International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	ARTIFICIAL INTELLIGENCE IN RADIOGRAPHIC FRACTURE DIAGNOSIS: CURRENT APPLICATIONS, CLINICAL UTILITY, AND LIMITATIONS: A NARRATIVE REVIEW
DOI	https://doi.org/10.31435/ijitss.2(50).2026.6148
RECEIVED	18 April 2026
ACCEPTED	23 June 2026
PUBLISHED	30 June 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

ARTIFICIAL INTELLIGENCE IN RADIOGRAPHIC FRACTURE DIAGNOSIS: CURRENT APPLICATIONS, CLINICAL UTILITY, AND LIMITATIONS: A NARRATIVE REVIEW

Mateusz Onopiuk *(Corresponding Author, Email: mateusz400@op.pl)*
Bielański Hospital named after Rev. Jerzy Popiełuszko, Warsaw, Poland
ORCID ID: 0009-0003-9916-1292

Jowita Wiktoria Maksymiuk
Bielański Hospital named after Rev. Jerzy Popiełuszko, Warsaw, Poland
ORCID ID: 0009-0004-3130-9913

Urszula Gadomska
Bielański Hospital named after Rev. Jerzy Popiełuszko, Warsaw, Poland
ORCID ID: 0009-0005-8381-9955

Natalia Dejewska
Bielański Hospital named after Rev. Jerzy Popiełuszko, Warsaw, Poland
ORCID ID: 0009-0006-4806-6780

Zuzanna Walewska
Bielański Hospital named after Rev. Jerzy Popiełuszko, Warsaw, Poland
ORCID ID: 0009-0001-0088-4294

Zuzanna Wiktoria Szumska
Szpital Powiatowy w Otwocku, Otwock, Poland
ORCID ID: 0009-0009-5287-9946

Mikołaj Daniluk
Centralny Szpital Kliniczny, Poland
ORCID ID: 0009-0005-6458-373X

Jonasz Żuk
Centralny Szpital Kliniczny, Poland
ORCID ID: 0009-0001-7500-7982

Kinga Bukała
Medical University of Lodz, Lodz, Poland
ORCID ID: 0009-0003-9587-4826

Szymon Klimaszewski
Praktyka Lekarska Dentystyczna, Poland
ORCID ID: 0009-0001-3505-2797

ABSTRACT

Artificial intelligence (AI) is increasingly investigated as a supportive tool in radiographic fracture diagnosis, particularly in emergency and musculoskeletal imaging. This narrative review aims to summarize current evidence on the use of AI in fracture detection, clinician-assisted interpretation, fracture classification, workflow optimization, and implementation challenges. A literature search was conducted in PubMed using combinations of terms related to artificial intelligence, deep learning, convolutional neural networks, fracture detection, radiography, emergency radiology, and clinical decision support. Priority was given to systematic reviews, meta-analyses, diagnostic accuracy studies, reader studies, and real-world clinical evaluations. Current evidence indicates that AI-based systems can achieve high diagnostic performance in selected fracture detection tasks, especially when applied to anatomically focused radiographic datasets. Studies assessing AI-assisted interpretation suggest that these tools may improve clinicians' sensitivity, specificity, and overall diagnostic performance, with the greatest benefit observed among less experienced readers and non-radiologist clinicians. AI may also support fracture classification and workflow optimization, although these applications remain less extensively validated. Important limitations include restricted generalizability, reduced performance in subtle or complex fractures, limited external validation, and a shortage of prospective real-world studies. Overall, AI represents a promising but still evolving adjunct in radiographic fracture diagnosis and should be developed as a decision-support tool integrated with clinical expertise.

KEYWORDS

Artificial Intelligence, Deep Learning, Fracture Detection, Fracture Detection, Musculoskeletal Imaging, Emergency Radiology, Clinical Decision Support, Missed Fractures

CITATION

Onopiuk, M., Maksymiuk, J. W., Gadomska, U., Dejewska, N., Walewska, Z., Szumska, Z. W., Daniluk, M., Żuk, J., Bukała, K., & Klimaszewski, S. (2026). Artificial intelligence in radiographic fracture diagnosis: Current applications, clinical utility, and limitations: A narrative review. *International Journal of Innovative Technologies in Social Science*, 2(50). https://doi.org/10.31435/ijitss.2(50).2026.6148

1. Introduction

Traumatic injuries and suspected fractures are among the most common reasons for patient visits to emergency departments, as well as for orthopaedic consultations requested in this setting [18, 24]. Despite the development of modern imaging techniques, plain radiography remains the first-line imaging modality and a fundamental diagnostic tool for the initial assessment of suspected fractures and musculoskeletal injuries [9, 22].

