



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	DEEP LEARNING-ASSISTED FRACTURE DETECTION ON PLAIN RADIOGRAPHS: DIAGNOSTIC ACCURACY, CLINICAL INTEGRATION, AND FUTURE DIRECTIONS
----------------------	--

DOI	https://doi.org/10.31435/ijitss.3(51).2026.6149
------------	---

RECEIVED	26 June 2026
-----------------	--------------

ACCEPTED	16 September 2026
-----------------	-------------------

PUBLISHED	21 September 2026
------------------	-------------------

LICENSE



The article is licensed under a **Creative Commons Attribution 4.0 International License**.

© The author(s) 2026.

This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

DEEP LEARNING–ASSISTED FRACTURE DETECTION ON PLAIN RADIOGRAPHS: DIAGNOSTIC ACCURACY, CLINICAL INTEGRATION, AND FUTURE DIRECTIONS

Monika Szlachta-Gubernat (Corresponding Author, Email: monika-gubernat@o2.pl)
Hospital of the Ministry of Interior and Administration in Rzeszów, Rzeszów, Poland
ORCID ID: 0009-0005-6994-9761

Nicol Baran
Medical Center in Łańcut, Łańcut, Poland
ORCID ID: 0009-0003-3607-6744

Marta Bajkowska-Piterak
Medical Center in Łańcut, Łańcut, Poland
ORCID ID: 0009-0000-6689-0788

Adrianna Buż
John Paul II Municipal Hospital in Rzeszów, Rzeszów, Poland
ORCID ID: 0009-0001-8202-5628

Przemysław Piterak
Hospital of the Ministry of Interior and Administration in Rzeszów, Rzeszów, Poland
ORCID ID: 0009-0002-3553-7353

Damian Danilczuk
John Paul II Municipal Hospital in Rzeszów, Rzeszów, Poland
ORCID ID: 0009-0005-5755-1115

Wiktor Adamiec
John Paul II Municipal Hospital in Rzeszów, Rzeszów, Poland
ORCID ID: 0009-0002-3322-5479

Jerzy Buszko
Medical Care Center in Jarosław, Jarosław, Poland
ORCID ID: 0009-0008-2708-1222

Michał Dyś
Health Care Complex in Dębica, Dębica, Poland
ORCID ID: 0009-0000-2644-7626

ABSTRACT

Introduction and Purpose. Fractures are a major global health burden, and a non-trivial fraction of fractures presenting to emergency and urgent care settings are initially missed on plain radiographs, causing diagnostic delay and avoidable morbidity. This review appraises the current state of deep learning for fracture detection on conventional radiography, addressing standalone accuracy, the effect of algorithmic assistance on readers, and the conditions of clinical integration.

Brief Description of the State of Knowledge. Across recent meta-analyses, pooled sensitivity and specificity for fracture detection on radiographs cluster in the low-to-mid nineties, with summary AUROC approaching 0.97 in the largest synthesis. The literature is unevenly distributed across the skeleton: the wrist and the hip dominate the evidence base, the pediatric elbow is moderately developed, and other regions remain represented by only a handful of studies. Algorithmic assistance lifts reader sensitivity by several percentage points without a specificity penalty, with the largest gains in clinicians who have less musculoskeletal imaging experience.

Summary. Deep learning is mature for narrow fracture-detection tasks in the appendicular skeleton, but not yet for broad deployment claims. External validation across institutions and vendors, prospective evaluation tied to clinical endpoints, and standards-based integration into PACS and reporting infrastructure are the conditions that would move these tools from accuracy benchmarks toward measurable patient benefit. The clearest near-term value lies in decision support that narrows the diagnostic gap for less experienced readers.

KEYWORDS

Deep Learning, Convolutional Neural Networks, Fractures, Bone, Radiography, Diagnostic Imaging, Artificial Intelligence, Clinical Decision Support Systems

CITATION

Szlachta-Gubernat, M., Baran, N., Bajkowska-Piterak, M., Buż, A., Piterak, P., Danilczuk, D., Adamiec, W., Buszko, J., & Dyś, M. (2026). Deep learning–assisted fracture detection on plain radiographs: Diagnostic accuracy, clinical integration, and future directions. *International Journal of Innovative Technologies in Social Science*, 3(51). [https://doi.org/10.31435/ijitss.3\(51\).2026.6149](https://doi.org/10.31435/ijitss.3(51).2026.6149)

COPYRIGHT

© The author(s) 2026. This article is published as open access under the **Creative Commons Attribution 4.0 International License (CC BY 4.0)**, allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

1. Introduction

1.1. Epidemiology of Bone Fractures as a Public Health Concern

Fractures are a major and growing global burden. In the Global Burden of Disease (GBD) 2019 analysis there were 178 million new fractures and 455 million prevalent cases worldwide in that single year, and fractures accounted for 25.8 million years lived with disability; although age-standardized rates declined slightly, the absolute number of fractures rose substantially (1). A later GBD analysis reached the same qualitative conclusion, describing fractures as a major contributor to disease burden and predicting that they will rise significantly in the years to come (2). The risk is concentrated in older adults and, in particular, in women, especially for hip and wrist fractures, with hip fracture incidence increasing steeply with advancing age (3). The burden is also less well characterized in lower-resource settings, where a systematic review found that few studies had directly measured fracture incidence or prevalence despite widely reported clinical rates (4).

Missed fractures add a second layer of burden, since they delay treatment and may lead to long-term disability. In emergency and urgent care settings, the cost of a missed fracture is therefore not merely a delayed diagnosis but prolonged pain, immobility, and a measurable risk of preventable downstream morbidity (5). With population ageing, the share of fragility fractures is expected to grow, which sharpens the case for any technology that can reduce diagnostic miss at the front line of imaging.

1.2. The Role of Conventional Radiography in Diagnostic Algorithms

Plain radiography remains the front-line test for most suspected fractures because it is quick, portable, and embedded in standard imaging algorithms. In the ACR Appropriateness Criteria for acute traumatic hip pain, radiography is described as the initial imaging modality of choice and can be performed portably in the trauma bay; the same guideline favors a cross-table lateral view over a frog-leg lateral view because of the risk of fracture displacement (6). Within the broader musculoskeletal pathway, radiography is the low-burden

starting point, while CT and MRI are typically reserved for problem-solving or for selected scenarios in which plain films are non-diagnostic (7). Current guidelines also emphasize judicious imaging in order to avoid unnecessary radiation exposure, minimize cost, and preserve diagnostic precision (8), a logic that is reflected in pathways such as the Ottawa knee and ankle rules and in selective first-line MRI for suspected scaphoid fracture once clinical examination has been performed (9). A cost-minimization analysis in primary care further confirmed that local radiographic diagnosis and initial treatment cost substantially less than transport to and treatment in hospital, reinforcing the central position of plain films in the early management of musculoskeletal trauma (10).

The principal limitation of radiography is sensitivity, especially for occult or subtle fractures. Reviews repeatedly note concerns about the ability of radiographs to demonstrate certain fractures in the acute phase, and a systematic review of older patients sustaining low-energy falls found that plain radiography missed a substantial proportion of pelvic ring, rib, and thoracolumbar fractures (11). Technique and viewing conditions matter as well: a study using a trauma-meeting digital projector showed that subtle hip fractures were considerably less detectable on a non-diagnostic projector than on a validated monitor, so projection geometry and display quality can materially change what plain films reveal (12). These constraints set the practical horizon against which decision support tools must be evaluated.

1.3. Characterization of Diagnostic Errors in Musculoskeletal Radiology

Musculoskeletal radiology errors are usually grouped into two broad categories: perceptual errors, in which the relevant abnormality is not detected at all, and cognitive or interpretive errors, in which a finding is seen but its significance is misread. In practice, these two categories account for the bulk of missed diagnoses on plain films, with Bruno and colleagues separating perceptual from cognitive errors and Busby and colleagues distinguishing a miss from a misinterpretation (13, 14).

