

Article

AI in Diagnostic Radiology: What Happens When Algorithms Are Updated

Martine Rustøen Skregelid ^{1,†}, Kasim Ibrahim-Pur ^{1,†} , Flemming Skjøth ^{1,2} ,
Malene Roland Vils Pedersen ^{1,3,4,5,6,*}  and Helle Precht ^{1,3,4,6,7} 

- ¹ Department of Regional Health Research, University of Southern Denmark, 5230 Odense, Denmark
- ² Research Support Unit, Lillebaelt Hospital, University Hospitals of Southern Denmark, 7100 Vejle, Denmark
- ³ Radiology Department, Lillebaelt Hospital, University Hospitals of Southern Denmark, 7100 Vejle, Denmark
- ⁴ Department of Radiology, Lillebaelt Hospital, University Hospitals of Southern Denmark, 6000 Kolding, Denmark
- ⁵ Department of Radiography, VIA University College, 7400 Herning, Denmark
- ⁶ Discipline of Medical Imaging & Radiation Therapy, School of Medicine, University College Cork, T12 K8AF Cork, Ireland
- ⁷ Health Sciences Research Center, UCL University College, 5230 Odense M, Denmark
- * Correspondence: malp@via.dk
- † These authors contributed equally to this work.

Simple Summary

Hand and ankle fractures are common and can easily be missed in busy emergency departments. Therefore, many hospitals now use artificial intelligence (AI) to highlight possible fractures. AI software is frequently updated, and regular evaluation is warranted to assess whether new updates alter the diagnostic performance. In this study, we investigated the diagnostic accuracy of two different versions of the same AI fracture detection software on hand and ankle radiographs, using reporting radiographers' reports as the reference standard. We also assessed whether AI output led to changes in reporting radiographers' assessments in cases of disagreement.

Abstract

Background: Interpretation of radiographs is prone to diagnostic errors. Artificial intelligence (AI) has shown promising results in fracture detection, although systematic evaluation of software updates remains limited. This study compares the diagnostic performance of two versions of an AI-based fracture detection software in hand and ankle radiographs and assesses the influence of AI output on diagnostic decisions. **Methods:** This retrospective diagnostic accuracy study included 193 hand and ankle examinations obtained during routine clinical practice at Lillebaelt Hospital, Denmark. Radiographs were analysed using two versions of the same AI software and compared with the diagnostic report as the reference standard. Diagnostic performance of both versions was assessed using diagnostic accuracy metrics. Exploratory subgroup analyses were conducted to further investigate the difference in performance. The influence of AI was evaluated by the proportion of reports revised after review of AI output. **Results:** The newest software version demonstrated higher diagnostic performance than the older one (accuracy 0.933 vs. 0.824; $p < 0.001$). Similar improvements were observed across patient subgroups. Excluding radiographs containing casts resulted in only minimal changes in performance (accuracy in version 2: 0.930 vs. 0.933). In 8 of 15 discordant cases, reporting radiographers revised the initial assessment upon reassessment. **Conclusions:** The newest version demonstrated higher overall diagnostic performance, indicating that software updates can enhance the accuracy of AI-assisted fracture detection. The proportion of revised assessments suggests that radiographers' decisions may be influenced by AI output.



Academic Editor: Sam Payabvash

Received: 3 January 2026

Revised: 18 January 2026

Accepted: 23 January 2026

Published: 26 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: artificial intelligence; fracture detection; hand fractures; ankle fractures; radiography; software update

1. Introduction

Misdiagnosed fractures in plain radiographs can lead to delayed treatment, long-term disability and reduced quality of life [1]. Artificial intelligence (AI)-based software has shown promise in reducing the number of missed fractures [1,2]; however, given that AI software is frequently updated, its performance should be evaluated regularly to avoid overreliance and to ensure that new software versions do not compromise diagnostic accuracy [3].

Hospitals worldwide, including the Nordic countries, have started integrating AI into clinical practice [4–7]. At Lillebaelt Hospital, University Hospitals of Southern Denmark, a total of 148,690 radiographic examinations were performed in 2024 [8]. To address the issue of missed fractures, the hospital implemented an AI-based software for automatic detection of trauma-related findings across the appendicular skeleton. The manufacturer reports that the current version achieves 94% accuracy and an 86% reduction in missed fractures, based on data from 319 cases from the Kettering General Hospital in the UK [9].

Several studies have demonstrated that AI can achieve high sensitivity and specificity for fracture detection on radiographs [2,10–12]. In addition, AI has been shown to reduce workload, support routine tasks and shorten reading time by 27%, according to a meta-analysis published in 2024 [13]. Radiology is the medical field with the most extensive application of AI [14], reflected by the recent increase in AI-related publications from 100–150 per year to 700–800 per year [15].

However, the use of AI in clinical settings may also raise important concerns. Algorithms often function as “black boxes” and do not provide insights into how decisions are made [16]. Also, performance may vary if training data is not representative of the investigated population [17]. Overreliance on AI may lead to false reassurance, and because AI tools are continuously updated, it is essential to determine whether new versions improve or compromise diagnostic performance.