Although X-ray imaging remains a widely available and fundamental diagnostic tool in trauma assessment, the interpretation of radiographs can be challenging, particularly in the emergency department setting. Fractures represent an important source of diagnostic error in this environment. In analyses of errors reported in emergency care, they accounted for a substantial proportion of cases involving incorrect or delayed diagnosis [11]. At the same time, missed or delayed fracture diagnosis on plain radiographs is described in the literature as a common error, with an estimated frequency of approximately 3–10% [17]. This is clinically relevant, as failure to establish the correct diagnosis may delay treatment, worsen patients' functional outcomes, and increase healthcare costs [10, 14]

Emergency imaging is particularly vulnerable to diagnostic errors [21, 24]. Diagnoses often have to be made under time pressure, frequently in the context of incomplete medical history or limited patient cooperation [23]. Work in the emergency department is associated with high levels of stress, fatigue, workload overload, and difficult decision-making, all of which may contribute to occupational burnout. Emergency physicians and radiology residents appear to be especially vulnerable to this phenomenon. In these professional groups, as many as one in two physicians may experience symptoms of burnout, including physical and mental exhaustion related to work, which in turn may increase the risk of diagnostic errors [21, 25-27]. In addition, the growing number of radiographic examinations being performed has made immediate access to expert radiographic interpretation increasingly challenging, particularly during night shifts and on-call hours [17]. As a result, the initial diagnosis and decisions regarding further patient management often have to be based on radiograph interpretation performed by emergency physicians, on-call orthopaedic surgeons, or other clinicians, frequently before the final radiology report becomes available [4, 18].

Such a clinical environment naturally encourages the search for tools that could support the diagnostic process in trauma radiography. In this context, increasing attention has been directed toward artificial intelligence (AI)-based systems, particularly those based on deep learning models, which can be used for the automated analysis of radiographs for the presence of fractures [18]. Their intended role is not to replace physicians in image interpretation, but rather to serve as supportive tools, such as second readers, triage systems, or clinical decision support systems [17]. Beyond simply classifying an image as normal or abnormal, different AI models may highlight suspicious regions on radiographs or even automatically classify fractures, thereby facilitating decisions regarding further patient management.

Previous meta-analyses have demonstrated high sensitivity and specificity of AI systems in fracture detection [15, 17]. However, their actual clinical value depends on how these systems are integrated into physicians' workflow, the experience of their users, and the availability of external validation. Therefore, this literature review aims to summarize the current state of knowledge regarding the use of artificial intelligence in radiographic fracture diagnosis, with particular emphasis on automated fracture detection, clinical decision support, reduction of missed fractures, workflow optimization, and challenges related to implementation.

2. Methodology

A narrative literature review was conducted to identify relevant studies on the use of artificial intelligence in radiographic fracture diagnosis. The PubMed database was searched for articles published within the last ten years using combinations of following keywords: "artificial intelligence", "deep learning", "convolutional neural network", "fracture detection", "fracture diagnosis", "radiography", "emergency radiology". Additional relevant publications were identified by screening the reference lists of selected systematic review, meta-analyses and original studies. Priority was given to systematic reviews, meta-analyses, diagnostic accuracy studies and real-world clinical evaluations assessing AI-based fracture detection, clinical workflow impact and implementation challenges.

3. Results

3.1 Basics of Artificial Intelligence in Medical Imaging

Artificial intelligence (AI) encompasses a group of computational methods designed to perform tasks traditionally associated with human reasoning, such as pattern recognition and data classification [17]. In the context of diagnostic imaging, particularly useful AI-related tasks include pattern recognition, image classification, and support for diagnostic decision-making [8]. Although AI has found applications as a supportive tool in many areas of medicine [19], diagnostic imaging represents one of the fields particularly well suited to the development of this technology, as it is based on digital image data that can be analysed algorithmically.

Machine learning (ML) is a subfield of AI in which an algorithm is not programmed solely by fixed rules, but instead learns patterns and relationships from training data. Deep learning (DL), in turn, is a form of machine learning that uses multilayer artificial neural networks. In medical image analysis, convolutional neural networks (CNNs) are of particular importance, as they are designed to recognize visual features in images. Unlike many traditional machine learning methods, in which relevant image features had to be manually defined in advance, deep learning models can learn these features directly from the data [3, 17].

Convolutional neural networks analyse images through successive layers. Early layers identify basic features, such as edges, contours, and contrast differences, whereas subsequent layers combine this information into more complex patterns that may correspond to specific anatomical structures or pathological findings. In fracture diagnosis on radiographs, CNN-based models may detect features suggestive of bone injury, such as cortical discontinuity, fracture lines, displacement of bone fragments, deformity of the bone axis, or other post-traumatic abnormalities. The large number of radiographic examinations performed in clinical practice, together with the relative standardization of radiographic projections, makes plain radiographs a potentially useful source of data for training such models, provided that image quality is adequate and the training data are reliably labelled [3, 18].

Studies on fracture detection have also frequently used transfer learning, which involves applying networks pre-trained on large image datasets and subsequently adapting them to a specific medical task, such as analysis of a particular anatomical region or a specific radiographic projection [5]. This approach may be especially useful in medical imaging, where the creation of large, well-annotated, and validated datasets is often challenging.

In clinical practice, AI systems may provide different types of output. The simplest approach is binary classification, in which the model estimates the probability of a fracture being present on a given radiograph. More advanced systems may additionally indicate the likely location of pathology, for example by using heatmaps, saliency maps, or bounding boxes. This has practical relevance, as the clinician receives not only information about the possible presence of a fracture, but also an indication of which part of the image requires particularly careful assessment. For this reason, AI systems are most commonly considered as tools supporting image interpretation, functioning as a second reader, a triage system, or a component of clinical decision support, rather than as a replacement for physician assessment [9, 18].