In emergency department workflows, reported discrepancy or miss rates cluster in the low single digits. Studies in pediatric and mixed ED populations have reported roughly 1.8% to 5.4% overall discrepancy and approximately 3% for radiograph discordance, while broader reviews of trauma radiography document higher rates in selected high-risk settings, ranging from 9% to 25% depending on reader experience and case mix (15–18). The most consistent predisposing factors are fatigue, circadian disruption, time pressure, and limited supervision. A Radiology editorial on diurnal effects noted that overnight radiologists made more errors than daytime readers and that fatigue impairs attention and decision making, and resident discrepancy studies have likewise shown higher error rates for overnight or after-hours work (19, 20). Technique and case complexity further compound this picture, because subtle fractures, overlapping bony structures, and image artefacts are disproportionately difficult to detect (14, 21).

1.4. Evolution of Decision Support Systems in Radiology

Computer-aided detection in radiology has shifted from feature-engineered pipelines to end-to-end deep neural networks over roughly a decade. Early systems relied on hand-crafted descriptors, while modern approaches learn task-relevant representations directly from labelled images, a transition that began in earnest after the 2012 ImageNet competition and accelerated with the widespread adoption of transfer learning. Vision transformers and hybrid convolutional–attention architectures have entered the medical-imaging literature in the period 2020–2023, complementing rather than displacing convolutional backbones (22, 23). The trajectory matters because it sets the technical baseline against which fracture-detection systems are now benchmarked: a contemporary fracture model is typically a deep convolutional network or transformer-based detector trained with substantial pretraining, evaluated against expert consensus, and assessed with curve-based as well as classification metrics.

1.5. Rationale and Aims of the Present Review

Although recent systematic reviews have synthesized pooled accuracy figures across heterogeneous tasks, an up-to-date narrative that simultaneously addresses technical foundations, anatomical performance, human–AI interaction, and the conditions of clinical integration remains comparatively scarce. The present review therefore aims to provide a critical appraisal of the diagnostic accuracy and translational maturity of deep learning algorithms for fracture detection on conventional radiography. Four guiding questions structure the synthesis: what is the diagnostic accuracy of contemporary models across distinct anatomical regions; to what extent does algorithmic assistance modify the performance of clinicians; what are the principal barriers to and facilitators of clinical integration; and which research directions remain insufficiently explored.

2. Theoretical and Methodological Foundations of Deep Learning

Deep learning in medical image analysis has evolved from hand-crafted feature pipelines to end-to-end convolutional neural networks (CNNs), and subsequently to hybrid and transformer-based systems. Recent reviews consistently place CNNs, and especially residual and dense variants, at the center of classification and detection work, while U-Net-style encoder–decoder designs remain the default for segmentation; DeepLab-style atrous convolution and newer hybrid CNN–Transformer designs are increasingly used when broader context or boundary precision matters (24–26).

2.1. Convolutional Neural Networks — the Dominant Paradigm

Reference convolutional architectures established the technical idiom on which most fracture-detection systems still build. EfficientNet showed that a balanced scaling of depth, width, and input resolution improves the accuracy–efficiency trade-off compared with arbitrary scaling of any single dimension (27), while Xception reframed the Inception module as a stack of depth wise separable convolutions, providing a clean, computationally efficient backbone that has been widely adopted in medical imaging (28). A short review of the 2012–2017 generation of architectures situates these developments within a broader trajectory of ImageNet-driven progress, in which depth, parameter count, and learned feature hierarchies were repeatedly traded against computational cost (29). For fracture detection, the practical consequence is that residual, dense, and Xception-style backbones dominate published pipelines, with successive layers of convolution, normalization, and pooling extracting features that are then fed into either a classification head or a downstream detection module.

2.2. Object Detection and Semantic Segmentation Models

When the clinical task requires not only an image-level decision but also the precise location of a lesion, the literature relies on two architectural families. Two-stage detectors, of which Faster R-CNN remains the canonical example, generate region proposals and then classify and refine them; the region proposal network introduced in that work continues to be the standard reference point for two-stage detection (30). One-stage detectors such as YOLO, RetinaNet, and SSD trade some accuracy for inference speed by predicting bounding boxes directly from feature maps. On the segmentation side, U-Net and its successors remain the default architectures in medical imaging, and a comprehensive review confirms the persistent success of U-Net-derived designs across modalities (31); DoubleU-Net is one example of a stacked U-Net variant that has been used to refine boundary delineation in challenging segmentation tasks (32). For fracture-related tasks these segmentation backbones can support fracture-line delineation, bone segmentation as a preprocessing step, or attention-guided detection within an anatomically restricted region of interest.

2.3. Transformer-Based Architectures in Medical Image Analysis

Transformer-based architectures bring a fundamentally different inductive bias to image analysis, replacing the local receptive field of convolution with self-attention. The Swin Transformer introduced a hierarchical, window-based formulation that scales efficiently to dense prediction tasks and now serves as a common backbone in medical imaging (33), while Focal Attention provides a complementary framing in which local fine-grained interactions are combined with coarser global context (34). Self-attention is particularly attractive when long-range spatial dependencies matter, as is the case in detecting fracture patterns that span large portions of a radiograph or that are defined by relationships between non-adjacent anatomical landmarks. A systematic comparison of vision transformers with CNNs in medical image analysis found that transformer-based models often perform competitively with, and in some settings better than, conventional CNNs, although pretraining is especially important for transformers and the results depend on dataset size and task type (22, 35).

2.4. Learning Paradigms

Learning strategy matters at least as much as architecture. Supervised learning remains the prevailing standard in fracture detection, but the broader medical-imaging literature points consistently to self-supervised and semi-supervised methods as the most important response to limited labels, and to ImageNet pretraining, domain-specific pretraining, and foundation models as the main routes to improved robustness and data efficiency (25, 26). MedSAM is a representative example of the foundation-model turn: trained on a very large corpus of image–mask pairs spanning multiple anatomies and modalities, it was reported to deliver gains in accuracy and robustness over modality-specific specialist models (36). For musculoskeletal radiography this matters because real-world datasets are often small, heterogeneous, and label-noisy; the practical question is usually how to adapt a pretrained model to a narrow anatomical target, not how to train a large model from scratch on every site. A large transfer-learning benchmark in medical imaging reinforces this conclusion, showing that ConvNets transferred better and were more annotation-efficient than vision transformers, while self-supervised and domain-adaptive pretraining improved transfer performance for both families (37).

2.5. Data Augmentation and Regularization Strategies

Data augmentation has become a routine component of fracture-detection pipelines and is no longer regarded as optional. A review focused on medical-imaging augmentation describes augmentation primarily as a robustness and regularization tool, mitigating the consequences of limited and unbalanced data rather than functioning merely as a way to enlarge the training set (38). Classical operations such as rotation, scaling, and reflection coexist with more sophisticated techniques including CutMix, MixUp, and generative-adversarial synthesis, while standard regularization devices, namely dropout, weight decay, and early stopping, are typically combined with augmentation rather than used as alternatives.

2.6. Evaluation Metrics for Diagnostic Performance

Evaluation should be tied to the task. Classification and detection studies usually report sensitivity, specificity, AUROC, and, for imbalanced detection problems, precision–recall or FROC-style curves, whereas segmentation studies rely more heavily on the Dice coefficient, intersection-over-union, and boundary-aware measures such as Hausdorff distance (24, 35). A segmentation guideline paper warns that metric choice and implementation can materially change conclusions, especially under class imbalance, and recommends standardized reporting rather than reliance on a single headline score (39). A systematic review and meta-analysis of diagnostic accuracy in medical imaging similarly highlights heterogeneity in methods and reference standards, which complicates direct comparisons of sensitivity, specificity, and AUROC across studies (40). In a fracture-detection paper this implies that AUROC alone is insufficient: a model can look strong on curve-based metrics yet still miss clinically important fractures, so sensitivity at the chosen operating point, false positives per image, and reader-setting or referral-standard validation are the numbers that matter most. Reference standards typically combine radiologist consensus, correlation with computed tomography or MRI, and, in some studies, clinical verification through follow-up.

Table 1. Overview of deep learning architectures commonly used in fracture detection on plain radiographs.

