Authors	Title	Purpose	Study Design	Study Population	Methods	Limitations	Key Findings
Anttila et al. (2023)	Detecting Distal Radius Fractures Using a Segmentation-Based Deep Learning Model	To assess the performance of convolutional neural networks (CNNs) in detecting distal radial fractures (DRFs) on plain radiographs.	Retrospective Cohort Study	961 patients with wrist radiographs, 755 frontal wrist radiographs, 578 lateral wrist radiographs	Data from four institutions, including 503 patients with distal radial fractures and 289 without fractures, applied convolutional neural networks trained on frontal and lateral wrist radiographs to evaluate their performance in comparison with hand orthopedic surgeons for accurate fracture detection.	The limitations of the study include a binary classification without localization of the pathological region, a relatively small dataset of 1633 radiographs, a focus on adult wrist radiographs without evaluation for pediatric distal radial fractures, and the potential disadvantage of hand orthopedic surgeons judging small cropped images without additional clinical information.	The study demonstrated that convolutional neural networks (CNNs) exhibited high accuracy, sensitivity, and specificity in detecting distal radial fractures on plain radiographs, outperforming hand orthopedic surgeons and showing improved diagnostic ability when using both frontal and lateral views.
Aryasomayajula et al. (2023)	Developing an artificial intelligence diagnostic tool for pediatric distal radius fractures, a proof of concept study	To train an AI model to detect fractures in pediatric wrist radiographs. It is the first study of its kind using a specifically prepared dataset with pediatric wrist radiographs only.	Retrospective Cohort Study	5000 pediatric radiographs	Study utilized a dataset of 5000 pediatric wrist radiographs, collected retrospectively from 2014 to 2018, which were labeled as 'fracture' or 'no fracture' based on radiology text reports. A Convolutional Neural Network (CNN) was trained using transfer learning from VGG16 on this dataset, with data augmentation to enhance model generalization. The model was evaluated on a test dataset, achieving an 85% accuracy in detecting fractures.	Limitations of the study include the potential overfitting of the model, addressed by adding additional images through augmentation techniques. The dataset, though large, might not capture all real-life data variability.	The CNN, specifically VGG16 with data augmentation, demonstrated feasibility in diagnosing pediatric wrist fractures with an improved accuracy of 85%.

Ashkani-Esfahani et al. (2022)	Detection of ankle fractures using deep learning algorithms	Aimed to assess the performance of two different DCNNs in detecting ankle fractures using radiographs compared to the ground truth.	Retrospective Cohort Study	1050 patients with ankle radiographs	DCNNs were trained using radiographs obtained from 1050 patients with ankle fracture and the same number of individuals with otherwise healthy ankles. Inception V3 and Renet-50 pretrained models were used in our algorithms. Danis-Weber classification method was used. Out of 1050, 72 individuals were labeled as occult fractures as they were not detected in the primary radiographic assessment. Single-view (anteroposterior) radiographs was compared with 3-views (anteroposterior, mortise, lateral) for training the DCNNs.	The limitations include the inherent 2D nature of radiographs for a 3D ankle joint, suggesting that while the deep learning algorithm enhances 2D interpretation, 3D assessment through CT scans could surpass its accuracy; the reliance on data from a single center, indicating potential benefits from increased diversity in imaging sources and patient demographics in future studies; and the focused model training on distinguishing ankle fractures from normal images, highlighting the potential for a more comprehensive model with a larger dataset to differentiate various types of ankle fractures.	DCNNs showed that it can be used for developing the currently used image interpretation programs or as a separate assistant solution for the clinicians to detect ankle fractures faster and more precisely.
Blüthgen et al. (2023)	Detection and localization of distal radius fractures: Deep learning system versus radiologists	To evaluate a deep learning based image analysis software for the detection and localization of distal radius fractures.	Retrospective Cohort Study	524 wrist radiographs	A deep learning system (DLS) was trained on wrist radiographs. Performance was tested on internal (100 radiographs, 42 showing fractures) and external test sets (200 radiographs, 100 showing fractures). Single and combined views of the radiographs were shown to DLS and three readers. Readers were asked to indicate fracture location with regions of interest (ROI). The DLS yielded scores (range 0–1) and a heatmap.	Limitations include the absence of diagnostic details beyond the classification of "normal" or "abnormal" in the MURA dataset, potential selection bias favoring more easily detectable fractures in the external image set, indicated by the radiologists' slightly better performance on that dataset. Another limitation is the proprietary nature of the employed deep learning system, lacking disclosed technical background compared to publicly available architectures, though the utilized software demonstrated ease of use and effective predictive capabilities through image augmentation and systematic hyperparameter optimization.	The DLS was able to detect and localize wrist fractures with a performance comparable to radiologists, using only a small dataset for training.