3.2 Current Applications of AI in Fracture Diagnosis

3.2.1 Automated Fracture Detection

The first and most frequently investigated application of artificial intelligence in fracture diagnosis is the automated detection of post-traumatic abnormalities on radiographs. In this setting, AI models analyse an image, or another input imaging modality, and estimate the probability of a fracture being present. More advanced models may also indicate the regions in which the probability of fracture is highest. The performance of such systems has been evaluated both in studies focused on individual anatomical locations using dedicated AI models, and in meta-analyses including different anatomical regions, fracture types, and input data.

Kuo et al., in a meta-analysis of 42 studies, reported high diagnostic performance of AI in fracture detection, with a sensitivity of 92% in internal validation and 91% in external validation, as well as a specificity of 91% in both analyses. The performance achieved by AI did not differ significantly from, and was comparable to, that achieved by clinicians interpreting radiographs. In a more recent analysis including 66 studies, Jung et al. reported similarly high pooled results for AI, with a sensitivity of 91% and a specificity of 90%. Their analysis also showed that models based on imaging data achieved better performance, with 92% sensitivity and 91% specificity, than models based on tabular data, which achieved 81% sensitivity and 83% specificity. The included studies used different types of imaging modalities as input data, including CT, MRI, DXA, and plain radiography. Among these modalities, plain radiography appeared to be associated with the best diagnostic performance. AI models based on radiographs achieved the highest results, with 94% sensitivity and 92% specificity.

In addition to the pooled results reported in meta-analyses, the performance of automated detection systems has also been evaluated in studies of models developed for specific anatomical locations. In such studies, apart from sensitivity and specificity, the area under the receiver operating characteristic curve (AUC) is frequently reported as a measure of the model's ability to distinguish radiographs with fractures from normal images. The closer the AUC value is to 1.0, the better the discriminatory performance of the model. An AUC of 1.0 would indicate perfect differentiation between abnormal and normal images, whereas an AUC of 0.5 represents performance equivalent to random classification [18].

Lindsey et al. evaluated the performance of a deep learning model designed to detect wrist fractures on radiographs. The model achieved AUC values of 0.967 and 0.975 in two test datasets. In a subset of the second test set that included only cases considered by experts to be unambiguous and not subject to disagreement regarding the diagnosis, the AUC reached 0.994.

Chung et al. investigated the performance of a CNN-based model for detecting proximal humeral fractures. In a study including 1,891 radiographs, of which 1,376 showed a fracture, the model achieved an AUC of 0.996.

Studies on pelvic and hip fractures have also demonstrated high diagnostic performance. In the study by Krogue et al., a model designed for hip fracture detection achieved a sensitivity of 92.7%, a specificity of 95%, and an AUC of 0.973. These results were non-inferior to the performance achieved by clinicians. A good example of a more universal diagnostic tool is PelviXNet. This model was trained on 5,204 annotated pelvic and hip radiographs and enables simultaneous analysis for pelvic, hip, and proximal femoral injuries. In the study by Cheng et al., the algorithm achieved a sensitivity of 90.8%, a specificity of 93.2%, and an AUC of 0.973.

The results of studies evaluating commercial algorithms are also promising, although more heterogeneous. Huhtanen et al. compared two commercially available systems, BoneView and RBfracture, and reported similar sensitivity and specificity values for the detection of post-traumatic abnormalities on radiographs of different musculoskeletal regions. BoneView achieved a sensitivity of 89% and a specificity of

86%, whereas RBfracture achieved a sensitivity of 82% and a specificity of 89%. At the same time, the authors emphasized that both algorithms performed worse in subtle abnormalities, which accounted for the majority of false-negative findings. A similar conclusion was reached by Al Hajaj et al.; in their study, false-negative findings accounted for 9.6% of the analysed images despite high overall sensitivity and specificity. Luiken et al. showed that commercial AI systems may achieve satisfactory performance in the general detection of acute fractures, with AUC values ranging from 0.772 to 0.849 depending on the model. However, their performance was weaker in more complex cases, particularly in the presence of multiple fractures, where AUC values ranged from 0.641 to 0.734.

Overall, the available evidence indicates that AI achieves high diagnostic performance as a standalone fracture detection tool, particularly in the analysis of limb and hip radiographs. At the same time, this performance is not uniform across all anatomical locations and injury types. Subtle, multiple, and radiographically challenging fractures, as well as cases assessed in conditions different from those represented in the training data, remain the greatest challenges for automated fracture detection systems.

3.2.2 AI as a Clinical Decision Support Tool

Although automated fracture detection represents the basic application of AI, in clinical practice the use of these systems as tools supporting physicians in radiograph interpretation may be even more important. In this model, AI does not replace the clinician, but acts as a second reader, highlighting suspicious regions and supporting diagnostic decision-making [10, 17].