Architecture family	Representative models	Primary task	Strengths	Typical limitations
CNN (residual / dense / efficient)	ResNet, DenseNet, EfficientNet (27), Xception (28)	Image-level classification	Mature, computationally efficient, strong transfer learning from ImageNet (22, 29)	Limited receptive field; pretraining bias toward natural-image features (37)
Two-stage detectors	Faster R-CNN (30)	Object detection with region proposals	High localisation accuracy; well-suited to bounding-box fracture annotation	Slower inference; more complex training pipeline
One-stage detectors	YOLO, RetinaNet, SSD	Object detection	Real-time inference; appropriate for triage workflows	Reduced sensitivity for small or subtle lesions
Encoder–decoder (U-Net family)	U-Net, DoubleU-Net (32), nnU-Net	Semantic segmentation	Default architecture in medical imaging (31); precise boundary delineation	Memory-intensive at high input resolutions
Vision transformers	ViT, Swin Transformer (33), Focal Attention (34)	Classification and detection	Long-range spatial context; competitive with CNNs given sufficient pretraining (23, 35)	Data-hungry; performance dependent on pretraining strategy (37)
Foundation / multimodal	MedSAM (36), vision-language models	Universal segmentation; multimodal reasoning	Label efficiency through self-supervision; cross-domain transfer (25)	High compute cost; limited radiograph-specific validation

3. Review Methodology

3.1. Literature Search Strategy

PubMed/MEDLINE, Embase, Scopus, Web of Science, IEEE Xplore, and the arXiv preprint repository were searched for studies. PubMed and Google Scholar were the primary search venues, supplemented by hand-searching of references in retrieved articles and recent meta-analyses.

3.2. Eligibility Criteria

Inclusion was restricted to original studies that validated deep learning algorithms applied to conventional radiography and reported diagnostic performance metrics. Studies limited to CT or MRI modalities, conference abstracts without peer review, single-case reports, and editorial pieces were excluded.

3.3. Study Selection

Articles judged relevant to the four research questions stated in Section 1.5 were selected purposively, with priority given to recent meta-analyses, multi-centre studies, prospective workflow studies, and primary studies offering distinctive methodological or clinical insight.

3.4. Information Extracted from Included Studies

For each study discussed in this review, key information was extracted in a structured manner to support consistent narrative comparison. This included dataset characteristics (size, source, single- versus multi-center, pediatric versus adult, anatomical region), model architecture and training strategy, reported performance metrics (sensitivity, specificity, AUROC, and where available reader-with-AI versus reader-alone comparisons), the presence and design of external validation, and the clinical setting in which the study was conducted. Where studies reported workflow or human–AI interaction outcomes, additional information on reader experience level, deployment mode (safety net, triage, problem-solving, or autonomous), and operational endpoints (reading time, recall rate, length of stay) was recorded.

3.5. Methodological Quality Considerations

Because this article is presented as a narrative review rather than a systematic one, formal study-level risk-of-bias scoring with instruments such as QUADAS-2, QUADAS-AI, or PROBAST was not performed. The established reporting and quality frameworks for artificial intelligence in medical imaging — most notably the CLAIM checklist for AI-in-imaging studies and the TRIPOD-AI statement for prediction-model development and validation — were used as conceptual reference points rather than as scored instruments, informing which findings were emphasized and how strongly individual results were phrased.

Two recurrent methodological concerns shaped these judgements throughout the review. The first is data leakage between training and test partitions, which arises when patient-level splits are not enforced or when the same patient contributes images to multiple partitions, and which tends to inflate reported performance in ways that do not survive independent external validation. The second is spectrum bias arising from non-representative case mix, particularly the over-representation of unambiguous fractures and the under-representation of subtle, occult, or anatomically atypical presentations, which is the dominant failure mode observed when models trained on curated datasets are deployed in unselected clinical populations. Studies that did not enforce patient-level splits, that relied exclusively on internal validation, or that selected case mix in ways likely to inflate apparent performance, are flagged in the narrative wherever they are discussed.

External validation status was treated as the most informative single signal of evidence maturity: findings from single-center internal-validation studies are reported with appropriate caveats, while results supported by independent, multi-site external validation are given greater weight in the synthesis. Readers interested in formal, instrument-based risk-of-bias profiles for individual studies are referred to the recent systematic reviews and meta-analyses cited in Section 4 (in particular, the 2024 wrist and PLOS Digital Health meta-analyses and the 2025 pediatric updated systematic review), which apply QUADAS-2 and related instruments to overlapping subsets of this literature.

3.6. Synthesis Strategy

A narrative synthesis supported by tabular summaries was performed. Quantitative meta-analysis was not undertaken because of substantial methodological heterogeneity across the included studies, including variation in reference standards, anatomical scope, model architectures, and reporting conventions.

4. Diagnostic Performance by Anatomical Region

Deep learning studies have not been distributed evenly across the skeleton. A region-by-region synthesis must therefore acknowledge that the wrist and the hip dominate the literature, that the pediatric elbow is the next most visible niche, and that several other regions are represented by only a handful of studies. The headline accuracy figures should be read against this asymmetric coverage rather than as a uniform statement about fracture AI on radiographs (41-43).

4.1. Upper Extremity Fractures

The upper extremity, and the wrist in particular, is the best-represented anatomical region in the literature. A 2024 wrist meta-analysis reported pooled sensitivity and specificity of 92% and 93% across 33,026 radiographs, while a broader 2024 fracture meta-analysis spanning 100 studies found overall sensitivity and specificity around 91% and 92%, with radiologists generally more specific and AI leaning slightly more sensitive (41, 43, 44). The wrist meta-analysis was nevertheless built on only six included studies and exhibited substantial heterogeneity, with several studies judged at high risk of bias (41). Hand and phalangeal fractures remain comparatively thinly studied, which suggests that single-site performance figures for these subregions should be interpreted cautiously.

Elbow fracture work is notable less for raw volume than for the clinical problem it targets. A pediatric systematic review found that most eligible studies were radiograph-based, with the elbow the most common body part, but only two of nine studies performed external validation, so the apparent accuracy range of 88.8% to 97.9% should be read as promising but still institution-dependent (45, 46). The same pattern recurs in small-anatomy regions such as shoulder, clavicle, and hand: performance can look excellent in single-center datasets, yet the literature remains too sparse to support region-specific conclusions with the same confidence as for the wrist or hip — a setting in which transparent reporting against established AI imaging guidelines is particularly important to discipline claims that may otherwise rest on narrow datasets (45).

4.2. Lower Extremity Fractures

The lower extremity is dominated by the hip and femoral neck because the clinical cost of a miss is so high. In the 2024 PLOS Digital Health meta-analysis, hip fractures had the highest pooled sensitivity among fracture outcomes, and radiographs themselves had the best pooled sensitivity and specificity across modalities and data types; a 2021 external-validation study of paediatric fracture detection also showed strong performance on radiographs, and a 2025 femoral-neck study reported an AUC of 0.991 with external validation (47). The more important point for this region is not just accuracy but robustness: occult femoral-neck fractures are a recurring reason to escalate from negative radiographs to MRI or CT, and the literature on occult hip and femoral fractures consistently treats MRI as the best second-line test, with CT a practical alternative when speed or access matter (48). Femoral-shaft, periprosthetic, knee, and lower-leg fractures are represented but the studies are fewer and the evidence is more fragmented, so these subregions are better described as emerging rather than fully mature (42, 43, 49). Specific challenges include metallic artefact in periprosthetic imaging, tibial-plateau fractures with subtle articular involvement, avulsion injuries, and stress fractures whose radiographic signs may lag clinical presentation (49).

4.3. Axial Skeleton Fractures

For the axial skeleton, vertebral fractures stand out because they are both common and methodologically awkward. A 2026 CT-focused meta-analysis found lower sensitivity for spine fractures than for rib, skull, or pelvis fractures, and older fracture reviews have repeatedly noted that detection depends heavily on projection, reader experience, and the presence of subtle compression deformity rather than a clear cortical break (48). Rib fractures are another difficult region because oblique fracture lines may run parallel to the projection plane, and pelvic fractures are anatomically complex but clinically important in trauma. In the CT meta-analysis, pelvic fractures performed well, but that probably reflects the easier visibility of certain pelvic injuries on cross-sectional imaging rather than a general claim about plain films. Vertebral compression fractures, their correlation with osteoporosis, and the application of the Genant grading system therefore remain particularly difficult targets for radiography-based deep learning systems (50).