Bousson et al. (2023)	Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms	To assess the performance of three commercially available artificial intelligence (AI) algorithms for detecting acute peripheral fractures on radiographs in daily emergency practice.	Retrospective Cohort Study	1210 patients admitted for skeletal trauma	Three AI algorithms—SmartUrgence, Rayvolve, and BoneView—were used to analyze 13 body regions. Four musculoskeletal radiologists determined the ground truth from radiographs. The diagnostic performance of the three AI algorithms was calculated at the level of the radiography set. Accuracies, sensitivities, and specificities for each algorithm and two-by-two comparisons between algorithms were obtained. Analyses were performed for the whole population and for subgroups of interest (sex, age, body region).	Limitations include a single-center dataset, potentially limiting generalizability due to the cosmopolitan population and large radiography set, with variability from multiple radiographers. The ground truth is considered weak compared to systematic CT evaluation, and the study lacks evaluation of AI algorithms' impact on emergency physician decision-making. Additionally, the study provides a snapshot of three AI algorithms at a specific time without longitudinal assessment.	The performance of AI detection of acute peripheral fractures in daily radiological practice in an emergency department was good to high and was related to the AI algorithm, patient age, and body region examined.
Bulstra et al. (2022)	A Machine Learning Algorithm to Estimate the Probability of a True Scaphoid Fracture After Wrist Trauma	To identify predictors of a true scaphoid fracture among patients with radial wrist pain following acute trauma, train 5 machine learning (ML) algorithms in predicting scaphoid fracture probability, and design a decision rule to initiate advanced imaging in high-risk patients.	Prospective Study	422 patients with radial wrist pain following wrist trauma were combined.	Predictors of a scaphoid fracture were identified among demographics, mechanism of injury and examination maneuvers. Five ML algorithms were trained in calculating scaphoid fracture probability. ML-algorithms were assessed on ability to discriminate between patients with and without a fracture (area under the receiver operating characteristic curve), agreement between observed and predicted probabilities (calibration), and overall performance (Brier score). The best performing ML-algorithm was incorporated into a probability calculator. A decision rule was proposed to initiate advanced imaging among patients with negative radiographs.	Limitations include the lack of a consistent reference standard for a true fracture in all patients.	ML-algorithm accurately calculated scaphoid fracture probability based on scaphoid pain on ulnar deviation, sex, age, and mechanism of injury. The ML-decision rule may reduce the number of patients undergoing advanced imaging by a third with a small risk of missing a fracture.

Canoni-Meynet et al. (2022)	Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow	To compare radiologists' performance without and with artifi-cial intelligence (AI) assistance for the detection of bone fractures from trauma emergencies.	Retrospective Cohort Study	500 patients	Three radiologists independently interpreted radiographs without AI assistance after a 1-month minimum washout period. The ground truth was determined by consensus reading between musculoskeletal radiologists and AI results. Patient-wise sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for fracture detection and reading time were compared between unassisted and AI-assisted readings of radiologists. Their performances were also assessed by receiver operating characteristic (ROC) curves.	Limitations of the study include potential order and recall bias as the same radiologists interpreted radiographs both with and without AI. Lack of clinical information introduced a context bias, and the study's context may have led to a Hawthorne effect. The study had fewer readers compared to other studies, impacting the representativeness of radiologists. The ground truth's dependence on readers and AI introduces bias in performance assessment, and assessing medico-economic impact was indirect and not optimally studied, requiring a challenging case-control study from an ethical standpoint.	AI-assisted radiologists work better and faster compared to unassisted radiologists.