In addition to evaluating the overall performance of a deep learning model, Lindsey et al. investigated its impact on the accuracy of radiograph interpretation by 40 emergency physicians. Without AI assistance, physicians achieved a sensitivity of 80.8% and a specificity of 87.5%. The use of AI support increased these values to 91.5% and 93.9%, respectively, corresponding to a 47% relative reduction in misinterpretations.

A similar pattern of findings was reported by Duron, Guermazi, and Bachmann, who evaluated AI as a support tool for radiograph interpretation across different groups of clinicians. Duron et al. demonstrated an increase in sensitivity of 8.7% and specificity of 4.1% among a study group consisting of six radiologists and six emergency physicians. In the study by Guermazi et al., the use of AI increased the sensitivity of fracture detection from 64.8% to 75.2% without reducing specificity. Moreover, AI assistance did not prolong the interpretation process; on the contrary, the mean radiograph reading time was shortened. Bachmann et al. reported an increase in sensitivity from 72% to 80% and a 29% reduction in the number of missed fractures. In their study, they compared the performance of radiograph interpretation across different anatomical regions among physicians of various specialties, as well as nurses, both without and with AI support.

A particularly important aspect of AI as an assistive tool is its potentially greater usefulness among clinicians with less experience in interpreting musculoskeletal radiographs. In the study by Anderson et al., the results showed that clinicians with limited experience had a higher proportion of missed fractures than experienced radiologists when working without AI assistance: 20% versus 14%, respectively. With the support of a deep learning model, the proportion of missed fractures not only improved in both groups, but also became more similar, reaching 9% among clinicians and 10% among radiologists. Duron et al. also reported a greater improvement in sensitivity among emergency physicians than among radiologists, at 13% versus 4.3%. Similarly, Bachmann et al. noted that the largest reduction in the proportion of missed fractures occurred among emergency medicine residents and orthopaedic residents, with 33% and 35% fewer missed fractures, respectively, whereas the smallest improvement was observed among radiologists, at 21%.

The findings of individual studies therefore suggest that the greatest value of AI lies not in its standalone performance, but in its use as a tool supporting physicians' interpretation. The consistency of these observations was confirmed in the meta-analysis by Qin et al., which collectively assessed the impact of AI assistance on clinicians' performance in fracture detection on radiographs. Their analysis included 26 studies comprising 9,218 radiographs and 124,900 readings. AI assistance increased clinicians' overall sensitivity from 77% to 87%, specificity from 88% to 92%, and AUC from 0.90 to 0.95. At the same time, interpretation time did not change significantly.

3.2.3 Fracture Classification and Characterization

Beyond the assessment of radiographs for the presence or absence of a fracture, AI may also be used for fracture classification and characterization. In clinical practice, simply confirming the presence of a fracture is often insufficient to determine the appropriate therapeutic management. Further treatment depends on several factors, including fracture type, degree of displacement, anatomical location, involvement of the articular surface, and the presence of associated findings, such as dislocation or pre-existing implants.

Chung et al. evaluated their CNN-based model not only for the detection of proximal humeral fractures, but also for their classification into clinically relevant subtypes. The model analysed anteroposterior shoulder radiographs and distinguished normal images from four fracture types: greater tuberosity, surgical neck, three-part, and four-part fractures. The CNN demonstrated high classification performance, particularly for complex three-part and four-part fractures, with AUC values of 0.903 and 0.938, respectively. Its performance exceeded that of most clinicians; however, when compared with orthopaedic surgeons specializing in shoulder surgery, the difference was no longer statistically significant.

The clinical relevance of fracture classification is particularly evident in hip fractures. Krogue et al. emphasized that the choice of implant and surgical strategy largely depends on fracture type. The model developed by the authors classified hip and pelvic radiographs into several categories, including displaced and non-displaced femoral neck fractures, intertrochanteric fractures, and the presence of previous fixation or arthroplasty. In this multiclass classification task, the model achieved an accuracy of 90.4%, meaning that in most cases it correctly assigned the image to the appropriate diagnostic category. However, it should be emphasized that despite the high overall performance, the model achieved a markedly lower sensitivity for non-displaced femoral neck fractures, which are more difficult to detect and remain challenging even for experienced clinicians.

3.2.4 Workflow Optimization and Triage

Beyond improving diagnostic performance, an important area of AI application in fracture diagnosis is workflow optimization. In this context, AI may function as a triage tool by automatically flagging examinations with suspected fractures, prioritizing them for earlier review, and supporting workflow organization in high-pressure emergency department and radiology department settings.

One practical aspect of using AI as a tool supporting radiograph interpretation is its impact on physicians' reading time. In the emergency department setting, even a slight increase in the time required to assess a single examination could limit the usefulness of such systems, particularly when large numbers of radiographs are performed and rapid clinical decision-making is required. Therefore, it is important that the improvement in diagnostic performance observed in studies of AI models is not accompanied by longer image interpretation time; ideally, it should be associated with a reduction in reading time. In the study by Guermazi et al., the use of AI not only increased the sensitivity of fracture detection, but also shortened the mean radiograph interpretation time. A similar trend was observed by Duron et al., in whose study the mean reading time decreased from 67 to 57 seconds, although this difference did not reach statistical significance. In turn, the meta-analysis by Qin et al. showed that AI assistance improved clinicians' diagnostic performance without a statistically significant effect on interpretation time.

