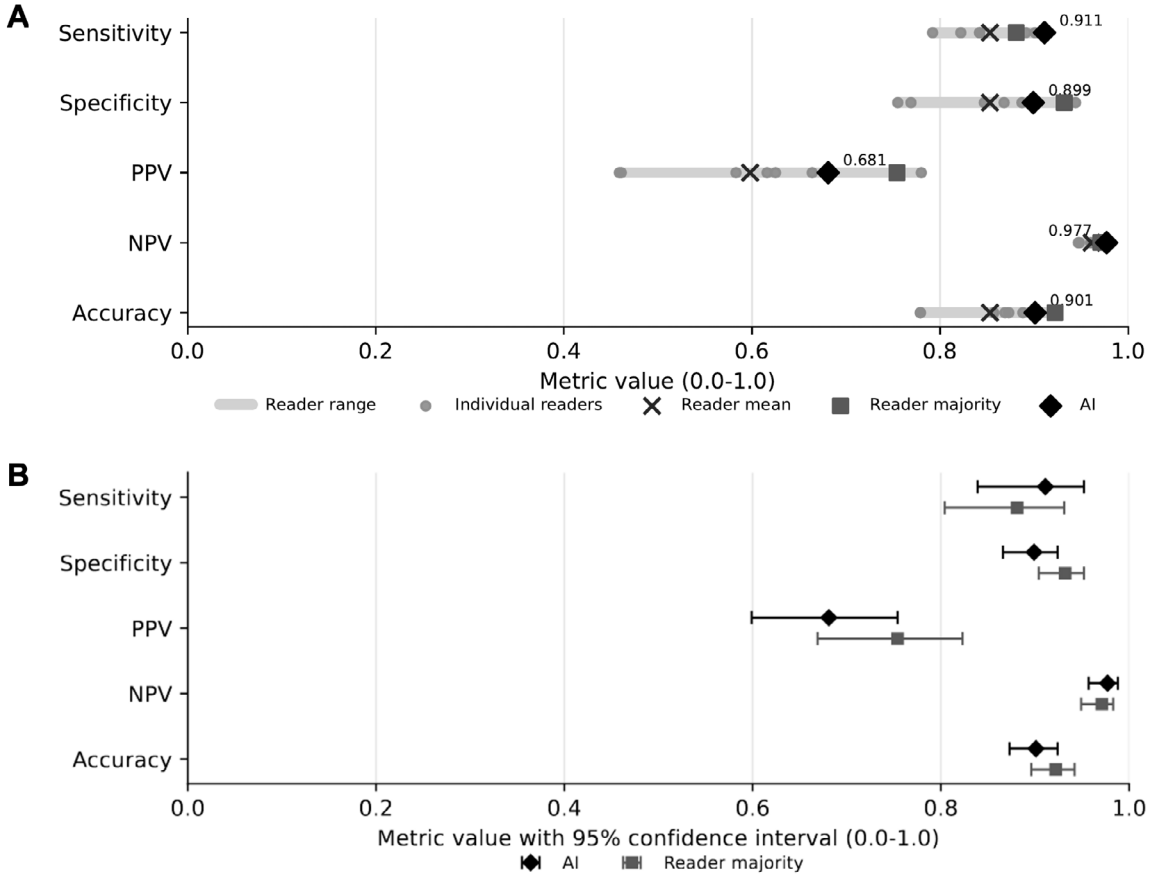


Figure 2 – study-level diagnostic metrics. (A) AI, seven individual readers, their arithmetic mean, and reader-majority vote; grey spans show the individual-reader range. (B) AI and reader-majority point estimates with Wilson 95% confidence intervals. These marginal intervals do not test paired AI–reader differences.

3.2 Individual reader performance

Individual reader sensitivity ranged from 0.792 to 0.901, and specificity from 0.755 to 0.944 (Table 2). The AI sensitivity point estimate of 0.911 exceeded that of every individual reader, whereas RAD 1 had higher specificity, PPV, and accuracy than the AI. These comparisons are descriptive; reader identities remain anonymized.