4.4. Particularities of the Pediatric Population

Pediatric fracture detection has its own failure modes. The pediatric systematic review noted above found that most studies were single-center, elbow-heavy, and only sparsely externally validated (45). This fits clinical intuition: growth plates, incomplete ossification, greenstick and torus (buckle) fractures, plastic deformation, and normal developmental variants create an error landscape that is materially different from that in adults. For this reason, pediatric datasets are best discussed separately rather than folded into adult performance summaries, and the Salter–Harris classification provides an essential anatomical framework for interpreting model output around the physical region (45, 46, 51).

4.5. Comparative Synthesis of Diagnostic Performance

Taken together, research maturity tracks both anatomy and clinical consequence. The wrist and the hip have the deepest evidence base; the pediatric elbow has a moderate but still limited base; and the hand, clavicle, shoulder, periprosthetic, and several pediatric subregions remain underrepresented (41, 42). This is the principal pattern that the literature supports from the present scan, and it implies that any future tabular synthesis crossing anatomical region with architecture, performance metrics, and the presence of external validation will reveal as much about the geography of the evidence as about the geography of the skeleton.

Table 2. Diagnostic performance of deep learning models for fracture detection on radiographs, by anatomical region.

Anatomical region	Evidence base	Reported performance	External validation	Translational maturity
Wrist	Largest evidence base (e.g. 33 026 radiographs in 2024 meta-analysis) (41)	Pooled sensitivity $\approx$ 92%, specificity $\approx$ 93% (41)	Limited; substantial between-study heterogeneity (41)	High for narrow detection task
Hip / femoral neck	Deep base; clinically high-impact target (42)	Sensitivity 94–99%; AUC up to 0.991 in femoral-neck study (42, 47)	Increasing; selected external studies available (48)	High
Pediatric elbow	Moderate; most-studied pediatric site (45)	Reported accuracy range 88.8–97.9% (45)	Only 2 of 9 studies externally validated (45)	Moderate; institution-dependent
Hand, phalanges, shoulder, clavicle	Sparse; predominantly single-center (41, 46)	Often $> 90\%$ on internal data; estimates unstable	Rare	Emerging
Vertebrae / axial skeleton	Sparse on plain films; methodologically awkward (51)	Lower sensitivity than appendicular sites (43)	Rare	Early
Pelvis and ribs	Limited on radiographs (more developed on CT) (42)	Mixed; performance depends on injury complexity	Rare	Early
Periprosthetic, lower-leg, stress fractures	Fragmented evidence base (42, 43, 48)	Few studies, narrow datasets, no pooled estimates	Rare	Early

5. Human–AI Comparative Performance

5.1. Standalone Algorithmic Performance

Standalone performance is generally high. The largest meta-analysis identified in the present search, covering 100 studies, reported pooled sensitivity of 91.43% and specificity of 92.12% with a summary AUROC of 0.968 (43), and a separate PLOS Digital Health meta-analysis found pooled radiograph performance of 94% sensitivity and 92% specificity, with hip fractures performing especially well (42). These summary estimates nevertheless come with the usual caveats: patient selection and reference standards remained major sources of bias, and the authors repeatedly emphasized that external validation is the weak point of the literature (42, 44).

5.2. AI-Assisted Reader Performance

The cleanest human–AI comparison in the present material is a 2022 multicenter multi-reader, multi-case (MRMC) study by Anderson and colleagues, involving 24 clinicians across 12 anatomic regions and conducted with an FDA-cleared deep learning system. With algorithmic assistance, diagnostic AUC rose from 0.90 to 0.94, sensitivity from 82% to 90%, and specificity from 89% to 92%; the benefit was larger for clinicians with limited musculoskeletal imaging training, in whom the miss rate fell from 20% unaided to 9% with assistance, nearly matching the performance of radiologists (52). A 2023 prospective workflow study (Oppenheimer et al.) with two residents showed a smaller but real gain: human sensitivity rose from 84.74% to 91.28% when AI output was presented, without a specificity penalty, and 33 of 35 post-AI diagnosis changes were correct (53).

Table 3. Selected human–AI reader studies for fracture detection on radiographs.

Study	Design	Readers and scope	Reader performance (unaided → AI-assisted)	Principal finding
Anderson et al. (52)	Multicenter MRMC, FDA-cleared system	24 clinicians, 12 anatomic regions	AUC 0.90 → 0.94; sensitivity 82% → 90%; specificity 89% → 92%	Largest absolute gain in clinicians with limited MSK training; miss rate fell from 20% to 9%, approaching radiologist performance
Prospective workflow study, 2023 (53)	Prospective resident study	Residents, mixed anatomic regions	Sensitivity 84.74% → 91.28%; no specificity penalty	33 of 35 post-AI diagnosis changes were correct
Herpe et al. (54)	ED operational study	ED clinicians, limb radiographs	Improved fracture recognition with AI assistance	Supports safety-net deployment in emergency settings
Chan et al., Singapore ED (55)	Operational implementation	Mixed ED clinicians, limb fractures	Reduced recall rates; time and manpower savings	Modest but consistent workflow benefits
Loeffen et al. (56)	Feasibility, four deployment modes	Radiology team, fracture diagnosis	Best results when used as safety net rather than triage / problem-solving	Triage and problem-solving modes increased false negatives or false positives; safety-net mode recommended

5.3. Modulation of Benefits According to Observer Experience

Algorithmic assistance does not deliver an equal benefit to every reader. The MRMC evidence is consistent: the largest absolute gains accrue to clinicians with less musculoskeletal imaging experience, and the practical effect is to narrow rather than to eliminate the diagnostic gap between non-radiologists and subspecialty radiologists (52). This pattern is clinically meaningful because much of the front-line interpretation of musculoskeletal radiographs is performed not by radiologists but by emergency physicians, orthopedic surgeons, and trainees, especially during overnight and weekend hours. The corollary is that the largest expected workflow benefit lies in settings where AI is positioned to support the less specialized reader rather than to refine the performance of those who are already operating near ceiling (57).

5.4. Reduction of Inter-Observer Variability

Formal kappa-based evidence on inter-observer variability under conditions with and without algorithmic assistance is comparatively limited in the present search, and direct estimates of Cohen's kappa or Fleiss's kappa with and without AI assistance for fracture detection on plain radiographs were not consistently surfaced. The available human–AI studies suggest that algorithmic input tends to bring less experienced readers closer to the level of subspecialty radiologists, which is a plausible mechanism for reduced inter-observer variability. Direct quantification of this effect, however, remains an open empirical question (52).

5.5. Impact on the Efficiency of the Diagnostic Process

Direct evidence of reduced interpretation time and reduced time-to-treatment intervals is also more limited than the headline accuracy figures might suggest. In the 24-reader study referenced above, AI assistance reduced average reading time by 7.1 seconds per examination, alongside the gain in fracture-detection accuracy (52). The available workflow evidence in emergency settings, discussed in Section 6, points to modest but real efficiency gains at the system level, including reductions in discrepancy rates and mean ED length of stay. Claims of substantial reductions in interpretation time or time-to-treatment beyond these examples should therefore be made cautiously and conditioned on local workflow factors (54).

5.6. Risks of Automation Bias and Deskilling

A separate strand of the human–AI literature is concerned with the behavioral risks of decision support. Reviews of human–AI interaction in radiology highlight automation bias, algorithmic aversion, deskilling, workload management, and burnout as recurring concerns, with specific vulnerabilities for trainees who may form their diagnostic habits in environments saturated with algorithmic prompts (58). These risks are not yet quantified in fracture-specific work in the present search, but the conceptual framing is clear: if AI assistance becomes a routine reader habit, the design of training programs, the structure of feedback loops, and the architecture of the user interface all become safety-relevant variables, not merely usability concerns.

