



Artificial Intelligence for Fracture Detection: Current Methods, Diagnostic Performance, and Clinical Use

Nishant

M.S. Ortho, D-Ortho, F.A.G.E., Assistant Professor in Orthopaedic Surgery RDGM Medical College, Muzaffarpur Bihar India; Email: nishantspine@gmail.com, ORCID: 0000-0002-6206-7732

**Received:-26/06/2026, Revised:- 30/07/2026, Accepted:- 17/08/2026,
Published:- 29/08/2026**

Abstract

Fractures are still overlooked on initial imaging, particularly when findings are subtle or when radiographs are interpreted before specialist review. Deep-learning systems have therefore been developed to support fracture recognition on radiographs and computed tomography, with a smaller body of work addressing magnetic resonance imaging. This narrative review examines the principal model types, the datasets used for training and validation, and the performance reported across common anatomical regions. It also considers how these systems are being introduced into clinical practice. On curated test sets, many algorithms achieve diagnostic accuracy close to that of clinicians. The more clinically persuasive evidence, however, comes from assisted-reading studies, in which software improves sensitivity while largely preserving specificity. Results remain difficult to compare because datasets, reference standards, case prevalence, and validation methods vary widely. External validation, demographic fairness, automation bias, workflow design, cost, and professional accountability therefore remain central concerns. At present, fracture-detection AI is best regarded as decision support for triage or concurrent interpretation rather than an independent diagnostic service.

Keywords: artificial intelligence, fracture detection, deep learning, radiography, computed tomography, clinical decision support

1. Introduction

Plain radiographs make up a large share of acute musculoskeletal imaging. In many emergency departments, the first interpretation is made by an emergency physician or another non-radiologist, especially overnight. Small cortical breaks, minimally displaced injuries, overlapping anatomy, and immature or osteoporotic bone can all make fractures difficult to recognize. Scaphoid fractures, rib fractures, vertebral compression fractures, and pediatric buckle fractures are familiar examples. When these injuries are missed, treatment may be delayed and patients may require repeat imaging or unplanned follow-up.

Computer-assisted interpretation is intended to reduce this gap. A fracture-detection program can analyze an examination soon after acquisition, mark a suspicious site, or raise the priority of the study on a reporting list. This role is narrower than autonomous diagnosis: the clinician remains responsible for relating the imaging result to the history, examination, and other investigations. The distinction matters because most regulatory clearances describe these products as detection, triage, or decision-support devices [25].

Research in this area has expanded from single-site experiments to multicenter evaluations and commercial deployment. Algorithms now address injuries of the wrist, hip, ribs, spine, face, and pediatric skeleton. This review describes the technical approaches behind those systems, summarizes reported performance, and considers the practical conditions required for safe clinical use.

2 Technical Foundations

2.1 Model development

Early computer-aided detection relied on features selected by the developer, such as cortical edges, texture measurements, or changes in bone contour. A conventional classifier then used those measurements to distinguish normal from abnormal images. Such pipelines were sensitive to positioning, anatomy, and image quality because the chosen features did not transfer well across the full range of fracture appearances.

Convolutional neural networks changed the task by learning image features during training. Classification models, including ResNet, DenseNet, EfficientNet, and Inception families, return a probability that a fracture is present within an image or predefined region [2,8,9]. Detection architectures such as Faster R-CNN, RetinaNet, and the YOLO family add spatial information by drawing a box around a

suspected lesion [1,9,12]. U-Net derivatives provide pixel-level segmentation and may be useful when the fracture line or fragment outline must be defined [14,15]. More recent studies have used vision transformers or combined transformer and convolutional components to model relationships across a larger field of view [13,30].

2.2 Training data and reference standards

A model reflects the data on which it was trained. Dataset size is important, but the range of patient ages, injury patterns, imaging equipment, projections, and institutions is equally consequential. Labels are usually based on review by one or more experienced radiologists. Some studies use consensus reading or computed tomography to resolve uncertain cases [6]. Differences in these reference standards can alter both the apparent prevalence of fractures and the measured accuracy of the algorithm.