Reader	Sensitivity	Specificity	PPV	NPV	Accuracy
RAD 1	0.842	0.944	0.780	0.962	0.924
RAD 2	0.822	0.769	0.459	0.948	0.779
RAD 3	0.901	0.847	0.583	0.973	0.857
RAD 4	0.842	0.899	0.664	0.960	0.888
RAD 5	0.881	0.755	0.461	0.964	0.779
RAD 6	0.891	0.868	0.616	0.971	0.873
RAD 7	0.792	0.887	0.625	0.947	0.869

Table 2 – individual study-level reader performance.

3.3 Centre-stratified performance

At AGEL Šternberk, 69 of 309 examinations were fracture-positive according to the reference standard. AI identified 62 TP, 218 TN, 22 FP, and 7 FN, with sensitivity of 0.899 and specificity of 0.908. At AGEL Vítkovice, 32 of 217 examinations were fracture-positive. AI identified 30 TP, 164 TN, 21 FP, and 2 FN, with sensitivity of 0.938 and specificity of 0.886.

Centre	N (positive)	TP/TN/FP/FN	Sensitivity	Specificity	Accuracy
AGEL Šternberk	309 (69)	62/218/22/7	0.899	0.908	0.906
AGEL Vítkovice	217 (32)	30/164/21/2	0.938	0.886	0.894

Table 3 – exploratory centre-stratified AI performance in the final cohort.

3.4 Illustrative examinations

3.4.1 True-positive examples

Case	Radiographic finding	Reference standard	AI result	Positive readers
TP-01	Bimalleolar ankle fracture with tibiotalar subluxation	Positive	Positive	7/7
TP-02	Nondisplaced intra-articular distal radius fracture	Positive	Positive	6/7
TP-03	Left thumb tuft fracture	Positive	Positive	3/7

Table 4 – true-positive examples: radiographic findings, reference standard, AI results, and reader assessments.

3.4.2 True-negative examples

Case	Radiographic finding	Reference standard	AI result	Positive readers
TN-01	Intact pelvic ring without fracture; bilateral mild to moderate coxarthrosis	Negative	Negative	0/7
TN-02	Degenerative hand changes without acute traumatic findings	Negative	Negative	0/7
TN-03	Gonarthrosis and osteophytes without fracture	Negative	Negative	0/7

Table 5 – true-negative examples: radiographic findings, reference standard, AI results, and reader assessments.

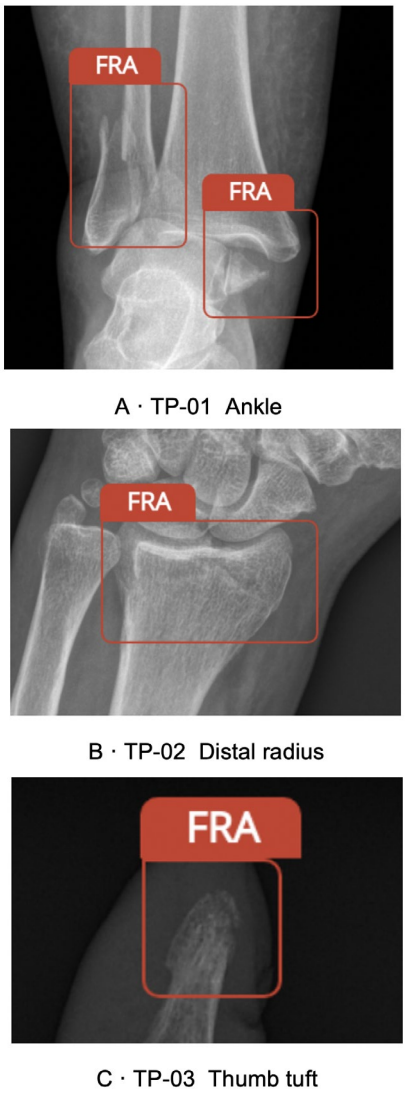


Figure 3 – AI true-positive examinations: (A) bimalleolar ankle fracture, (B) distal-radius fracture, and (C) thumb tuft fracture. Red FRA boxes show AI detections.