6. Clinical Integration and Implementation

6.1. Deployment Models

The literature supports a physician-in-the-loop model more than full automation. A recent review describes integration strategies along the imaging pathway from acquisition and triage to interpretation, reporting, and teaching, outlining common interaction models and physician-in-the-loop workflows (58), and the best human-in-the-loop study in the present material reported that AI assistance improved diagnostic performance, with reader AUC rising from 0.90 to 0.94 (52). In practice, this makes triage and second-reader workflows the most defensible near-term deployments, while fully autonomous interpretation remains a narrower, higher-governance use case. A feasibility study published in 2025 found that AI worked best as a safety net, while triage and problem-solving modes worsened false negatives or false positives enough that the authors explicitly advised against those uses for now (56).

Table 4. Deployment models for AI-assisted fracture detection in the radiology workflow.

Deployment model	Position in workflow	Best-evidenced use case	Principal risk	Governance requirement
Safety net (second reader)	After the primary read (57, 58)	Catching missed fractures, particularly during overnight and weekend shifts (53, 58)	Automation bias when AI output is presented without uncertainty estimates (57)	Reader-time monitoring; drift surveillance; transparent display (60, 77)
Triage / worklist prioritization	Before the primary read (57)	Prioritizing examinations likely to contain a fracture	Possible increase in false negatives in low-prevalence flow (58)	Operating-point calibration; prospective evaluation (60)
Problem-solving (on demand)	Selectively, on equivocal cases (57)	Adjudicating radiographically subtle findings	False positives if applied beyond intended indications (58)	Explicit indication criteria; case-level audit
Autonomous interpretation	Replaces the primary read	Not currently supported for fracture detection on plain radiographs (58)	Erodes accountability; risk of deskilling among trainees (57)	Higher regulatory bar; not currently advised in the literature (57, 58)

6.2. Technical Integration Architecture

Standards-based interoperability is the main technical prerequisite. Tejani et al. state that standards-based interoperability facilitates AI integration with systems from different vendors into a single environment by enabling seamless exchange between information systems in the radiology workflow, and the same review emphasizes DICOM, HL7, and IHE profiles as the practical backbone for integration (59). A PACS-focused review adds that AI models should demonstrate their effectiveness across a wide range of scenarios and must be validated by prospective studies before deployment (60, 61), which is the right standard if a tool is intended to write back into PACS or downstream reporting systems. Deployment models in published implementations include on-premise installations, cloud-based services, and hybrid configurations, with the choice driven as much by local information-governance constraints as by technical performance.

6.3. Impact on Clinical Pathways in Emergency Departments

The workflow signal is real but modest. In one emergency-department implementation study, the discrepancy rate decreased by 17% after AI software deployment, and mean ED length of stay fell by 9 minutes, although the change in clinically relevant discrepancies was not statistically significant. A Singapore emergency-department study likewise concluded that AI-based fracture detection offers measurable potential for cost, manpower, and time savings (55). Taken together, these findings indicate a coherent, if modest, pattern in which AI support contributes to operational efficiency without yet delivering large changes in the clinically most consequential miss rates.

6.4. Commercially Available Solutions with Regulatory Certification

Detailed product-by-product claims about specific commercial fracture-detection systems are best made cautiously. A recent Radiology review notes as of early 2025 that the FDA has cleared more than 770 AI medical devices focused on radiology, primarily based on deep learning (61), although a reliable product-by-product registry for fracture-specific systems was not surfaced in the present search. Specific CE class and FDA pathway details, including 510(k) and De Novo classifications under the EU Medical Device Regulation and the FDA Good Machine Learning Practice framework, are therefore best handled as a separate tabular summary or supplement underpinned by primary regulatory sources (such as the FDA list of AI/ML-enabled medical devices), rather than as broad unqualified claims in the main text.

6.5. Cost-Effectiveness Analysis

The present evidence on cost-effectiveness reads more like early budget-impact work than mature QALY or ICER analysis. The clearest model identified projected net NHS savings and reduced downstream fracture callbacks, but the authors explicitly cautioned that additional evidence on radiograph-review accuracy and resource use is still needed (54). A fracture-specific paper anchored in QALYs or incremental cost-effectiveness ratios from the perspective of public or private payers was not identified in the present search, which means that any substantive cost-effectiveness subsection requires a more targeted search than the one underpinning this review.

6.6. Human and Organizational Factors

Acceptance is conditional rather than automatic. A systematic review of radiology stakeholders found that clinical accuracy is a condition for acceptance and that resistance is strongest toward fully automated tasks central to radiologists' core competencies (62, 63). A separate human–AI review adds concerns about automation bias, algorithmic aversion, deskilling, workload management, and burnout, with specific vulnerabilities for trainees (58), which is precisely where theoretical frameworks such as the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) belong, alongside discussions of training, change management, and patient-safety culture. The implementation argument, in short, is not only that the model performs well, but that teams can use it safely, consistently, and without eroding diagnostic skill.

7. Limitations, Systematic Biases, and Methodological Challenges

7.1. Representativeness and Quality of Training Datasets

Dataset representativeness is the central methodological constraint in fracture AI. The published literature remains skewed toward narrow institutional cohorts, and the reviews repeatedly flag unrepresentative training data, limited generalizability, and algorithmic bias as recurring barriers to deployment (24, 62, 64). In a large transfer-learning benchmark in medical image analysis, ConvNets were more transferable and more annotation-efficient than vision transformers, and transformer performance depended heavily on the pretraining strategy (37). In fracture-specific work, a pediatric review found that nearly all studies were single-institution and only two of nine had performed external validation (45). Underrepresentation of pediatric, geriatric, and ethnic minority populations is a persistent concern in the broader medical-imaging literature and is not, in general, resolved by larger architectures.

7.2. Algorithmic Bias

Demographic determinants of algorithmic performance, including age, sex, ethnicity, and socioeconomic status, are repeatedly highlighted in stakeholder and systematic reviews of clinical AI (62, 65). The evidence identified in the present search does not yet quantify demographic fairness for fracture detection specifically, so these variables are best framed as plausible but undermeasured sources of bias rather than as fully characterized determinants of performance. The implication for algorithmic fairness is that systematic, prospectively designed audits of subgroup performance should accompany deployment, rather than rely solely on retrospective post-hoc analyses of test-set fairness.

7.3. External Generalizability

External generalizability is where the limits of the literature become most visible in practice. Standards-based interoperability is presented as the practical route to moving AI across heterogeneous PACS and RIS environments, because custom integrations create maintenance burden and points of failure; the same literature emphasizes that clinical AI should be validated prospectively across diverse scenarios before deployment (59, 60, 66). Fracture-specific experiments isolating laterality markers or exposure settings as shortcut cues were not surfaced in the present search, but the technical-error literature is clear that imaging artefacts and acquisition pitfalls can distort interpretation and should be treated as plausible sources of spurious learning (67). Multi-center, multi-vendor validation across the heterogeneity of radiographic equipment manufacturers and acquisition parameters is therefore essential before any general claim of robustness can be sustained.

7.4. The Challenge of Rare Classes and Occult Fractures

Rare classes and occult fractures deserve their own paragraph, not a footnote. Medical-imaging reviews keep returning to class imbalance and data scarcity as core training problems, and fracture-specific pediatric work shows how normal anatomy, overlapping bones, and external artefact can be mistaken for fracture (68). This is the right place to discuss focal loss, cost-sensitive learning, and similar remedies if the section is to remain methodologically honest rather than speculative; in particular, the long-tail distribution of fracture types means that the most clinically consequential cases are often those for which the training signal is weakest.

7.5. Lack of Standardizations in Reporting

For reporting and governance, the literature is more directive than quantitative. One review explicitly recommends publishing criteria and guidelines to enhance standardization and reproducibility, while stakeholder reviews stress that clinical accuracy is only one condition for acceptance: AI literacy, transparency, regulation, and workflow fit matter as well, and the human-factor literature names automation bias, algorithmic aversion, deskilling, and trainee vulnerability directly (62, 69). In practical terms, insufficient adherence to the CLAIM, STARD-AI, and TRIPOD-AI guidelines compromises the reproducibility and comparability of findings, and a measurable share of the heterogeneity that prevents quantitative meta-analysis in this field is procedural rather than substantive.