Public resources have made independent research easier. MURA contains a large collection of musculoskeletal radiographs [21]; GRAZPEDWRI-DX focuses on pediatric wrist trauma [19]; FracAtlas includes labels for classification, localization, and segmentation [22]; and RibFrac supports detection and instance segmentation on chest computed tomography [14,15]. Commercial developers also use proprietary collections that may be considerably larger. When those datasets and labeling procedures are not described in detail, reproducibility and subgroup assessment become difficult [25].

2.3 Outputs and interpretability

A probability score alone offers limited help at the point of care. Many systems therefore display a bounding box, contour, or heatmap that directs the reader to the suspected abnormality [1,12]. Other outputs include confidence estimates and flags for worklist prioritization. These visual cues can improve efficiency, but they should not be mistaken for a complete explanation of model reasoning. Saliency maps may emphasize an irrelevant structure even when the final classification is correct, and a convincing overlay does not establish that the underlying decision is reliable.

3 Performance by Anatomical Region

Published performance differs by anatomy, modality, study design, and case mix. The

ranges in Table 1 summarize commonly reported results and should not be read as direct head-to-head comparisons. Curated datasets often contain clearer images and a

higher fracture prevalence than routine practice, both of which can raise apparent performance.

Table 1 Representative performance reported for fracture detection

Anatomical region	Modality	Reported performance	Clinical context	Sources
Wrist and distal radius	Radiography	Sensitivity about 90-96%; specificity about 90-94%; AUC 0.94-0.98	One of the most extensively studied applications, including pediatric buckle fractures	[1,3,4,8-10]
Hip and proximal femur	Radiography	Sensitivity about 91-98%; AUC about 0.90-0.98	High consequence of an occult injury and delayed treatment	[11-13,27,28]
Ribs	CT	Sensitivity gains of about 10-20 percentage points for subtle fractures in assisted-reading studies	Large CT series are time-consuming to review manually	[14-16]
Vertebral compression	Radiography CT and DXA	AUC about 0.85-0.95	Potential role in opportunistic osteoporosis assessment	[17,18]
Pediatric skeleton	Radiography	Sensitivity about 88-96%	Variable normal development makes interpretation difficult for non-specialists	[19,28]
Skull and face	CT or radiography	Sensitivity about 85-95%	Performance varies with anatomy and fracture type	[26]

Across these studies, stand-alone algorithms frequently approach the average performance of radiologists on selected test sets [6,17]. Assisted interpretation is the more relevant comparison for practice. Reader studies commonly report absolute sensitivity improvements of roughly 5-15 percentage points when clinicians can see the algorithm output, with little change in specificity and, in some settings, shorter reading times [16,23,24,28]. The size of that benefit depends on reader experience, fracture prevalence, and how the output is presented.

4 Clinical Integration and Regulation

Fracture-detection products have obtained clearance in several jurisdictions, usually as aids to detection or triage rather than as autonomous readers. OsteoDetect received US De Novo authorization in 2018 for distal-radius

fracture detection. BoneView later obtained European and US clearances, including an extension to pediatric use. NICE guidance has also examined the use of fracture-detection software in urgent and emergency care [25]. Product indications and regulatory status can change, so institutions should verify the current authorization for the intended patient group and imaging pathway before deployment.

Three patterns of integration are used most often:

- Worklist triage places examinations with a suspected fracture earlier in the reporting queue. The algorithm changes priority, not the final interpretation.
- Concurrent reading shows the software result while the clinician reviews the images. This is the usual design in studies that assess AI as a second reader [4,6,16,24].

- Point-of-care support gives emergency or other non-radiologist clinicians an additional check before the formal report is available, particularly where continuous radiology coverage is not feasible.

Retrospective validation remains more common than prospective implementation research. Reported gains also tend to be smaller after software enters routine care, where image quality, disease prevalence, reader behavior, and incomplete follow-up differ from a benchmark dataset [16,25]. Evaluation should therefore continue after implementation. Useful measures include false-positive burden, correction of initial interpretations, reporting delay, time to treatment, repeat attendance, and performance across demographic and clinical subgroups.