The aim of this study was to compare the diagnostic performance of two different software versions (v1 and v2) of an AI-based algorithm for detecting hand and ankle fractures on radiographs, using the reporting radiographers’ diagnostic report as the reference standard. In addition, the study aimed to assess the influence of AI output on diagnostic decisions.

2. Materials and Methods

2.1. AI Software and Image Management

The AI software used in this study was RBfracture (Radiobotics, Copenhagen, Denmark), a CE-marked class IIa medical device [18]. RBfracture is a clinical decision-supporting tool designed to assist clinicians with diagnoses such as fractures, lipohemarthroses, effusions and dislocations when there is a clinical suspicion of a new fracture. Two versions of the AI software, 1.8.1 (v1) and 2.1.1 (v2), were investigated. Radiographic images were acquired in the emergency departments at Lillebaelt Hospital (Kolding and Vejle), University Hospitals of Southern Denmark.

Medical imaging data was anonymised, and radiographs were reviewed using a DICOM viewer. For this study, Weasis medical viewer (version 4.6.3; Weasis Team, Geneva, Switzerland) was primarily used, with cross-checking performed in either Bee Dicom Viewer (version 2.6.2; Sainuo United Medical Technology, Beijing, China) for macOS or MicroDicom (version 2025.3; MicroDicom Ltd., Sofia, Bulgaria) for Windows.

2.2. Study Design

This diagnostic accuracy study was designed in accordance with the STARD guidelines (Standards for Reporting Diagnostic Accuracy Studies) [19]. See Supplementary Table S1.

2.3. Study Population and Materials

All radiographic examinations were acquired in 2023 as part of routine clinical diagnostics at Lillebaelt Hospital, Vejle and Kolding. Inclusion criteria were as follows: (i) patients aged ≥ 18 years; (ii) clinical suspicion of hand or ankle fracture; and (iii) availability of complete hand or ankle radiographs with a diagnostic report. Exclusion criteria were as follows: (i) patients < 18 years; (ii) examinations performed as follow-up of older fractures; (iii) examinations with inconclusive or insufficient reports by reporting radiographers; or (iv) missing v1 AI-generated output. Each examination comprised multiple projections, acquired according to routine imaging protocols. To ensure random selection within each anatomical group, all examinations were assigned a unique computer-generated random number in Microsoft Excel. The list was then sorted by this value, and the first 100 examinations in each group were selected, reaching a total of 200 patients.

Fractures were defined according to the radiographers' report and included displaced fractures, non-displaced fractures and avulsion fractures. Although all fracture types were eligible for inclusion, the available descriptions did not allow for a consistent classification of fracture morphology across the dataset. Information on patient comorbidities was not available, and image acquisition parameters were not standardised for the purpose of this study.

At the time of image acquisition in 2023, RBfracture v1 was used as part of routine clinical diagnostics. Seven patients were excluded due to missing v1 AI-generated output, resulting in a final dataset of 193 patients: 94 hand examinations and 99 ankle examinations (Figure 1). Hand radiographs comprised the distal radius and ulna, carpal bones and phalanges, while ankle radiographs included the distal tibia and fibula, malleolar region and foot [20].

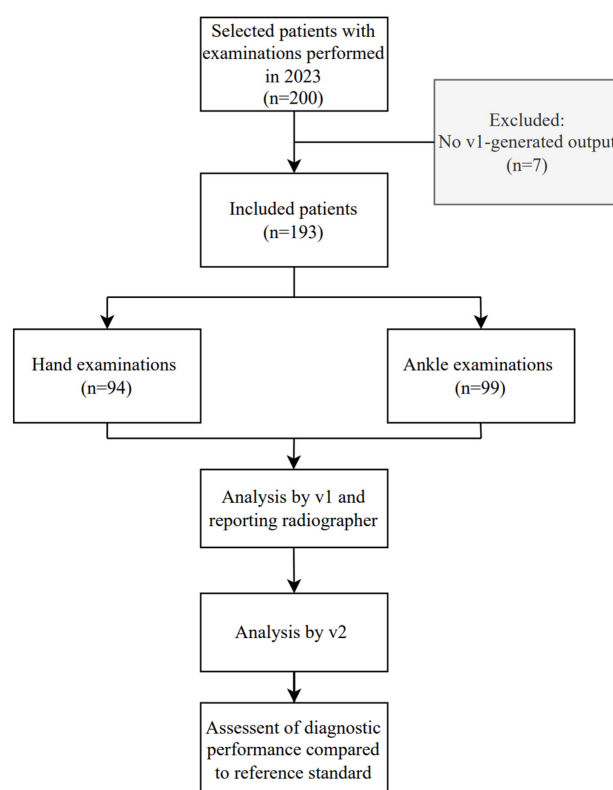


Figure 1. Flow diagram of population selection. Radiographs were analysed by version 1 and by reporting radiographers in 2023 and re-analysed by version 2 in 2025.

2.4. Study Procedure

During image acquisition in 2023, all radiographic examinations were analysed by v1 as part of the usual workflow. v1 produced an output image with AI-generated analysis, providing an output for each projection (Figure 2). Examinations were then analysed by a reporting radiographer, i.e., a radiographer who has completed an additional two-year advanced competency programme [21]. The v1 output was available during this analysis and could be used as decision support before the reporting radiographer finalised the report. Each radiographic examination was assessed by one reporting radiographer. Detailed information regarding the exact number of radiographers involved, individual years of experience, or specific reporting workflow characteristics was not systematically recorded.