7.6. Ethical and Legal Considerations

The principal ethical and legal obligations identified in the literature are validated deployment, transparent reporting, privacy-preserving data handling, and clear allocation of responsibility for errors; the exact legal framing depends on jurisdiction and device classification (70). Considerations of informed consent, data protection under the EU General Data Protection Regulation and the United States Health Insurance Portability and Accountability Act, and the status of medical devices under the EU Medical Device Regulation and the European Union AI Act all sit within this frame. These obligations are not opposed to clinical usefulness; rather, they specify the conditions under which clinical usefulness can be claimed. Because the present manuscript draws its primary evidence from the technical and clinical literature, readers seeking detailed regulatory guidance are directed to dedicated regulatory and legal commentaries on these instruments.

7.7. Risk of Overfitting and Technical Artefacts

The risk of overfitting and the role of technical artefacts deserve specific attention. The technical-error literature in musculoskeletal imaging emphasizes that laterality markers, projection orientation, exposure parameters, and characteristic acquisition artefacts can all distort interpretation (71, 72). Shortcut learning, in which a model learns to predict the target from spurious cues that happen to correlate with the label, is a recognized phenomenon in medical image analysis and is particularly insidious in datasets in which acquisition parameters are correlated with patient origin or clinical pathway. Combining standardized audits, careful train–test separation by institution and scanner, and explicit attention to potential shortcuts in the interpretation of saliency maps is the practical antidote suggested by the literature.

8. Future Directions

8.1. Multimodal Models

Multimodal models should be the first near-term extension of fracture AI. Recent reviews of medical foundation models describe multimodal systems as a way to combine images with radiology reports and clinical notes, and current applications of multimodal large language models in radiology already include preliminary report generation, visual question answering, and interactive diagnostic support (73, 74). The caveat is equally clear: a recent scoping review found that, despite strong niche performance, none of the available systems yet outperformed radiologists in diagnostic performance, so the field still needs clinically grounded pretraining and region-linked reasoning rather than scale alone (75). Vision-language pretraining holds particular promise for fracture detection because radiographs are routinely accompanied by structured and free-text clinical information that can support, and discipline, the visual model.

8.2. Foundation Models and Large-Scale Self-Supervised Learning

Foundation models and large-scale self-supervised learning are best framed as a response to label scarcity and portability, rather than as a generic upgrade. A systematic review describes multimodal foundation models as a paradigm shift toward label-efficient, holistic patient modelling, while a complementary review situates generative AI in radiology as a response to challenges of scalability, multimodal alignment, and data diversity (25, 76). For musculoskeletal radiography this points toward parameter-efficient fine-tuning, transfer learning, and adaptation to local acquisition protocols, rather than the retraining of large models from scratch at every site.

8.3. Federated Learning

Federated learning is the most plausible privacy-preserving route to multicenter fracture AI. A recent review states that federated learning is transforming healthcare by enabling collaborative AI model development across institutions while ensuring data privacy, but the same literature warns that post-deployment data shifts can degrade performance and that model parameters can still leak information (77, 78). The practical research agenda for fracture detection is therefore not only privacy but also communication efficiency, secure aggregation, and heterogeneity-aware calibration across vendors and institutions (79).

8.4. Explainable Artificial Intelligence

Explainable artificial intelligence should remain concrete. Reviews of XAI in medical imaging describe it as a means of making model decisions understandable to clinicians, with Grad-CAM, LIME, and SHAP still the most visible families of methods (80). In fracture detection the useful claim is not that explanations resolve trust on their own, but that they can help detect shortcut learning, demonstrate whether the model is attending to anatomy rather than to markers or overlay text, and support a structured review of failure modes. Saliency maps and feature-attribution techniques therefore function less as a regulatory endpoint and more as part of an ongoing safety toolkit.

8.5. Prospective Validation and Randomized Clinical Trials

Prospective validation should become the standard for any claim of clinical utility. A PACS-AI review states that AI models must demonstrate their effectiveness across a wide range of scenarios and must be validated by prospective studies, which is precisely the right bar for fracture workflows (60). The emergency-department implementation evidence already discussed is encouraging but modest: the discrepancy rate fell, yet clinically relevant discrepancies did not change significantly, even as length of stay improved. This argues for endpoints such as missed-fracture callbacks, time to treatment, expert-consultation burden, and ED throughput, rather than retrospective AUROC alone, and it supports a stronger role for randomized clinical trials of the kind already familiar in oncological imaging.

8.6. Continuous Learning and Post-Market Surveillance

Continuous learning and post-market surveillance belong in the core narrative rather than the appendix. Once a model is deployed, drift from new scanners, new patient mix, or new protocols can erode performance, which is why uncertainty quantification and ongoing monitoring matter (77). The same logic supports adaptive regulatory frameworks, but any such framework should remain measured: the requirement is for monitored updates, traceable changes, and pre-planned revalidation, rather than open-ended model drift in production. The FDA Predetermined Change Control Plan is one example of the regulatory thinking that would govern such a model lifecycle.

8.7. Standardizations and Regulation

Standardizations and regulation are the conditions that make all of the above deployable. The interoperability literature is explicit that standards-based interoperability facilitates AI integration with systems from different vendors into a single environment, and PACS-AI authors note that AI deployment raises significant legal and ethical issues around privacy, bias, safety, and regulatory compliance (59, 60). For fracture AI the future is therefore harmonized rather than isolated: standards-based PACS/RIS integration, multicenter validation, drift surveillance, and transparent reporting that lets hospitals compare systems across sites. The roles of the FDA, EMA, MDR, and the European Union AI Act sit at the intersection of these requirements, and the harmonization of international standards for the clinical deployment of AI will be central to whether the present generation of fracture-detection systems can be translated from accuracy benchmarks into routine clinical use.

9. Conclusions

The available evidence supports a measured rather than triumphant interpretation of deep learning for fracture detection on plain radiographs. The field is mature for narrow detection tasks, particularly in the appendicular skeleton, with pooled sensitivity and specificity in the low-to-mid nineties and summary AUROC values approaching 0.97 in the largest meta-analysis (42-44). It is not yet mature for broad deployment claims, because the literature still over-represents curated, retrospective, often single-center datasets, and external validation remains the weak point (42). The clearest near-term value lies in decision support that narrows the diagnostic gap for less experienced readers, with sensitivity gains in the order of several percentage points and proportionally larger reductions in miss rate for clinicians who do not interpret musculoskeletal radiographs daily (52, 53).

From the perspective of technological readiness, the most useful next step is no longer another accuracy paper. It is external validation in diverse sites, reader-impact studies tied to workflow outcomes, and better reporting of localization and failure modes (42, 56). For researchers this implies adherence to CLAIM, STARD-AI, and TRIPOD-AI in reporting, with explicit attention to subgroup performance, dataset shift, and shortcut learning. For clinicians and policymakers it implies that the impact of fracture AI will be determined as much by integration architecture, training programs, and post-market surveillance as by headline accuracy. Deep learning has changed what radiographic interpretation looks like at the technical level; whether it changes patient outcomes at the system level will depend on whether the deployment science catches up with the modelling science.

Author Contribution: Conceptualization, M.S-G., N.B., and M.B-P.; methodology, M.S.-G., P.P., D.D., and J.B.; literature search, A.B., M.D., and W.A.; data curation, M.S-G., P.P., and A.B.; writing—original draft preparation, M.S-G., N.B., M.B-P., P.P., D.D., J.B., A.B., M.D., and W.A.; writing—review and editing, M.S-G., N.B., M.B-P., P.P., D.D., J.B., A.B., M.D., and W.A.; visualization, N.B., D.D., and M.D.; supervision, M.B-P. and J.B.; project administration, W.A. All authors have read and agreed to the published version of the manuscript.

Funding Statement: This research received no external funding.