Cheng et al. (2021)	A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs	Show a universal algorithm can detect most types of trauma-related radiographic findings on Pelvic radiograph	Retrospective Cohort Study	1888 patient pelvic radiographs	Trauma data from Chang Gung Memorial Hospital to develop PelviXNet, a deep learning algorithm trained on pelvic anteroposterior radiographs (PXRs). An independent clinical test set from ER data in 2017 evaluated the algorithm's performance. PelviXNet was compared with physicians in detecting fractures on a separate test set (PXR150). The algorithm utilized DenseNets, point supervision, and feature pyramid networks, with an ensemble method during inference. Statistical analyses included AUROC and AUPRC calculations to assess diagnostic performance.	Limitations of the study include insufficient training data, particularly for rare conditions like hip dislocation, potentially affecting algorithm performance. The retrospective nature of the study, conducted in a single institute, introduces bias, and the findings may not directly apply to other institutes with different population distributions. Selective bias may be present due to the random selection of images based on clinical diagnosis. Additionally, the study evaluated the algorithm and physicians based on radiographic findings alone, not accounting for clinical information used in real-world scenarios. The study emphasized the need for future prospective studies to assess the algorithm's benefits in a clinical environment with a focus on trauma-related detection.	PelviXNet demonstrates comparable performance with radiologists and orthopedics in detecting pelvic and hip fractures.

Cheng et al. (2020)	A Human-Algorithm Integration System for Hip Fracture Detection on Plain Radiography: System Development and Validation Study	To develop and validate a human-algorithm integration (HAI) system to improve the accuracy of hip fracture diagnosis in a real clinical environment.	Retrospective Cohort Study	3605 pelvic radiographs	The HAI system was developed using a deep learning algorithm trained on trauma registry data and 3605 PXRs from August 2008 to December 2016. 34 physicians were recruited to compare diagnostic performance before and after HAI system assistance using an independent testing dataset. Physicians' accuracy, sensitivity, specificity, and agreement with the algorithm were analyzed.	Limitations include potential selection bias due to defining the system as a Human-AI (HAI) tool, the inability to detect soft tissue changes, lack of integration with clinical information, a limited number of participating physicians, potential selection bias in the validation dataset from a single institute, and the preliminary nature of the study with a limited number of cases at three trauma centers.	Integrating this technology into emergency departments is feasible. The developed HAI system can enhance physicians' hip fracture diagnostic performance.
Cheng et al. (2019)	Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs	To identify the feasibility of using a deep convolutional neural network (DCNN) for the detection and localization of hip fractures on plain frontal pelvic radiographs (PXR).	Retrospective Cohort Study	25,505 limb radiographs	The accuracy, sensitivity, false-negative rate, and area under the receiver operating characteristic curve (AUC) were evaluated on 100 independent PXRs acquired during 2017. The authors also used the visualization algorithm gradient-weighted class activation mapping (Grad-CAM) to confirm the validity of the model.	Limitations of the study include the inherent challenge of understanding features learned by the deep convolutional neural networks (DCNNs), potential misactivations and difficulty in explaining them, uncertainty about the exact features used for fracture identification, and the exclusion of other pathological presentations on radiographs. The algorithm, trained specifically for discriminating between healthy bones and fractures, might not identify various lesions, introducing potential selective bias. Additionally, differences in age, gender, and Injury Severity Score (ISS) between fracture and nonfracture patients may contribute to selective bias.	DCNN not only detected hip fractures on PXRs with a low false-negative rate but also had high accuracy for localizing fracture lesions.

Cheng et al. (2023)	Evaluation of ensemble strategy on the development of multiple view ankle fracture detection algorithm	To identify the feasibility and efficiency of deep convolutional neural networks (DCNNs) in the detection of ankle fractures and to explore ensemble strategies that applied multiple projections of radiographs.	Retrospective Cohort Study	3102 ankle radiographs	DCNN was trained using a trauma image registry, then separately trained the DCNN on anteroposterior (AP) and lateral (Lat) AXRs. Different ensemble methods, such as “sum-up,” “severance-OR,” and “severance-Both,” were evaluated to incorporate the results of the model using different projections of view.	Limitations include the inherent uncertainty about features learned by the deep convolutional neural networks (DCNNs), the inability to fully understand the exact features used for fracture identification, and the potential exclusion of certain features not recognized by humans. The study's reliance on oblique views, not routinely used in the hospital, limits available data, and the algorithm, focused on discriminating between healthy and fractured bones in radiographs, may not identify other pathological presentations. Additionally, the study lacks the detection of other abnormalities in abdominal X-rays, crucial for routine diagnosis, and integrating the automatic detection algorithm into clinical pathways poses a challenge, necessitating further prospective studies for clinical impact validation.	Ankle fracture in the AXR could be identified by the trained DCNN algorithm.