Herpe et al. evaluated the implementation of an AI-based fracture detection system in a real-world emergency department setting using a before-and-after design. After implementation of the AI-based system, a reduction in the rate of diagnostic discrepancies was observed, from 6.6% to 5.1%. However, it should be emphasized that no change was found in the frequency of clinically significant discrepancies. The median length of stay in the emergency department also decreased slightly, from 190 minutes to 181 minutes. No significant reduction in the rate of specialist consultation requests was observed.

Taken together, the available studies evaluated AI-based tools across four main domains: standalone fracture detection, AI-assisted clinician interpretation, fracture classification and characterization, and workflow optimization. Reported outcomes included diagnostic accuracy, clinician performance with and without AI support, classification of specific fracture types, reading time, discrepancy rates, and emergency department length of stay. These findings provide the basis for discussing the clinical utility, limitations, and implementation challenges of AI in radiographic fracture diagnosis.

4. Discussion

The findings summarized in this review indicate that artificial intelligence has substantial potential in radiographic fracture diagnosis. Across the included studies, AI-based systems achieved high diagnostic performance in automated fracture detection, particularly in anatomically focused tasks and selected imaging datasets. However, the reviewed evidence also suggests that the most clinically relevant role of AI may not be fully autonomous diagnosis, but rather its use as an adjunctive tool supporting clinicians during radiograph interpretation. Studies evaluating AI-assisted reading generally reported improvements in clinician performance, especially among less experienced readers or non-radiologist clinicians. In addition, emerging evidence shows that AI may contribute to fracture classification, characterization, and workflow support, although these applications remain less extensively validated than binary fracture detection. These findings provide the basis for discussing the clinical utility, generalizability, implementation challenges, and limitations of AI in fracture diagnosis.

One of the most important conclusions emerging from the available evidence is the high performance of AI models in detecting fractures on plain radiographs [15, 17]. In many analyses, algorithms achieved very good AUC, sensitivity, and specificity values, particularly in tasks limited to specific anatomical locations, such as wrist, hip, proximal humeral, or pelvic fractures [6, 7, 16, 18]. These findings indicate that deep learning models are capable of recognizing radiographic features of fractures and may achieve performance comparable to that of clinicians under selected study conditions [17]. However, it should be emphasized that most of the best-documented models have been evaluated in relatively well-defined diagnostic tasks, which limits the extent to which these results can be directly transferred to the full spectrum of trauma radiographs encountered in everyday clinical practice.

Therefore, the high performance of anatomically targeted models should not be equated with the existence of a single universal system capable of analysing all trauma radiographs with comparable accuracy. In routine clinical practice, emergency physicians interpret examinations involving different anatomical regions, acquired in various projections and in heterogeneous patient populations, with variable image quality and the presence of factors that may hinder interpretation, such as implants, degenerative changes, or atypical injury mechanisms. Meanwhile, many of the best-described AI models have been optimized for a single, strictly defined diagnostic task, such as detecting fractures of the wrist, hip, proximal humerus, or pelvis [6, 7, 16, 18]. This means that an algorithm achieving excellent results in one anatomical region may not maintain comparable performance in other areas of the musculoskeletal system or across the full spectrum of cases encountered in clinical practice.

The performance of an AI model depends to a large extent on the similarity between the test data and the data on which the model was trained. An important issue is dataset shift, which occurs when a model is applied to data that differ from those used during training. In clinical practice, this may involve implementing an algorithm in a hospital that uses different X-ray equipment, imaging protocols, image quality standards, or serves a patient population with a different injury profile. Under such conditions, model performance may be lower than that reported in the original validation study, highlighting the limited generalizability of current AI systems [15, 17]. For this reason, external validation and multicenter studies are essential, as they allow algorithms to be assessed in more diverse and realistic clinical settings [15, 17, 20].

In addition to the problem of generalization across different institutions, variation in model performance within specific subgroups of cases is also important. For example, in the study by Lindsey et al., the model achieved a very high AUC for wrist fracture detection; however, the highest performance was observed in the subset of images for which there was no disagreement between experts regarding the diagnosis. This suggests that AI performance may be particularly high in more straightforward cases, whereas diagnostically uncertain cases remain a greater challenge. A similar issue can be observed in other studies, where high AUC values reported for the overall test set may have masked weaker algorithm performance in clinically important subcategories. Subtle, non-displaced, multiple, and complex fractures remain particularly challenging, as these are the cases in which the risk of diagnostic error is highest [1, 14, 16, 20].

Current evidence suggests that the greatest value of AI in fracture diagnosis lies not in replacing physicians, but in supporting their interpretation as a second reader or a clinical decision support tool [17, 24]. In this model, AI does not replace clinical reasoning, but may direct the physician's attention to suspicious areas on the radiograph and reduce the risk of overlooking important, often subtle, fracture features [9, 12]. Particularly important is the observation that the benefit of AI support may be most evident among users with less experience in interpreting musculoskeletal radiographs. In several studies, performance improvements were especially pronounced among non-radiologists, frontline clinicians, or less experienced readers. This is highly relevant in the emergency department setting, where the initial interpretation of a radiograph often precedes the final radiology report [2, 4, 9]. However, it should be emphasized that physician–AI collaboration is safest when the algorithm's output is treated as additional information rather than as a final diagnostic decision. The system should support the physician's diagnostic vigilance, but not replace it [10, 17].