Figure 4 – AI true-negative examinations: (A) pelvis, (B) hand, and (C) knee.

DOI: 10.35191/medsoft_2026_1_38_Kvak

3.4.3 False-positive examples

Case	Radiographic finding	Reference standard	AI result	Positive readers
FP-01	Ossification centre with sclerotic margins at the fifth-metatarsal base; no identifiable acute fracture line	Negative	Positive	1/7
FP-02	Olecranon spur and bone proliferation with soft-tissue changes or bursitis; no visible fracture line or separated fragment	Negative	Positive	3/7
FP-03	Os trigonum, os subfibulare, and marginal ankle osteophytes without clear acute traumatic changes	Negative	Positive	2/7

Table 6 – false-positive examples: radiographic findings, reference standard, AI results, and reader assessments.

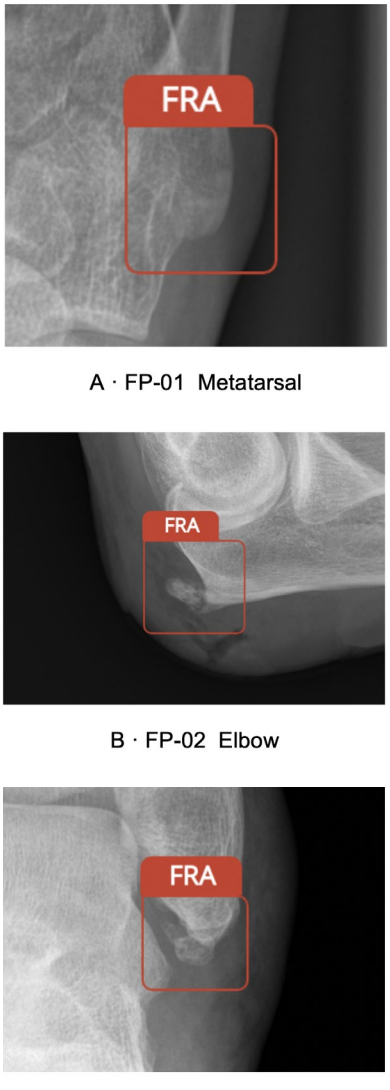


Figure 5 – AI false-positive examples: (A) fifth-metatarsal base, (B) olecranon region, and (C) ankle. All three examinations were reference-negative and AI-positive.

3.4.4 False-negative examples

Case	Radiographic finding	Reference standard	AI result	Positive readers
FN-01	Displaced periprosthetic proximal femoral fracture	Positive	Negative	5/7
FN-02	Nondisplaced subperiosteal distal radius fracture	Positive	Negative	3/7
FN-03	Nondisplaced avulsion involving the posterior margin of the proximal tibia and an osteophyte, with advanced gonarthrosis	Positive	Negative	5/7

Table 7 – false-negative examples: radiographic findings, reference standard, AI results, and reader assessments.