Choi et al. (2023)	Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography	To develop a dual-input convolutional neural net-work (CNN)–based deep-learning algorithm that utilizes both anteroposterior (AP) and lateral elbow radiographs for the automated detection of pediatric supracondylar fracture in conventional radiography, and assess its feasibility and diagnostic performance.	Retrospective Cohort Study	1266 elbow radiographs	The study conducted external tests, using two distinct datasets—temporally separated (from 2018) and geographically separated (from another hospital, 2016-2018)—with 258 and 95 pairs of radiographs, respectively. Images underwent preprocessing and were input into a dual-input neural network. A radiologist observer study was performed on the geographic test set, comparing the model's performance metrics (AUC, sensitivity, specificity, PPV, NPV) with human readers.	Limitations include the exclusion of other traumatic elbow injuries due to a focus on feasibility in pediatric musculoskeletal radiology. Data labeling relied on radiologist consensus, and though a reasonable reference, clinical outcomes or advanced imaging were not incorporated. Additionally, being a retrospective case-control study with limited data, further multi-institutional prospective studies are needed for model performance verification and clinical use.	The dual-input deep-learning model that interprets both AP and lateral elbow radiographs provided an accurate diagnosis of pediatric supracondylar fracture comparable to radiologists.

Chung et al. (2018)	Automated detection and classification of the proximal humerus fracture by using deep learning algorithm	To evaluate the ability of artificial intelligence (a deep learning algorithm) to detect and classify proximal humerus fractures using plain anteroposterior shoulder radiographs.	Retrospective Cohort Study	1891 radiographic images	The ability of the CNN, as measured by top-1 accuracy, area under receiver operating characteristics curve (AUC), sensitivity/specificity, and Youden index, in comparison with humans (28 general physicians, 11 general orthopedists, and 19 orthopedists specialized in the shoulder) to detect and classify proximal humerus fractures was evaluated.	Limitations include the fair-to-moderate reliability of the Neer classification, absence of a gold standard for proximal humerus fracture classification, and the need for a more reliable system.	Use of artificial intelligence can accurately detect and classify proximal humerus fractures on plain shoulder AP radiographs.
Cohen et al. (2023)	Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs	To compare the performances of artificial intelligence (AI) to those of radiologists in wrist fracture detection on radiographs.	Retrospective Cohort Study	637 patients with wrist trauma	The AI software used was a deep neuronal network algorithm. Ground truth was established by three senior musculoskeletal radiologists who compared the initial radiology reports (IRR) made by non-specialized radiologists, the results of AI, and the combination of AI and IRR (IR+AI)	Limitations include the single-center and retrospective design, a low proportion of final diagnoses confirmed by CT or MRI, and lower detection rates by non-specialized radiologists, possibly attributed to their relative inexperience and the complexity of carpal bone fracture detection.	Performance of AI in wrist fracture detection on radiographs is better than that of non-specialized radiologists.
Duron et al. (2021)	Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study	To assess the performance of an artificial intelligence (AI) system designed to aid radiologists and emergency physicians in the detection and localization of appendicular skeletal fractures.	Retrospective Cohort Study	60,170 radiographs	Radiographs with quality precluding human interpretation or containing only obvious fractures were excluded. Six radiologists and six emergency physicians were asked to detect and localize fractures with (n = 300) and fractures without (n = 300) the aid of software highlighting boxes around AI-detected fractures. Aided and unaided sensitivity, specificity, and reading times were compared by means of paired Student t tests after averaging of performances of each reader.	Limitations include the reliance on image analysis alone, without access to patients' clinical data, potentially introducing a context bias, the possible influence of the Hawthorne effect on readers' performance, the exclusion of examinations with obvious fractures affecting the underestimation of unaided readers' sensitivity, the artificial 50% prevalence in fracture stratification impacting predictive value calculations, and the chosen design aimed at avoiding reader-order bias, aligning more closely with real-world conditions.	The artificial intelligence aid provided a gain of sensitivity (8.7% increase) and specificity (4.1% increase) without loss of reading speed.