Beyond fracture detection itself, the ability of AI to classify and characterize fractures may become increasingly important. This is clinically relevant because therapeutic decisions depend not only on the presence of a fracture, but also on its type and location. AI models may therefore go beyond simple fracture/no-fracture classification and recognize more detailed injury categories. However, classification is a more complex task than detection, and good overall results may mask weaker performance in rare or clinically important subtypes [6, 7, 16].

Although retrospective studies and reader studies provide important data on the diagnostic performance of AI, they do not fully answer the question of its real-world clinical utility. In practice, implementing an algorithm for fracture diagnosis requires not only high sensitivity and specificity, but also effective integration into the everyday workflow of emergency departments and radiology units. An AI system must provide its output at the right time, in a clear format, and without overburdening clinicians with excessive alerts or additional required actions. For this reason, the technical performance of a model should be distinguished from its actual clinical effectiveness [13,17].

Previous studies suggest that AI-assisted interpretation does not necessarily prolong reading time and, in some analyses, may even shorten it [9, 12, 24]. However, evidence from real-world implementations remains limited. Herpe et al. showed that after AI implementation, the overall rate of diagnostic discrepancies decreased and the length of stay in the emergency department was slightly shortened, but there was no significant reduction in clinically significant discrepancies. Therefore, the ultimate value of AI will depend not only on model performance metrics, but also on integration with clinicians' everyday work, the way alerts are presented, user acceptance, and its impact on clinically relevant endpoints.

Methodological Limitations of the Current Evidence

Despite promising results, the current literature on AI in fracture diagnosis has important methodological limitations. Many studies are retrospective and rely on selected radiograph datasets that do not always reflect the full spectrum of cases encountered in everyday emergency department practice. Fracture prevalence, image quality, the number of diagnostically challenging cases, and the availability of clinical information may differ from real-world conditions, which may affect the assessment of algorithm performance.

Another limitation is the insufficient number of external validations and multicenter studies. Models trained and tested on data from a single institution or on similar image datasets may perform worse after implementation in a different clinical environment. This limits the direct generalizability of study findings and highlights the need to evaluate AI models in diverse, real-world clinical settings.

Moreover, most studies focus on diagnostic performance metrics such as AUC, sensitivity, and specificity, whereas fewer data are available on real-world endpoints relevant to patients and healthcare organizations. The number of prospective studies assessing the impact of AI on time to diagnosis, therapeutic decisions, emergency department revisits, complications, costs, and patients' functional outcomes remains limited. Without such data, it is difficult to determine whether improved radiograph interpretation translates into the full clinical value of AI systems.

Future Perspectives

The future development of AI should move toward more universal models capable of analysing different anatomical locations and injury types. It will be particularly important for these models to maintain high performance not only in typical and straightforward cases, but also in clinically challenging subgroups, such as subtle, non-displaced, multiple, and complex fractures.

Another important direction should be the development of multicenter studies and external validations conducted in diverse clinical settings. Models need to be evaluated on data originating from different hospitals, X-ray systems, patient populations, and imaging protocols. Only such studies will make it possible to determine whether the performance observed in selected test datasets is maintained after implementation in everyday clinical practice.

However, the most important step will be to assess the real-world impact of AI on patient care. Future studies should go beyond traditional diagnostic metrics and analyse outcomes such as time to diagnosis, therapeutic decisions, the number of missed fractures, emergency department revisits, costs, and patients' functional outcomes. At the same time, safe principles for integrating AI into clinical workflow will need to be developed, so that algorithms support physicians' decisions without replacing their clinical judgment.

5. Conclusions

Artificial intelligence represents a promising direction in radiographic fracture diagnosis. Current evidence indicates that AI-based systems can achieve high diagnostic performance in selected fracture detection tasks and may improve clinicians' interpretation when used as decision-support tools. The greatest clinical value of AI currently appears to lie not in replacing physicians, but in supporting their diagnostic process, particularly in emergency settings and among less experienced readers.

However, the widespread implementation of AI in fracture diagnosis still requires further development and validation. Important challenges include limited generalizability, variable performance in subtle or complex fractures, insufficient external validation, and the need for more prospective real-world studies assessing clinically meaningful outcomes. Future research should therefore focus on robust multicenter validation, integration with clinical workflow, and evaluation of AI's impact on patient care rather than diagnostic accuracy alone.

Overall, AI should be regarded as a highly promising but still evolving tool in radiographic fracture diagnosis. Its development should not be abandoned, but rather guided by careful validation, clinical oversight, and a clear understanding of its role as an adjunct to human expertise.

Author Contributions: All authors have read and agreed to the submitted version of the manuscript.

Funding Statement: This study received no external funding.

Conflict of Interest Statement: The authors declare no conflict of interest.