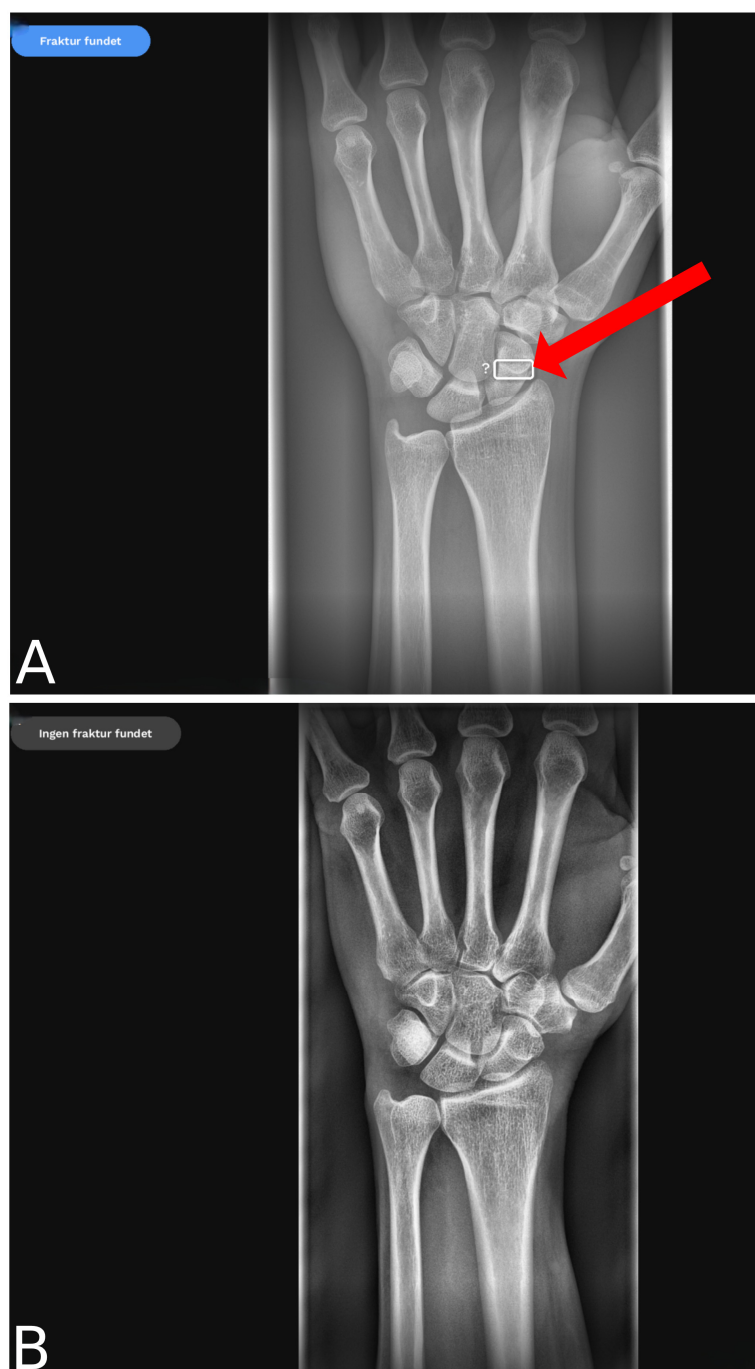


Figure 2. Example of an X-ray analysis by RBfracture version 1. (A) X-ray of left hand where the software detects a fracture and highlights the suspected region with a square (red arrow). A blue

notification box in the upper left corner indicates that a fracture has been detected ('Fraktur fundet', Danish for 'Fracture detected'). (B) X-ray of left hand where no fracture is detected by the programme ('Ingen fraktur fundet', Danish for 'No fracture detected').

In May 2025, the radiographs were re-evaluated for the present study using v2. In this version, a summary image was generated that integrated all projections from the examination, displaying the AI-generated output (Figure 3).