Gan et al. (2019)	Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments	Ability of a CNN, with a fast object detection algorithm previously identifying the regions of interest, to detect distal radius fractures (DRFs) on anterior–posterior (AP) wrist radiographs.	Retrospective Cohort Study	2340 wrist radiographs	CNN trained to analyze wrist radiographs in the dataset. Feasibility of the object detection algorithm was evaluated by intersection of the union (IOU). The diagnostic performance of the network was measured by area under the receiver operating characteristics curve (AUC), accuracy, sensitivity, specificity, and Youden Index; the results were compared with those of medical professional groups.	Limitations include a small original sample size, the use of anterior–posterior wrist radiographs for diagnostic assessment, and the training of Inception-v4 for simple image classification.	CNN exhibited a diagnostic ability similar to that of the orthopedists and a performance superior to that of the radiologists in distinguishing AP wrist radiographs with DRFs from normal images under limited conditions.
Gasmi et al. (2023)	Comparison of diagnostic performance of a deep learning algorithm, emergency physicians, junior radiologists and senior radiologists in the detection of appendicular fractures in children	To evaluate the performance of an AI algorithm based on deep neural networks toward detecting traumatic appendicular fractures in a pediatric population. To compare sensitivity, specificity, positive predictive value and negative predictive value of different readers and the AI algorithm.	Retrospective Cohort Study	878 patients	All radiographs of the shoulder, arm, elbow, forearm, wrist, hand, leg, knee, ankle and foot were evaluated. The diagnostic performance of a consensus of radiology experts in pediatric imaging (reference standard) was compared with those of pediatric radiologists, emergency physicians, senior residents and junior residents. The predictions made by the AI algorithm and the annotations made by the different physicians were compared.	Limitations include single-center and retrospective design with no control group, though it stands as one of the initial assessments of an AI algorithm for appendicular fracture detection in pediatric digital radiographs.	Deep learning algorithms can be useful in improving the detection of fractures in children.
Gipson et al. (2022)	Diagnostic accuracy of a commercially available deep-learning algorithm in supine chest radiographs following trauma	To evaluate the performance of a commercially available deep convolutional neural network – Annalise CXR V1.2 (Annalise.ai) – for detection of traumatic injuries on supine chest radiographs.	Retrospective Cohort Study	1404 chest radiographs	Annalise.ai assessment of the chest radiograph was compared to the radiologist report of the chest radiograph. Contemporaneous CT report was taken as the ground truth. Agreement with CT was measured using Cohen's κ and sensitivity/specificity for both AI and radiologists were calculated.	Limitations include the non-simultaneous performance of radiographs and CT, potential interventions between the two, and the exclusion of major traumas and cases reported after CT acquisition, raising uncertainties about radiologists' access to additional information or views during chest radiograph reporting.	AI performed comparably to radiologists in interpreting chest radiographs.

Guermazi et al. (2022)	Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence	To assess the effect of assistance by artificial intelligence (AI) on diagnostic performances of physicians for fractures on radiographs.	Retrospective Cohort Study	480 radiographic images	The ground truth was determined by two musculoskeletal radiologists, with discrepancies solved by a third. Twenty-four readers (radiologists, orthopedists, emergency physicians, physician assistants, rheumatologists, family physicians) were presented the whole validation data set (n = 480), with and without AI assistance, with a 1-month minimum washout period. The primary analysis had to demonstrate superiority of sensitivity per patient and the noninferiority of specificity per patient at 23% margin with AI aid. Stand-alone AI performance was also assessed using receiver operating characteristic curves.	Retrospective nature of the study, reliance on radiographs without clinical information, potential contextual bias, and the artificial prevalence of fractures limit the generalizability of our findings to real-life clinical settings and raise concerns about the representativeness of the study sample, preventing the calculation of predictive values.	AI assistance improved the sensitivity and may even improve the specificity of fracture detection by radiologists and non radiologists, without lengthening reading time.
Hayashi et al. (2022)	Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning	To perform an external validation of an existing commercial AI software program (BoneViewTM) for the detection of acute appendicular fractures in pediatric patients.	Retrospective Cohort Study	300 pediatric patients	The Ground Truth was defined by experienced radiologists. A deep learning algorithm interpreted the radiographs for fracture detection, and its diagnostic performance was compared against the Ground Truth, and receiver operating characteristic analysis was done. Statistical analyses included sensitivity per patient (the proportion of patients for whom all fractures were identified) and sensitivity per fracture (the proportion of fractures identified by the AI among all fractures), specificity per patient, and false-positive rate per patient.	AI software's standalone performance evaluated without assessing its role in assisting human readers, lack of clinical information input during AI interpretation, and artificially setting prevalence of fractures at 50%, all potentially affecting the generalizability of results to real-world scenarios in pediatric emergency settings.	The BoneViewTM deep learning algorithm provides high overall diagnostic performance for appendicular fracture detection in pediatric patients.