Declaration of the Use of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this manuscript, the authors used AI to improve language, readability, organization, and formatting of the text. The tool was not used to replace the authors' scientific judgment, interpretation of the literature, or responsibility for the content. Following its use, the authors thoroughly reviewed and edited the manuscript and accept full responsibility for the final version.

REFERENCES

1. Al Hajaj, S. W., Soliman, K., Zafar, M., Garnham, C., Al Hajaj, D., Elshafie, O., Alsswah, A., & Elwan, M. H. (2026). From subtle breaks to missed diagnoses: Real-world evaluation of an artificial intelligence fracture detection tool. *World Journal of Orthopedics*, *17*(4). https://doi.org/10.5312/wjo.v17.i4.113710
2. Anderson, P. G., Baum, G. L., Keathley, N., Sicular, S., Venkatesh, S., Sharma, A., Daluiski, A., Potter, H., Hotchkiss, R., Lindsey, R. V., & Jones, R. M. (2023). Deep learning assistance closes the accuracy gap in fracture detection across clinician types. *Clinical Orthopaedics & Related Research*, *481*(3), 580–588. https://doi.org/10.1097/CORR.0000000000002385
3. Avanzo, M., Stancanello, J., Pirrone, G., Drigo, A., & Retico, A. (2024). The evolution of artificial intelligence in medical imaging: From computer science to machine and deep learning. *Cancers*, *16*(21), 3702. https://doi.org/10.3390/cancers16213702
4. Bachmann, R., Gunes, G., Hangaard, S., Nexmann, A., Lisouski, P., Boesen, M., Lundemann, M., & Baginski, S. G. (2023). Improving traumatic fracture detection on radiographs with artificial intelligence support: A multi-reader study. *BJR|Open*, *6*(1), tzae011. https://doi.org/10.1093/bjro/tzae011
5. Chen, H.-Y., Hsu, B. W.-Y., Yin, Y.-K., Lin, F.-H., Yang, T.-H., Yang, R.-S., Lee, C.-K., & Tseng, V. S. (2021). Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. *PLOS ONE*, *16*(1), e0245992. https://doi.org/10.1371/journal.pone.0245992
6. Cheng, C.-T., Wang, Y., Chen, H.-W., Hsiao, P.-M., Yeh, C.-N., Hsieh, C.-H., Miao, S., Xiao, J., Liao, C.-H., & Lu, L. (2021). A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. *Nature Communications*, *12*(1), 1066. https://doi.org/10.1038/s41467-021-21311-3
7. Chung, S. W., Han, S. S., Lee, J. W., Oh, K.-S., Kim, N. R., Yoon, J. P., Kim, J. Y., Moon, S. H., Kwon, J., Lee, H.-J., Noh, Y.-M., & Kim, Y. (2018). Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthopaedica*, *89*(4), 468–473. https://doi.org/10.1080/17453674.2018.1453714
8. Currie, G., Hawk, K. E., Rohren, E., Vial, A., & Klein, R. (2019). Machine learning and deep learning in medical imaging: Intelligent imaging. *Journal of Medical Imaging and Radiation Sciences*, *50*(4), 477–487. https://doi.org/10.1016/j.jmir.2019.09.005
9. Duron, L., Ducarouge, A., Gillibert, A., Lainé, J., Allouche, C., Cherel, N., Zhang, Z., Nitche, N., Lacave, E., Pourchot, A., Felter, A., Lassalle, L., Regnard, N.-E., & Feydy, A. (2021). Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: A multicenter cross-sectional diagnostic study. *Radiology*, *300*(1), 120–129. https://doi.org/10.1148/radiol.2021203886