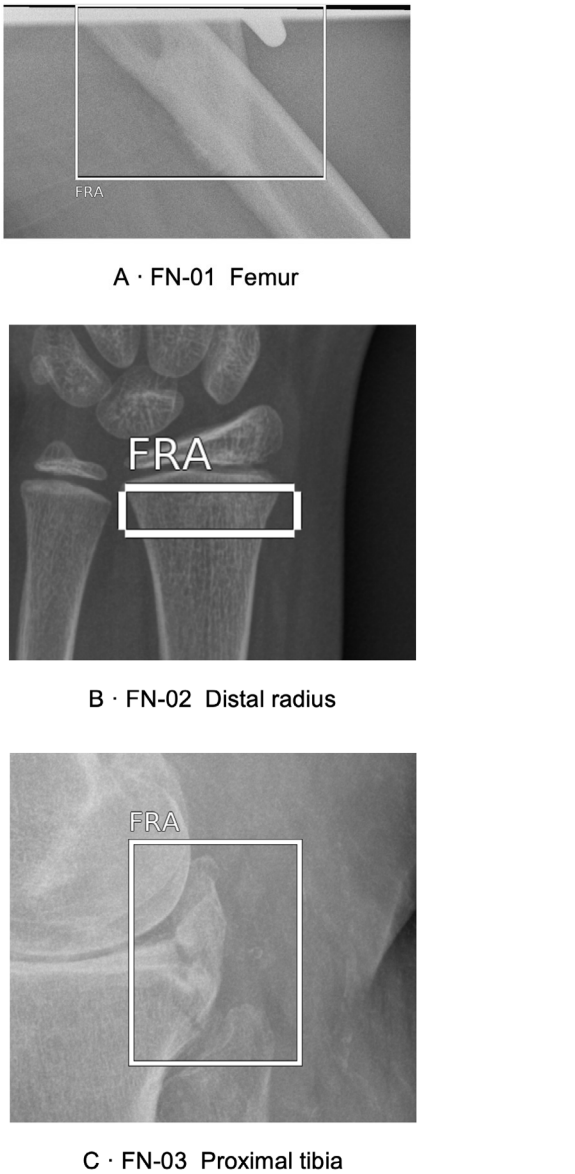


Figure 6 – AI false-negative examples: (A) periprosthetic femoral fracture, (B) subperiosteal distal-radius fracture, and (C) posterior proximal-tibial avulsion. White boxes mark the reported fracture regions. All three examinations were reference-positive and AI-negative.

DOI: 10.35191/medsoft_2026_1_38_Kvak

4 Discussion

AI achieved a sensitivity of 0.911 and an NPV of 0.977 for fracture detection at the examination level. These results indicate that AI detected most fracture-positive examinations and that most examinations classified as negative by AI were also negative according to the reference standard. Assessment across complete examinations accounted for differences in fracture visibility between projections, although it did not evaluate fracture localization or the detection of every fracture within an examination. AI point estimates exceeded the arithmetic mean of the seven individual readers across all five principal diagnostic metrics. Reader-majority voting achieved higher specificity, PPV, and accuracy. This pattern highlights the importance of considering both fracture detection and false-positive findings when comparing diagnostic performance. The results demonstrate strong standalone AI performance in this cohort, while the descriptive reader comparisons do not establish statistical superiority. The false-negative and false-positive examples provide clinical context for these findings. The missed periprosthetic femoral fracture illustrates that a negative AI result can occur despite a clinically relevant injury. Conversely, false positive examinations included cases whose reports described accessory ossicles, an ossification centre, or degenerative changes.

The next step is to determine whether AI assistance improves reader performance in practice, particularly by reducing missed fractures while limiting additional false-positive findings. Prospective assisted-reading studies are needed to assess this benefit and its effect on reporting workflow.

5 Conclusion

AI demonstrated strong performance in fracture detection, with results exceeding the average performance of individual readers in this cohort. These findings support its potential role as an aid to radiologists. Prospective evaluation is needed to determine whether AI assistance improves diagnostic performance and clinical workflow.

Literature

[1.] Kuo RYL, Harrison C, Curran TA, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology. 2022;304:50–62. doi:10.1148/radiol.211785.

[2.] 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;14:23053. doi:10.1038/s41598-024-73058-8.

[3.] 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;300:120–129. doi:10.1148/radiol.2021203886.

Kontakt

Mgr. Daniel Kvak, Ph.D.
Carebot s.r.o.
Libušina 49/3, 128 00
Praha 2 – Vyšehrad
+420 739 174 316
daniel.kvak@carebot.com

Faculty of Medicine
Masaryk University
kvak@mail.muni.cz

RNDr. Marek Biroš, Ph.D.
Carebot s.r.o.
marek.biros@carebot.com

MUDr. Jindřiška Judasová
Department of Radiodiagnostics
AGEL Šternberk Hospital
jindriska.judasova@nst.agel.cz