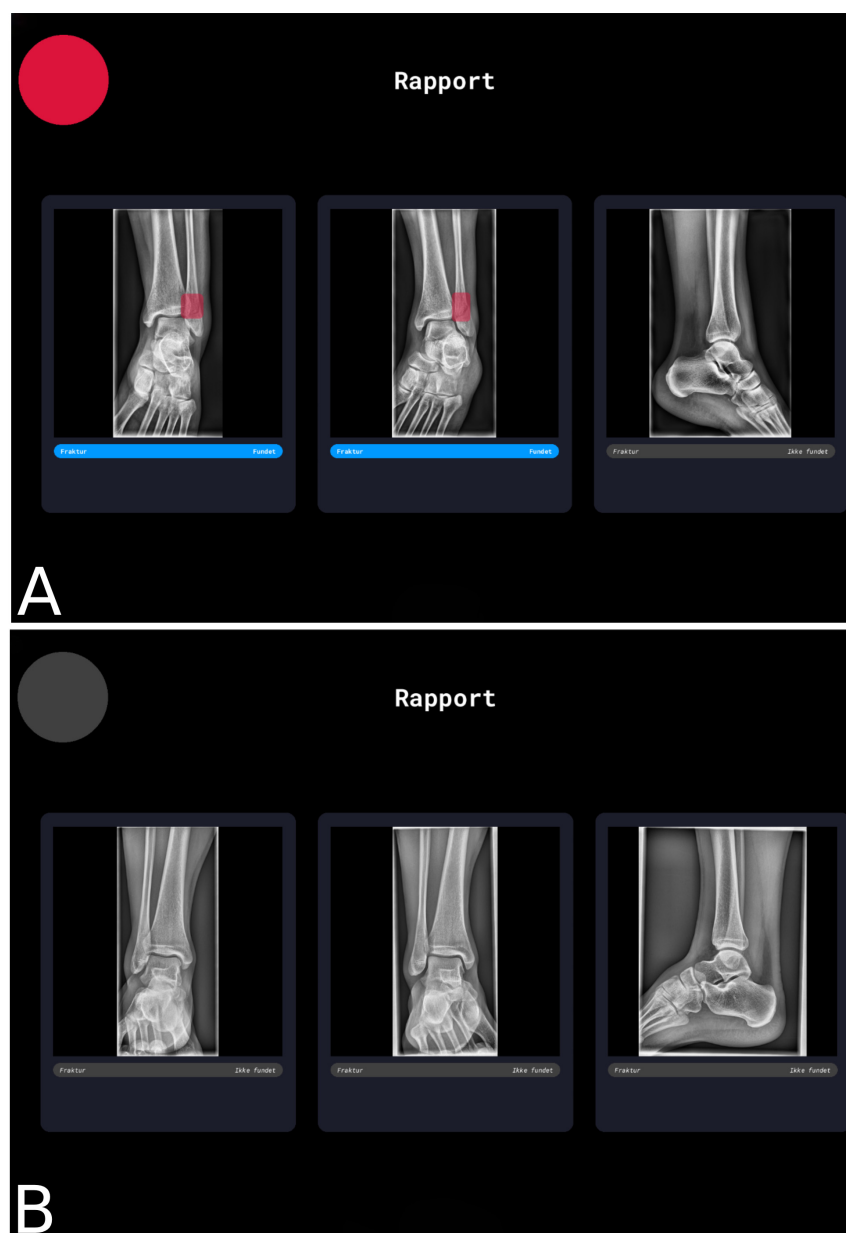


Figure 3. Example of an X-ray analysis by RBfracture version 2. (A) X-ray of right ankle. Summary view combining all X-ray images. In this example, fractures were detected in two of the images, indicated by the red notification dot on the left side of the image. (B) X-ray of left ankle. Summary view combining all X-ray images, showing no fractures detected in this example, as indicated by the labels under each image ("Fraktur Ikke fundet", Danish for "Fracture Not detected").

The AI outputs were interpreted by reviewing the output images, and results from both v1 and v2 were compared with the original radiographer reports, which served as the reference standard for evaluating algorithm performance.

2.5. Analysis

AI outputs from both software versions classified each radiograph as positive, negative or inconclusive for the presence of a fracture. The system does not provide further details on the likelihood of a fracture. For the analysis, inconclusive outputs were treated as negative to obtain a binary outcome for the diagnostic accuracy analyses. To evaluate the robustness of this approach, an additional sensitivity analysis was performed, in which inconclusive outputs were classified as positive. Data tabulation and initial descriptive analyses were calculated in Microsoft Excel (version 16.101.2) and cross-checked in R, while all statistical analyses were performed in R using RStudio (version 2025.05.1 + 513).

2.5.1. Overall Performance

Diagnostic performance was quantified using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy, calculated by the following formulas:

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}),$$

where TP are true positives and FN are false negatives.

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}), \quad (1)$$

where TN are true negatives.

$$\text{PPV} = \text{TP} / (\text{TP} + \text{FP}), \quad (2)$$

where FP are false positives.

$$\text{NPV} = \text{TN} / (\text{TN} + \text{FN}), \quad (3)$$

where TN are true negatives.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}). \quad (4)$$

Comparison of sensitivity, specificity and accuracy between the versions was performed in R using McNemar's chi-squared test for paired binary outcomes, restricted to reference positive for sensitivity and reference negative for specificity; accuracy was assessed across all samples. A significance level of 0.05 was used for all statistical tests. Comparisons of differences in performance metrics were conducted in R using bootstrap percentile confidence intervals (1000 resamples), as these measures may not be normally distributed. Calculation was performed by resampling patients at the examination level, thereby preserving the paired structure between v1 and v2.

2.5.2. Exploratory Subgroup Analyses

To further investigate the difference between the versions, exploratory subgroup analyses were performed. Potential demographic and anatomical variations were assessed by stratifying the data by sex (male and female), age groups (18–30, 31–60 and 61–99 years) and anatomical location (hand and ankle). Diagnostic accuracy metrics were calculated in R for each subgroup.