Institutional Review Board Statement: Not applicable. This is a narrative review of previously published literature and does not involve any new studies of human or animal subjects.

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

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this work, the authors used AI for the purpose of language editing. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the substantive content of the publication.

REFERENCES

1. Wu, A.-M., Bisignano, C., James, S. L., Abady, G. G., Abedi, A., Abu-Gharbieh, E., et al. (2021). 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. *The Lancet Healthy Longevity*, 2(9), e580–e92.
2. Yan, J., Li, F., Zhou, J., Ding, Y., Qin, Q., & Jin, C. (2025). The global burden of fractures and its underlying etiologies: Results from and further analysis of the Global Burden of Disease Study 2021. *Archives of Osteoporosis*, 20(1), 111.
3. De Laet, C. E., & Pols, H. A. (2000). Fractures in the elderly: Epidemiology and demography. *Baillieres Best Pract Res Clin Endocrinol Metab*, 14(2), 171–9.
4. Cordero, D. M., Miclau, T. A., Paul, A. V., Morshed, S., Miclau, T., 3rd, Martin, C., et al. (2020). The global burden of musculoskeletal injury in low and lower-middle income countries: A systematic literature review. *OTA Int*, 3(2), e062.
5. Pinto, A., Berritto, D., Russo, A., Riccitiello, F., Caruso, M., Belfiore, M. P., et al. (2018). Traumatic fractures in adults: Missed diagnosis on plain radiographs in the Emergency Department. *Acta Biomed*, 89(1-s), 111–23.
6. Ross, A. B., Lee, K. S., Chang, E. Y., Amini, B., Bussell, J. K., Gorbachova, T., et al. (2019). ACR appropriateness criteria® acute hip pain-suspected fracture. *J Am Coll Radiol*, 16(5s), S18–s25.
7. Aagesen, A. L., & Melek, M. (2013). Choosing the right diagnostic imaging modality in musculoskeletal diagnosis. *Prim Care*, 40(4), 849–61, viii.
8. Expert Panel on Musculoskeletal Imaging. (2020). ACR appropriateness criteria® acute trauma to the knee. *J Am Coll Radiol*, 17(5S), S12–S25.

9. National Institute for Health and Care Excellence. (2016). *Fractures (non-complex): Assessment and management* (NICE guideline [NG38]). NICE. <https://www.nice.org.uk/guidance/ng38>
10. Vabo, S., Kjerstad, E., Hunskaar, S., Steen, K., Brudvik, C., & Morken, T. (2023). Acute management of fractures in primary care - a cost minimisation analysis. *BMC Health Serv Res*, 23(1), 1291.
11. Pedersen, V., Lampart, A., Bingisser, R., & Nickel, C. H. (2020). Accuracy of plain radiography in detecting fractures in older individuals after low-energy falls: Current evidence. *Trauma Surg Acute Care Open*, 5(1), e000560.
12. Chellam, W. B. (2016). Missed subtle fractures on the trauma-meeting digital projector. *Injury*, 47(3), 674–6.
13. Bruno, M. A., Walker, E. A., & Abujudeh, H. H. (2015). Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6), 1668–76.
14. Busby, L. P., Courtier, J. L., & Glastonbury, C. M. (2018). Bias in radiology: The how and why of misses and misinterpretations. *Radiographics*, 38(1), 236–47.
15. Bisset, G. S., 3rd, & Crowe, J. (2014). Diagnostic errors in interpretation of pediatric musculoskeletal radiographs at common injury sites. *Pediatr Radiol*, 44(5), 552–7.
16. Taves, J., Skitch, S., & Valani, R. (2018). Determining the clinical significance of errors in pediatric radiograph interpretation between emergency physicians and radiologists. *Cjem*, 20(3), 420–4.
17. Ackermann, O., Wetter, A., Chelangattucherry, E., Emmanouilidis, I., & Rüländer, C. (2011). [Efficacy of x-ray assessment in emergency surgical departments: An evaluation in a level I trauma center]. *Unfallchirurg*, 114(1), 41–6.
18. Weber, M. A. (2025). Easily missed pathologies of the musculoskeletal system in the emergency radiology setting. *Rofo*, 197(3), 277–87.
19. Kung, J., Melenevsky, Y., Hochman, M., Didolkar, M., Yablon, C., Eisenberg, R., et al. (2013). On-call musculoskeletal radiographs: Discrepancy rates between radiology residents and musculoskeletal radiologists. *AJR American journal of roentgenology*, 200, 856–9.
20. Patel, A. G., Pizzitola, V. J., Johnson, C. D., Zhang, N., & Patel, M. D. (2020). Radiologists make more errors interpreting off-hours body CT studies during overnight assignments as compared with daytime assignments. *Radiology*, 297(2), 374–9.
21. Li, W., Stimec, J., Camp, M., Pusic, M., Herman, J., & Boutis, K. (2022). Pediatric musculoskeletal radiographs: Anatomy and fractures prone to diagnostic error among emergency physicians. *J Emerg Med*, 62(4), 524–33.
22. Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., et al. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems*, 48(1), 84.
23. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–44.
24. Zhang, X., Wang, J., Cheng, X., & Li, J. (2025). A review of deep learning-based medical image segmentation. *IET Image Processing*, 19(1), e70163.
25. Chen, Q., Liu, A., Zhang, J., Yang, C., & Zhang, Y. (2026). Foundation models in medical imaging: A review. *EngMedicine*, 3(2), 100123.
26. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med Image Anal*, 42, 60–88.
27. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks.
28. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. 1800–7 p.
29. Sarkar, D., Choudhury, P. G., Dutta, A., Sarkar, S., & Mandal, S. N. (2022). A brief journey of convolutional neural networks from 2012 to 2017. *Reason-A Technical Journal (Formerly Reason-A Technical Magazine)*, 20, 84–96.
30. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–49.
31. Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., et al. (2024). Medical image segmentation review: The success of U-Net. *IEEE Trans Pattern Anal Mach Intell*, 46(12), 10076–95.
32. Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., & Johansen, H. D. (Eds.). (2020). DoubleU-Net: A deep convolutional neural network for medical image segmentation. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 28–30 July 2020.
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. 9992–10002 p.
34. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., et al. (2021). Focal attention for long-range interactions in vision transformers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems* (Article 2297). Curran Associates Inc.
35. Aburass, S., Dorgham, O., Al Shaqsi, J., Abu Rumman, M., & Al-Kadi, O. (2025). Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases. *Journal of imaging informatics in medicine*, 38.
36. Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1), 654.