5 Limitations and Risks

External validity is the first concern. Performance may fall when a model encounters images from a new hospital, a different detector, an unfamiliar acquisition protocol, or a population unlike the training cohort. Multicenter testing reduces this uncertainty, but external validation is still inconsistent [10,17,30]. A local silent trial before clinical activation can reveal problems that are not visible in the vendor's aggregate metrics.

The difficult cases also matter more than the average score. Nondisplaced injuries, stress fractures, severe osteoporosis, postoperative anatomy, and pediatric growth plates remain challenging [6,18]. A system that performs well on obvious adult fractures may still fail in the situations for which clinicians most want support. Institutions should review error patterns by fracture type rather than relying only on a single sensitivity or AUC.

Bias may enter through case selection, labeling, or underrepresentation of particular ages, bone densities, ethnic groups, and patients with implants [27]. Human factors create a separate risk. Repeated agreement with the software can encourage automation bias, whereas frequent false alerts may lead users to ignore the system. Display design, training, audit, and clear escalation rules are therefore part of safety, not merely implementation details.

Responsibility for an incorrect assisted interpretation is not settled uniformly across jurisdictions. The clinician, institution, and vendor may each control different parts of the pathway. Contracts and governance policies should define software updates, incident

reporting, downtime procedures, and the handling of disagreement between the reader and the model. Cost and technical integration also remain substantial barriers, particularly where picture archiving and radiology information systems are fragmented.

6 Research Priorities

Future work should move beyond narrow detection tasks. Multitask systems could identify a fracture, define its extent, assign a clinically relevant classification, and estimate healing on follow-up imaging [15,29]. Such outputs will be useful only if each component is validated against an appropriate reference standard and if uncertainty is communicated clearly.

Larger and more diverse training cohorts are also needed. Federated learning may allow institutions to contribute to model development without transferring images to a central repository, although it does not by itself correct inconsistent labels or sampling bias. Foundation models trained on broad musculoskeletal datasets may reduce the amount of annotation required for uncommon injuries [30], but adaptation and external testing remain necessary before clinical use.

The most informative trials will be prospective and embedded in routine care. They should measure consequences that matter to patients and services: delayed treatment, complications after a missed injury, additional imaging, reporting time, length of stay, and resource use. Research in settings with limited specialist coverage is particularly important because the potential value is high, but infrastructure and maintenance constraints differ from those in well-resourced academic centers.

7 Conclusion

Fracture-detection AI has progressed from small experimental studies to regulated products used in clinical imaging pathways [25]. The evidence is strongest when the software supports a clinician by prioritizing an examination or marking a possible injury, rather than attempting to replace expert interpretation [16,23,24]. Reported accuracy is encouraging, but safe adoption depends on external and local validation, attention to subgroup performance, careful interface design, and continuing audit after deployment. Future progress should be judged less by small gains on benchmark datasets and more by whether the technology reduces missed injuries and improves timely patient care.