To assess whether visible bandages or casts affected algorithm performance, all such examinations were excluded to form a modified dataset. Diagnostic performance metrics were recalculated using R, and differences in accuracy between the original and modified datasets were compared descriptively.

2.5.3. Influence of AI

Finally, to assess the influence of AI on radiographers, cases with disagreement between v2 and the original radiographer report were re-evaluated by a single reporting

radiographer. The radiographer decided whether to revise or maintain the original reports, and the number of revised cases was recorded using Microsoft Excel. Disagreement with v1 was not reassessed, as v1 findings were already available to the radiographers upon initial assessment.

3. Results

A total of 193 patients were included (82 males [42%], 111 females [58%], with an age range of 18–99 years; mean age = 50 years). Clinical diagnoses included fractures, dislocations, effusions, arthrosis, chondrocalcinosis, traumatic amputations and non-pathological findings.

3.1. Overall Performance

Of the 193 examinations, 88 (46.6%) contained one or more fractures according to the reference standard. The remaining 105 examinations were classified as negative. v1 identified 90 positive, 85 negative and 18 inconclusive examinations, whereas v2 identified 85 positive, 108 negative and zero inconclusive.

Compared with v1, v2 demonstrated higher diagnostic performance. Overall, v2 achieved a sensitivity of 0.909, specificity 0.952, PPV 0.941, NPV 0.925 and accuracy 0.933, compared with 0.818, 0.828, 0.800, 0.844 and 0.824 for v1. The largest improvements were observed for PPV (difference 0.14, 95% CI 0.06–0.21), followed by specificity (difference 0.12, $p = 0.002$) and accuracy (difference 0.11, $p < 0.001$). Sensitivity increased by 0.09, but this difference did not reach statistical significance ($p = 0.061$). Detailed diagnostic metrics are shown in Tables 1 and 2.

Table 1. Diagnostic performance metrics for two versions of AI fracture-detection software against the radiographer reference standard ($n = 193$). p -values from the McNemar test; 95% confidence intervals reflect bootstrap percentile estimates of between-version estimates.

	v1	v2	p -Value	Difference (95% CI)
True positives	72	80		
True negatives	87	100		
False positives	18	5		
False negatives	16	8		
Sensitivity	0.818	0.909	0.061	–
Specificity	0.828	0.952	0.002	–
PPV	0.800	0.941	–	0.14 (0.06–0.21)
NPV	0.844	0.925	–	0.08 (0.03–0.15)
Accuracy	0.824	0.933	<0.001	–

PPV = positive predictive value; NPV = negative predictive value; CI = confidence interval; v1 = version 1; v2 = version 2. Difference ($v2 - v1$).

Table 2. Contingency table underlying the McNemar test for sensitivity and specificity, restricted to reference-positive (sensitivity, $n = 88$) and reference-negative (specificity, $n = 105$) examinations.

Sensitivity ($n = 88$)	v2 Fracture	v2 no Fracture
v1 fracture	69	3
v1 no fracture	11	5
Specificity ($n = 105$)	v2 no Fracture	v2 Fracture
v1 no fracture	86	1
v1 fracture	14	4

v1 = version 1; v2 = version 2.

In the additional sensitivity analysis in which inconclusive results were treated as positive, absolute performance metrics differed from the primary analysis, but the relative performance between AI versions remained unchanged (Supplementary Table S2).

3.2. Exploratory Subgroup Analyses

3.2.1. Anatomical Location

Diagnostic performance was higher for ankle radiographs than for hand radiographs across both AI versions. Within each anatomical location, v2 consistently demonstrated a numerically higher sensitivity, specificity, PPV, NPV and accuracy compared with v1 (Table 3).

Table 3. Diagnostic accuracy for fracture detection of version 1 and version 2 of the AI software across anatomical location, sex, age groups and in the modified dataset without visible casts.

Subgroup	Version	Sensitivity	Specificity	PPV	NPV	Accuracy
Location						
Hand (n = 94)	v1	0.714	0.808	0.750	0.778	0.766
	v2	0.881	0.942	0.925	0.907	0.915
Ankle (n = 99)	v1	0.913	0.849	0.840	0.918	0.879
	v2	0.935	0.962	0.956	0.944	0.949
Sex						
Male (n = 82)	v1	0.765	0.812	0.743	0.830	0.793
	v2	0.882	0.917	0.882	0.917	0.902
Female (n = 111)	v1	0.852	0.842	0.836	0.857	0.874
	v2	0.926	0.982	0.980	0.933	0.955
Age group						
18–30 (n = 44)	v1	1.000	0.900	0.824	1.000	0.930
	v2	0.929	0.667	0.929	0.967	0.953
31–60 (n = 85)	v1	0.778	0.857	0.800	0.840	0.819
	v2	0.889	0.959	0.94	0.922	0.928
61–99 (n = 64)	v1	0.789	0.692	0.789	0.692	0.750
	v2	0.921	0.923	0.946	0.889	0.922
No cast (n = 173) *	v1	0.794	0.829	0.750	0.861	0.815
	v2	0.897	0.952	0.942	0.935	0.930

PPV = positive predictive value; NPV = negative predictive value. * Modified dataset excluding radiographs containing visible bandages or casts.