37. Hosseinzadeh Taher, M. R., Haghighi, F., Gotway, M. B., & Liang, J. (2025). Large-scale benchmarking and boosting transfer learning for medical image analysis. *Medical Image Analysis*, 102, 103487.
38. Garcea, F., Serra, A., Lamberti, F., & Morra, L. (2023). Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152, 106391.
39. Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15.
40. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., et al. (2021). Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *npj Digital Medicine*, 4(1), 65.
41. Hansen, V., Jensen, J., Kusk, M. W., Gerke, O., Tromborg, H. B., & Lysdahlgaard, S. (2024). Deep learning performance compared to healthcare experts in detecting wrist fractures from radiographs: A systematic review and meta-analysis. *European Journal of Radiology*, 174, 111399.
42. 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 Digit Health*, 3(1), e0000438.
43. Nowroozi, A., Salehi, M. A., Shobeiri, P., Agahi, S., Momtazmanesh, S., Kaviani, P., et al. (2024). Artificial intelligence diagnostic accuracy in fracture detection from plain radiographs and comparing it with clinicians: A systematic review and meta-analysis. *Clin Radiol*, 79(8), 579–88.
44. Yang, S., Yin, B., Cao, W., Feng, C., Fan, G., & He, S. (2020). Diagnostic accuracy of deep learning in orthopaedic fractures: A systematic review and meta-analysis. *Clinical Radiology*, 75(9), 713.e17–e28.
45. Shelmerdine, S. C., White, R. D., Liu, H., Arthurs, O. J., & Sebire, N. J. (2022). Artificial intelligence for radiological paediatric fracture assessment: A systematic review. *Insights into Imaging*, 13(1), 94.
46. Ashworth, E., Allan, E., Pauling, C., Laidlow-Singh, H., Arthurs, O. J., & Shelmerdine, S. C. (2025). Artificial intelligence (AI) in radiological paediatric fracture assessment: An updated systematic review. *Eur Radiol*, 35(9), 5264–86.
47. Wang, L., Zhang, C., Wang, Y., Yue, X., Liang, Y., & Sun, N. (2025). Development and validation of a deep learning system for the detection of nondisplaced femoral neck fractures. *Bioengineering*, 12, 466.
48. Husarek, J., Fuchss, A. M. C., Ruder, T. D., Mougiakakou, S., Exadaktylos, A., Wahedi, K., et al. (2026). Artificial intelligence for fracture detection on computed tomography: A comprehensive systematic review and meta-analysis of diagnostic test accuracy in non-commercial and commercial solutions. *Emergency Radiology*, 33(2), 351–75.
49. Wright, A. A., Hegedus, E. J., Lenchik, L., Kuhn, K. J., Santiago, L., & Smoliga, J. M. (2016). Diagnostic accuracy of various imaging modalities for suspected lower extremity stress fractures: A systematic review with evidence-based recommendations for clinical practice. *Am J Sports Med*, 44(1), 255–63.
50. Kulig, L., Malec, J. A., Wiercioch, E., Winiarska, Z., Woźniak, W. M., Zapalska, M., et al. (2026). Applications of artificial intelligence in opportunistic screening and diagnostic imaging of osteoporotic fractures: A systematic review. *Journal of Education, Health and Sport*, 89, 70382.
51. Meshaka, R., Pinto Dos Santos, D., Arthurs, O. J., Sebire, N. J., & Shelmerdine, S. C. (2022). Artificial intelligence reporting guidelines: What the pediatric radiologist needs to know. *Pediatr Radiol*, 52(11), 2101–10.
52. Anderson, P., Baum, G., Keathley, N., Sicular, S., Venkatesh, S., Sharma, A., et al. (2022). Deep learning assistance closes the accuracy gap in fracture detection across clinician types. *Clinical Orthopaedics & Related Research*, Publish Ahead of Print.
53. Oppenheimer, J., Lüken, S., Hamm, B., & Niehues, S. M. (2023). A prospective approach to integration of AI fracture detection software in radiographs into clinical workflow. *Life*, 13(1), [223 p.].
54. Herpe, G., Nelken, H., Vendeuvre, T., Guenezan, J., Giraud, C., Mimos, O., et al. (2024). Effectiveness of an artificial intelligence software for limb radiographic fracture recognition in an emergency department. *Journal of Clinical Medicine*, 13(18), [5575 p.].
55. Chan, H. Y., Tang, Y. P., Wong, Z. Y., Koh, S. H., Nickalls, O., Steven, W., et al. (2025). Artificial Intelligence (AI) in a Singaporean Emergency Department: Detecting fractures and reducing recalls. *Med J Malaysia*, 80(4), 462–5.
56. Loeffen, D. V., Zijta, F. M., Boymans, T. A., Wildberger, J. E., & Nijssen, E. C. (2025). AI for fracture diagnosis in clinical practice: Four approaches to systematic AI-implementation and their impact on AI-effectiveness. *European Journal of Radiology*, 187, 112113.
57. Lo Mastro, A., Grassi, E., Berritto, D., Russo, A., Reginelli, A., Guerra, E., et al. (2025). Artificial intelligence in fracture detection on radiographs: A literature review. *Japanese Journal of Radiology*, 43(4), 551–85.
58. Kocak, B., & Cuocolo, R. (2026). Human-AI interaction and collaboration in radiology: From conceptual frameworks to responsible implementation. *Diagn Interv Radiol*.
59. Tejani, A., Cook, T., Hussain, M., Schmidt, T., O'Donnell, K., & Arzen, S. (2024). Integrating and adopting AI in the radiology workflow: A primer for standards and integrating the healthcare enterprise (IHE) profiles. *Radiology*, 311.

60. Theriault-Lauzier, P., Corbin, D., Tastet, O., Langlais, E. L., Taji, B., Kang, G., et al. (2024). A responsible framework for applying artificial intelligence on medical images and signals at the point of care: The PACS-AI platform. *Canadian Journal of Cardiology*, 40(10), 1828–40.
61. Sosna, J., Joskowicz, L., & Saban, M. (2025). Navigating the AI landscape in medical imaging: A critical analysis of technologies, implementation, and implications. *Radiology*, 315(3), e240982.
62. Verwey, J., Zwart, B., M. I. J., Visser, J. J., & Sülz, S. (2026). Factors influencing AI acceptance in radiology: A systematic review across the radiology workflow. *Eur Radiol*, 36(5), 3687–96.
63. Jungmann, F., Jorg, T., Hahn, F., Pinto Dos Santos, D., Jungmann, S. M., Düber, C., et al. (2021). Attitudes toward artificial intelligence among radiologists, IT specialists, and industry. *Acad Radiol*, 28(6), 834–40.
64. Najjar, R. (2023). Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17), [2760 p.].
65. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A*, 117(23), 12592–4.
66. Bhandari, A. (2024). Revolutionizing radiology with artificial intelligence. *Cureus*, 16(10).
67. Yan, Y. Y., & Peh, W. C. G. (Eds.). (2025). *Technical errors and artifacts causing mistakes in musculoskeletal imaging*. Thieme Medical Publishers, Inc.
68. Nguyen, T., Maarek, R., Hermann, A.-L., Kammoun, A., Marchi, A., Khelifi-Touhami, M. R., et al. (2022). Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. *Pediatric Radiology*, 52(11), 2215–26.
69. Nair, A., Ong, W., Lee, A., Leow, N. W., Makmur, A., Ting, Y. H., et al. (2025). Enhancing radiologist productivity with artificial intelligence in magnetic resonance imaging (MRI): A narrative review. *Diagnostics*, 15(9), 1146.
70. Kutbi, M. (2024). Artificial intelligence-based applications for bone fracture detection using medical images: A systematic review. *Diagnostics*, 14(17), 1879.
71. Masciocchi, C., Conti, L., D’Orazio, F., Conchiglia, A., Lanni, G., & Barile, A. (2012). Errors in musculoskeletal MRI. In *Errors in radiology* (pp. 209–17). Springer.
72. Choi, Y. Y., Lee, J. S., & Yang, S.-O. (2017). Musculoskeletal lesions: Nuclear medicine imaging pitfalls. In *Pitfalls in musculoskeletal radiology* (pp. 951–76). Springer.
73. Nam, Y., Kim, D. Y., Kyung, S., Seo, J., Song, J. M., Kwon, J., et al. (2025). Multimodal large language models in medical imaging: Current state and future directions. *Korean J Radiol*, 26(10), 900–23.
74. Huang, S. C., Jensen, M., Yeung-Levy, S., Lungren, M. P., Poon, H., & Chaudhari, A. S. (2025). A systematic review and implementation guidelines of multimodal foundation models in medical imaging. *Res Sq*.
75. Seo, I., Bae, E., Jeon, J.-Y., Yoon, Y.-S., & Cha, J. (2024). The era of foundation models in medical imaging is approaching : a scoping review of the clinical value of large-scale generative AI applications in radiology.
76. Lee, J. G., Kyung, S., & Kim, N. (2026). Generative AI for developing foundation models in radiology and imaging: Engineering perspectives. *Biomed Eng Lett*, 16(1), 1–10.
77. Koutsoubis, N., Waqas, A., Yilmaz, Y., Ramachandran, R. P., Schabath, M. B., & Rasool, G. (2025). Privacy-preserving federated learning and uncertainty quantification in medical imaging. *Radiology: Artificial Intelligence*, 7(4), e240637.
78. Sun, M., Yang, Z., Huang, Y., Yu, H., Chen, Y., Qi, S., et al. (2025). Federated learning for large models in medical imaging: A comprehensive review. *arXiv preprint arXiv:250820414*.
79. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3(1), 119.
80. Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput Biol Med*, 140, 105111.