10. Elbahi, M. K., Muhammed, A., Fadlelmola Abdalla Mohamednour, M., & Mukhtar, F. S. (2025). Artificial intelligence in fracture diagnosis on radiographs: Evidence, pitfalls, and pathways for clinical integration (2020–2025). *Cureus*. https://doi.org/10.7759/cureus.93124
11. Fernholm, R., Pukk Härenstam, K., Wachtler, C., Nilsson, G. H., Holzmann, M. J., & Carlsson, A. C. (2019). Diagnostic errors reported in primary healthcare and emergency departments: A retrospective and descriptive cohort study of 4830 reported cases of preventable harm in sweden. *European Journal of General Practice*, *25*(3), 128–135. https://doi.org/10.1080/13814788.2019.1625886
12. Guermazi, A., Tannoury, C., Kompel, A. J., Murakami, A. M., Ducarouge, A., Gillibert, A., Li, X., Tournier, A., Lahoud, Y., Jarraya, M., Lacave, E., Rahimi, H., Pourchot, A., Parisien, R. L., Merritt, A. C., Comeau, D., Regnard, N.-E., & Hayashi, D. (2022). Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*, *302*(3), 627–636. https://doi.org/10.1148/radiol.210937
13. Herpe, G., Nelken, H., Vendeuvre, T., Guenezan, J., Giraud, C., Mimoz, O., Feydy, A., Tasu, J.-P., & Guillevin, R. (2024). Effectiveness of an artificial intelligence software for limb radiographic fracture recognition in an emergency department. *Journal of Clinical Medicine*, *13*(18), 5575. https://doi.org/10.3390/jcm13185575
14. Huhtanen, J. T., Nyman, M., Blanco Sequeiros, R., Koskinen, S. K., Pudas, T. K., Kajander, S., Niemi, P., Aronen, H. J., & Hirvonen, J. (2025). Comparative accuracy of two commercial AI algorithms for musculoskeletal trauma detection in emergency radiographs. *Emergency Radiology*, *32*(4), 569–580. https://doi.org/10.1007/s10140-025-02353-2
15. Jung, J., Dai, J., Liu, B., & Wu, Q. (2024). Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. *PLOS Digital Health*, *3*(1), e0000438. https://doi.org/10.1371/journal.pdig.0000438
16. Krogue, J. D., Cheng, K. V., Hwang, K. M., Toogood, P., Meinberg, E. G., Geiger, E. J., Zaid, M., McGill, K. C., Patel, R., Sohn, J. H., Wright, A., Darger, B. F., Padrez, K. A., Ozhinsky, E., Majumdar, S., & Pedoia, V. (2020). Automatic hip fracture identification and functional subclassification with deep learning. *Radiology: Artificial Intelligence*, *2*(2), e190023. https://doi.org/10.1148/ryai.2020190023
17. Kuo, R. Y. L., Harrison, C., Curran, T.-A., Jones, B., Freethy, A., Cussons, D., Stewart, M., Collins, G. S., & Furniss, D. (2022). Artificial intelligence in fracture detection: A systematic review and meta-analysis. *Radiology*, *304*(1), 50–62. https://doi.org/10.1148/radiol.211785
18. Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., & Potter, H. (2018). Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, *115*(45), 11591–11596. https://doi.org/10.1073/pnas.1806905115
19. Liu, P., Lu, L., Zhang, J., Huo, T., Liu, S., & Ye, Z. (2021). Application of artificial intelligence in medicine: An overview. *Current Medical Science*, *41*(6), 1105–1115. https://doi.org/10.1007/s11596-021-2474-3
20. Luiken, I., Lemke, T., Komenda, A., Marka, A. W., Kim, S. H., Graf, M. M., Ziegelmayer, S., Weller, D., Mertens, C. J., Bressem, K. K., Makowski, M. R., Adams, L. C., Prucker, P., & Busch, F. (2025). Evaluation of commercial AI algorithms for the detection of fractures, effusions, and dislocations on real-world clinical data: A prospective registry study. *Radiography*, *31*(6), 103189. https://doi.org/10.1016/j.radi.2025.103189
21. Newman-Toker, D. E., Peterson, S. M., Badihian, S., Hassoon, A., Nassery, N., Parizadeh, D., Wilson, L. M., Jia, Y., Omron, R., Tharmarajah, S., Guerin, L., Bastani, P. B., Fracica, E. A., Kotwal, S., & Robinson, K. A. (2022). *Diagnostic errors in the emergency department: A systematic review*. Agency for Healthcare Research and Quality (AHRQ). https://doi.org/10.23970/AHRQEPCCER258
22. Pinto, A., Berritto, D., Russo, A., Riccitiello, F., Caruso, M., Belfiore, M. P., Papapietro, V. R., Carotti, M., Pinto, F., Giovagnoni, A., Romano, L., & Grassi, R. (2018). Traumatic fractures in adults: Missed diagnosis on plain radiographs in the emergency department. *Acta Biomedica Atenei Parmensis*, *89*(1-S), 111–123. https://doi.org/10.23750/abm.v89i1-S.7015
23. Pinto, A., Reginelli, A., Pinto, F., Lo Re, G., Midiri, F., Muzj, C., Romano, L., & Brunese, L. (2016). Errors in imaging patients in the emergency setting. *The British Journal of Radiology*, *89*(1061), 20150914. https://doi.org/10.1259/bjr.20150914
24. Qin, H., Ding, Y., Ju, J., Qu, Z., & Peng, L. (2026). Enhanced fracture detection on radiographs with AI assistance for clinicians: A systematic review and meta-analysis. *Annals of Medicine*, *58*(1), 2610079. https://doi.org/10.1080/07853890.2025.2610079
25. Somville, F., Van Bogaert, P., Wellens, B., De Cauwer, H., & Franck, E. (2024). Work stress and burnout among emergency physicians: A systematic review of last 10 years of research. *Acta Clinica Belgica*, *79*(1), 52–61. https://doi.org/10.1080/17843286.2023.2273611
26. Wan, Z., Tang, J., Bai, X., Cao, Y., Zhang, D., Su, T., Zhou, Y., Qiao, L., Shen, K., Wang, L., Tian, X., & Wang, J. (2023). Burnout among radiology residents: A systematic review and meta-analysis. *European Radiology*, *34*(2), 1399–1407. https://doi.org/10.1007/s00330-023-09986-2
27. Zha, N., Patlas, M. N., & Duszak, R. (2019). Radiologist burnout is not just isolated to the United States: Perspectives from Canada. *Journal of the American College of Radiology*, *16*(1), 121–123. https://doi.org/10.1016/j.jacr.2018.07.010