3.2.2. Sex

Overall diagnostic performance was generally higher for radiographs from female patients than from male patients. Across both sex categories, v2 outperformed v1 in all evaluated metrics (Table 3).

3.2.3. Age Groups

Across all age strata, v2 demonstrated higher accuracy than v1. In the 31–60 and 61–99 age groups, v2 showed improvement across all diagnostic metrics. In contrast, in the youngest age group (18–30 years), v1 demonstrated higher sensitivity, specificity and NPV than v2, whereas PPV and accuracy were higher for v2. Both AI versions performed best in the youngest patient group and poorest in the oldest patient group (Table 3).

3.2.4. Bandages/Casts

After excluding 20 examinations containing visible bandages or casts, 173 radiographs remained for analysis. v2 continued to demonstrate higher diagnostic performance than v1 across all evaluated metrics in this modified dataset (Table 3). Overall diagnostic accuracy

differed only marginally between the original dataset and the modified dataset without visible casts for both AI versions (Supplementary Table S3), with absolute differences in accuracy of 0.008 for v1 and 0.002 for v2.

3.3. Radiographer Reassessment of Discordant Cases

A total of 15 examinations were re-evaluated due to discrepancies between v2 and the original radiographer's report. Following reassessment, the radiographer revised the initial interpretation in 8 of the 15 cases, while 7 of the 15 evaluations remained unchanged. In most instances, the interpretation was revised from positive to negative. Reassessment was not performed for v1, as v1 outputs were available during original reporting.

4. Discussion

This study compared two versions of RBfracture AI software for detecting fractures in hand and ankle radiographs and demonstrated high overall diagnostic performance. We found that v2 outperformed v1 in the overall performance metrics, with exploratory subgroup analyses showing only minor variations but a consistently favourable pattern. Additionally, more than half of the cases with disagreement between the AI output and the reporting radiographer's initial interpretation resulted in a revised interpretation.

Previous studies have assessed RBfracture software in clinical settings and have reported diagnostic accuracies comparable to those observed in the present study. Ziegner et al. reported a sensitivity of 92%, a specificity of 83% and an accuracy of 87% for fracture detection across the axial skeleton [12]. Chan et al. demonstrated a specificity of 95.5% in fractures across the body in an emergency setting [22]. These studies evaluated a single software version, whereas the present study compared two versions. To our knowledge, no previous studies have compared two versions of RBfracture software.

Several factors may explain the improved performance of v2. For instance, v1 appeared more likely to classify radiographs containing osteosynthesis material as inconclusive, whereas v2 more often provided a definitive classification. However, the presence of osteosynthetic material was not systematically recorded in the present study. Interestingly, excluding radiographs with casts resulted in only small performance differences, suggesting that improvement cannot be explained solely by cast handling. As the distribution of fracture types was not assessed, fracture type-specific analyses were not feasible. Therefore, it cannot be determined whether the improved performance observed in v2 is related to the detection of specific fracture patterns. Furthermore, as radiographs were acquired under routine clinical conditions and information on bone density and acquisition parameters was unavailable, it remains unclear whether the improved performance of v2 reflects enhanced fracture detection or increased robustness to variations in radiographic attenuation, particularly in older patients.

The handling of inconclusive outputs from v1 represents an important methodological consideration. Inconclusive results do not provide a definitive diagnosis and offer limited clinical utility, as they neither confirm nor exclude the presence of a fracture. To enable statistical analysis, inconclusive results were classified into a binary outcome. Reclassification of inconclusive results as positive affected the absolute diagnostic performance, but the comparative advantage of v2 over v1 was preserved. Accordingly, inconclusive outputs were classified as negative, acknowledging the inherent limitation of the analysis. The absence of inconclusive results in v2, despite analysis of the same radiographs, suggests improved algorithm robustness.

The higher predictive values observed by v2 may have practical relevance in emergency situations with a high patient load, where AI-based systems could help prioritise cases and reduce workflow pressure. Nevertheless, the continued need for human verifica-

tion underscores that AI should complement, rather than replace, radiological expertise. The reassessment of discordant cases indicates that AI output may influence radiographers' interpretations. However, as reassessment was performed without an independent reference standard, it remains unclear whether the observed changes reflect improved diagnostic accuracy or decision modification following exposure to AI output. This analysis should therefore be interpreted as an assessment of decision modification rather than diagnostic improvement.

One potential advantage of AI in diagnostic imaging is its immunity to the "satisfaction of search" effect, in which detection of one abnormality reduces the likelihood of identifying additional ones. Berbaum et al. demonstrated this phenomenon by showing reduced diagnostic accuracy when simulated lesions were added to chest radiographs [23]. At the same time, excessive reliance on AI introduces the risk of automation bias, defined by Khera et al. as the tendency to over-rely on assistive technologies, potentially leading to inappropriate diagnostic changes when the algorithm is incorrect [24].

This study has several strengths. It focused on two anatomical locations—hand and ankle—that are common fracture sites [17] and can be diagnostically challenging, making them suitable for evaluating diagnostic accuracy. Additionally, the dataset represents a real-world case mix from routine clinical practice across two hospitals, increasing heterogeneity and thereby enhancing external validity.