References

1. Radiology Artificial Intelligence. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence*. 2019. <https://pubs.rsna.org/doi/full/10.1148/ryai.2019180001>
2. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*. 2018;73(5):439-445. Discussed in Deep learning in fracture detection: a narrative review. PMC7144272.
3. Musculoskeletal radiologist-level performance using deep learning for detection of scaphoid fractures on conventional multiview hand and wrist radiographs. *European Radiology*. 2023. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9935716/>
4. Duron L, Ducarouge A, Gillibert A, et al. Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study. *Radiology*. 2021;300(1):120-129.
5. Guermazi A, et al. AI assistance and improvement in fracture-detection sensitivity among residents and junior radiologists. Summarized in *Artificial Intelligence in Fracture Diagnosis on Radiographs: Evidence, Pitfalls, and Pathways for Clinical Integration 2020-2025*. PMC12551796.
6. Jacques T, Cardot N, Ventre J, et al. Commercially available AI algorithm improves radiologists' sensitivity for wrist and hand fracture detection on X-ray compared with a CT-based ground truth. *European Radiology*. 2024;34(5):2885-2894.
7. Dell'Aria A, Tack D, Saddiki N, et al. Radiographic detection of post-traumatic bone fractures: contribution of artificial intelligence software to the analysis of senior and junior radiologists. *Journal of the Belgian Society of Radiology*. 2024;108(1):44.
8. Detecting distal radial fractures from wrist radiographs using a deep convolutional neural network with an accuracy comparable to hand orthopedic surgeons. *Journal of Imaging Informatics in Medicine*. 2021. <https://link.springer.com/article/10.1007/s10278-021-00519-1>
9. Deep learning assisted diagnosis system: improving the diagnostic accuracy of distal radius fractures. 2023. PMC10471443.
10. Critical evaluation of deep neural networks for wrist fracture detection. arXiv:2012.02577.
11. Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. *Radiology: Artificial Intelligence*. 2020;2(2):e190023.
12. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. PMC6717182.
13. Proximal femur fracture detection on plain radiography via feature pyramid networks. PMC11130146.
14. Jin L, et al. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: development and validation of FracNet. *eBioMedicine*. 2020. [https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964\(20\)30482-5/fulltext](https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(20)30482-5/fulltext)
15. Yang J, et al. Deep rib fracture instance segmentation and classification from CT in the RibFrac Challenge. arXiv:2402.09372. 2024.
16. AI-assisted radiologists versus standard double reading for rib fracture detection on CT images: a real-world clinical study. *PLOS ONE*. 2025. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0316732>
17. Deep learning for opportunistic vertebral fracture detection on routine thoraco-abdominal computed tomography: a systematic review and hierarchical summary receiver operating characteristic meta-analysis of patient-level diagnostic test accuracy. *PubMed* 42373383.
18. Automatic AI tool for opportunistic screening of vertebral compression fractures on chest frontal radiographs: a multicenter study. *Bone*. 2024. *PubMed* 39549901.
19. Nagy E, Janisch M, Hrzic F, Sorantin E, Tschauner S. A pediatric wrist trauma X-ray dataset GRAZPEDWRI-DX for machine learning. *Scientific Data*. 2022;9:222.
20. Deep learning in scaphoid fracture detection and healing prediction: a systematic review of artificial intelligence applications in hand surgery. 2025. PMC12658349.
21. Rajpurkar P, Irvin J, et al. MURA: large dataset for abnormality detection in musculoskeletal radiographs.

- arXiv:1712.06957. 2017.
22. Abedeen I, et al. FracAtlas: a dataset for fracture classification, localization, and segmentation of musculoskeletal radiographs. *Scientific Data*. 2023;10:521.
23. Artificial intelligence as a simultaneous second reader in diagnostic radiology: an umbrella review of systematic reviews and meta-analyses. *PubMed* 41850944. 2026.
24. Improving traumatic fracture detection on radiographs with artificial intelligence support: a multi-reader study. *BJR Open*. 2023;6(1):tzae011.
25. Artificial Intelligence in Fracture Diagnosis on Radiographs: Evidence, Pitfalls, and Pathways for Clinical Integration 2020-2025. *PMC12551796*. arXiv:2602.03076.
26. Diagnosis of nasal bone fractures on plain radiographs using convolutional neural networks. *PMC9747951*.
27. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*. 2019. <https://www.nature.com/articles/s41746-019-0105-1>
28. AI-Assisted Fracture Detection in Orthopedic and Trauma Imaging: Where It Works, Where It Fails, and Principles for Safe Clinical Deployment. *PMC13206237*.
29. Deep learning to detect triangular fibrocartilage complex injury in wrist MRI: retrospective study with internal and external validation. 2022. *PMC9322609*.
30. A generalizable large-scale foundation model for musculoskeletal radiographs. arXiv:2602.03076.