Several limitations should be acknowledged. The overall sample size was relatively modest, and the number of discordant cases was small, particularly after stratification into subgroups. As a result, the subgroup analyses were exploratory in nature, and their findings should be interpreted with caution. A key structural methodological limitation of this study is the reference standard, which was based on reporting radiographers who had access to v1 output during initial reporting. This may have introduced incorporation bias [25], as the AI output from v1 could have influenced the radiographer's interpretation. Consequently, the diagnostic performance of v1 may have been overestimated, thereby attenuating the observed performance differences between the two versions and potentially underestimating the performance gap. The use of multiple reporting radiographers with varying levels of expertise reflects routine clinical practice but may have introduced inter-reader variability, which should be considered when interpreting the results. Also, as the number of reporting radiographers was not assessed, it is unclear whether the same individuals contributed at both the original assessment and reassessment. This may have influenced the likelihood of changing diagnoses and the reassessment outcomes. Furthermore, as the reference standard was human-derived, diagnostic errors such as missed or occult fractures cannot be entirely excluded. The reference standard was also based on reports from reporting radiographers rather than radiologists. Although this reflects routine clinical practice in Denmark [26], it differs from many other healthcare systems. However, previous studies have demonstrated comparable diagnostic accuracy between reporting radiographers and consulting radiologists in musculoskeletal imaging, suggesting that this approach is unlikely to have introduced substantial bias [27].

The observed performance differences between software versions highlight the clinical relevance of algorithm updates. These changes may not be apparent in routine clinical use unless systematically evaluated. It may therefore be recommended to perform local performance assessments following the implementation of new versions. Future studies are needed to validate these observed findings between v1 and v2 and examine if these findings are seen in other anatomical regions.

5. Conclusions

AI demonstrated high diagnostic accuracy for the detection of fractures in hand and ankle radiographs and may therefore be a valuable tool in clinical practice. Diagnostic performance differed between software versions, with v2 showing improved accuracy compared with v1, highlighting the potential impact of algorithm updates on clinical performance. The proportion of revised interpretations following reassessment suggests that radiographers' clinical decision-making may be influenced by AI-generated findings.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/radiation6010004/s1>, Table S1: STARD 2025 checklist for reporting of diagnostic accuracy studies. Table S2: Diagnostic performance when classifying inconclusive AI outputs as positive. Table S3: Comparison of accuracy in original dataset including casts and in modified dataset excluding casts.

Author Contributions: Conceptualization, M.R.V.P.; methodology, M.R.V.P., M.R.S. and K.I.-P.; formal analysis, M.R.S. and K.I.-P.; investigation, M.R.S. and K.I.-P.; writing—original draft preparation, M.R.S. and K.I.-P.; writing—review and editing, all authors; supervision, M.R.V.P., H.P. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study received ethical approval from the hospital institution board and the region of Southern Denmark local data protection (25/71).

Informed Consent Statement: Patient consent was waived due to the retrospective data collection.

Data Availability Statement: The data presented in this study is not publicly available due to privacy and ethical restrictions.

Acknowledgments: During the preparation of this study, the authors used Generative AI (ChatGPT (GPT-5.2), OpenAI) for the purpose of grammar, structure and part of the analytical code used in Rstudio (version 2025.09.0+387). The authors have reviewed and edited the output and take full responsibility for the content of this publication. A big thanks to reporting radiographer Lena Lassen, Department of Radiology, Vejle for various assistance.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lakhani, P.; Prater, A.; Hutson, K.R.; Andriole, K.P.; Dreyer, K.J.; Morey, J.; Prevedello, L.M.; Clark, T.J.; Geis, J.R.; Itri, J.N.; et al. Machine Learning in Radiology: Applications Beyond Image Interpretation. *J. Am. Coll. Radiol.* **2018**, *15*, 350–359. [[CrossRef](#)] [[PubMed](#)]
2. Lindsey, R.; Daluiski, A.; Chopra, S.; Lachapelle, A.; Mozer, M.; Sicular, S.; Hanel, D.; Gardner, M.; Gupta, A.; Hotchkiss, R.; et al. Deep neural network improves fracture detection by clinicians. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 11591–11596. [[CrossRef](#)] [[PubMed](#)]
3. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Coriado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**, *17*, 195. [[CrossRef](#)] [[PubMed](#)]
4. Tucker, J. The Future Vision(s) of AI Health in the Nordics: Comparing the National AI Strategies. *Futures* **2023**, *149*, 103154. [[CrossRef](#)]
5. Nygren, J. AI Healthcare Map Will Create Opportunities for Collaboration. 2025. Available online: <https://www.hh.se/english/information-english/news/news/2025-04-15-ai-healthcare-map-will-create-opportunities-for-collaboration.html> (accessed on 19 September 2025).
6. Ministry of Foreign Affairs of Denmark. AI-Based Algorithm in Chest X-Rays Will Enable Quicker and Better Diagnoses. 2021. Available online: <https://investindk.com/insights/ai-based-algorithm-in-chest-x-rays-will-enable-quicker-and-better-diagnoses> (accessed on 18 September 2025).
7. Radiological Artificial Intelligence Testcenter. Who Is RAIT. Available online: <https://www.rait.dk/who-is-rait> (accessed on 19 September 2025).

8. Sundhedsdatabanken. Sygehus og Speciallæge. 2025. Available online: <https://sundhedsdatabank.dk/behandling-og-pleje/sygehus-og-speciallaege> (accessed on 30 September 2025).
9. Radiobotics. *Reducing Missed Fractures at Kettinger General Hospital*; Radiobotics: København, Denmark, 2025.
10. Nowroozi, A.; Salehi, M.A.; Shobeiri, P.; Agahi, S.; Momtazmanesh, S.; Kaviani, P.; Kalra, M.K. 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*, 579–588. [[CrossRef](#)] [[PubMed](#)]
11. Jung, J.; Dai, J.; Liu, B.; Wu, Q. Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. *PLoS Digit. Health* **2024**, *3*, e0000438. [[CrossRef](#)] [[PubMed](#)]
12. Ziegner, M.; Pape, J.; Lacher, M.; Brandau, A.; Kelety, T.; Mayer, S.; Wolfgang, F.; Hirsch, F.W.; Rosolowski, M.; Gräfe, D. Real-life benefit of artificial intelligence-based fracture detection in a pediatric emergency department. *Eur. Radiol.* **2025**, *35*, 5881–5890. [[CrossRef](#)]
13. Chen, M.; Wang, Y.; Wang, Q.; Shi, J.; Wang, H.; Ye, Z.; Xue, P.; Qiao, Y. Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. *npj Digit. Med.* **2024**, *7*, 349. [[CrossRef](#)] [[PubMed](#)]
14. Joshi, G.; Jain, A.; Araveeti, S.R.; Adhikari, S.; Garg, H.; Bhandari, M. FDA-Approved Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: An Updated Landscape. *Electronics* **2024**, *13*, 498. [[CrossRef](#)]
15. Pesapane, F.; Codari, M.; Sardanelli, F. Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur. Radiol. Exp.* **2018**, *2*, 35. [[CrossRef](#)] [[PubMed](#)]
16. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [[CrossRef](#)] [[PubMed](#)]
17. GBD 2019 Risk Factors Collaborators. Global, regional, and national burden of bone fractures in 204 countries and territories, 1990–2019: A systematic analysis from the Global Burden of Disease Study 2019. *Lancet Healthy Longev.* **2021**, *2*, e580–e592. [[CrossRef](#)]
18. Radiobotics. About Radiobotics—Our Vision. Available online: <https://radiobotics.com/company/#vision> (accessed on 22 January 2026).
19. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Glasziou, P.P.; Irwig, L.; Lijmer, J.G.; Moher, D.; Rennie, D.; de Vet, H.C.W.; et al. STARD2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. 2015. Available online: <https://www.equator-network.org/reporting-guidelines/stard/> (accessed on 22 January 2026).
20. Radiologisk Afdeling, Odense Universitets Hospital; Dansk Forening for Muskuloskeletal Radiologi. Knoglebogen. Available online: <https://knoglebogen.dk> (accessed on 12 December 2025).
21. Sygehus Lillebælt. Uddannelse til Beskrivende Radiograf. 2025. Available online: https://www.sdu.dk/da/uddannelse/efter_videreuddannelse/kurser/beskrivende-radiograf (accessed on 22 September 2025).
22. Chan, H.Y.; Tang, U.P.; Wong, Z.Y.; Koh, S.H.; Nickalls, O.; Steven, W.; Tan, M.O. Artificial Intelligence (AI) in a Singaporean Emergency Department: Detecting fractures and reducing recalls. *Med. J. Malays.* **2025**, *80*, 462–465.
23. Berbaum, K.S.; Franken, E.A.; Dorfmann, D.D.; Rooholamini, S.A.; Kathol, M.H.; Barloon, T.; Bejlike, F.M.; Sato, Y.; Lu, C.H.; el-Khoury, G.Y.; et al. Satisfaction of search in diagnostic radiology. *Investig. Radiol.* **1990**, *25*, 133–140. [[CrossRef](#)] [[PubMed](#)]
24. Khera, R.; Simon, M.A.; Ross, J.S. Automation Bias and Assistive AI—Risk of Harm from AI-Driven Clinical Decision Support. *JAMA* **2023**, *330*, 2255–2257. [[CrossRef](#)] [[PubMed](#)]
25. Lynøe, N.; Eriksson, A. Disguised incorporation bias and meta-analysis of diagnostic accuracy studies. *Acta Paediatr.* **2024**, *113*, 503–505. [[CrossRef](#)] [[PubMed](#)]
26. Radiograf Rådet. Bliv Radiograf. 2025. Available online: <https://radiograf.dk/fag-og-viden/bliv-radiograf/> (accessed on 22 September 2025).
27. Cain, G.; Pittock, L.J.; Piper, K.; Venumbaka, M.R.; Bodoceanu, M. Agreement in the reporting of General Practitioner requested musculoskeletal radiographs: Reporting radiographers and consultant radiologists compared with an index radiologist. *Radiography* **2022**, *28*, 288–295. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.