Hendrix et al. (2022)	Musculoskeletal radiologist-level performance by using deep learning for detection of scaphoid fractures on conventional multi-view radiographs of hand and wrist	To assess how an artificial intelligence (AI) algorithm performs against five experienced musculoskeletal radiologists in diagnosing scaphoid fractures and whether it aids their diagnosis on conventional multi-view radiographs.	Retrospective Cohort Study	4796 patients with fractures	Four datasets of conventional hand, wrist, and scaphoid radiographs were retrospectively acquired at two hospitals (hospitals A and B). Dataset 1 (12,990 radiographs from 3353 patients, hospital A) and dataset 2 (1117 radiographs from 394 patients, hospital B) were used for training and testing a scaphoid localization and laterality classification component. Dataset 3 (4316 radiographs from 840 patients, hospital A) and dataset 4 (688 radiographs from 209 patients, hospital B) were used for training and testing the fracture detector. The algorithm was compared with the radiologists in an observer study. Evaluation metrics included sensitivity, specificity, positive predictive value (PPV), area under the characteristic operating curve (AUC), Cohen's kappa coefficient (κ), fracture localization precision, and reading time.	Limitations include potential mislabeling of occult scaphoid fractures as negative cases due to the absence of follow-up imaging in some patients and the trade-off problem between reference standard quality and selection bias. Additionally, the simplified model architecture processed various radiographic views with a single CNN, possibly impacting performance, suggesting potential improvement with separate CNNs for each view in future research.	The AI algorithm detects scaphoid fractures on conventional multi-view radiographs at the level of five experienced musculoskeletal radiologists and could significantly shorten their reading time.
Hrzic et a. (2022)	Fracture Recognition in Paediatric Wrist Radiographs: An Object Detection Approach	To test a machine learning model based on the YOLOv4 method for fracture recognition in pediatric wrists radiographs.	Retrospective Cohort Study	19,700 pediatric radiographic images	YOLOv4 model compared to the U-Net model for wrist fracture detection, with adjustments made to the U-Net model based on previous literature. The U-Net model included pixel-wise heat map generation and fracture probability calculation, while the YOLOv4 model, a one-stage object detector, consisted of input data preprocessing, a backbone network (CSPDarknet53), a PANet for feature aggregation, and a YOLOv3 head for fracture detection. Various YOLOv4 models with different input sizes and anchor configurations were trained and tested.	Limitations include potential bias in the dataset due to data collection from a single hospital, with implications for generalization to data from other hospitals.	YOLOv4-based model obtained significantly better results in comparison to the state-of-the-art method based on the U-Net model. Additionally, three out of five radiologists significantly improved their performance when aided by the AI model.

Kim et al. (2023)	Transfer learning-based ensemble convolutional neural network for accelerated diagnosis of foot fractures	To develop and assess the effectiveness of an AI assistant system for foot fracture diagnosis, particularly targeting interns and non-experts, using an ensemble model based on transfer learning-based convolutional neural networks.	Retrospective Cohort Study	1099 radiographs	The study employed preprocessing techniques, including resizing radiographs and applying contrast-limited adaptive histogram equalization, and utilized the ImageDataGenerator API for data augmentation. Transfer learning was employed with three pre-trained models, and an ensemble method was applied to improve fracture detection. Model evaluation involved ROC curve analysis, AUC calculation, and F1-Score assessment. Fracture localization was visualized using the Score-CAM method.	Limitations include challenges posed by deformities in the human foot, such as hallux valgus, limiting model training, and the difficulty of identifying fractures in anteroposterior view radiographs, suggesting the need for a prospective study with radiographs from various views for improved accuracy and reliability.	A transfer learning-based ensemble Convolutional Neural Network (CNN) significantly improves foot fracture detection accuracy and diagnosis time across different experience levels, suggesting the potential for efficient operational strategies and applicability to other fractures with limited datasets.
Kim et al. (2023)	Detection of incomplete atypical femoral fracture on anteroposterior radiographs via explainable artificial intelligence	Develop a transfer learning-based ensemble model to accurately detect and localize incomplete atypical femoral fractures, addressing the challenge of potential misdiagnosis, and demonstrated the effectiveness of this artificial intelligence-assisted diagnostic application in supporting decision-making and reducing clinician workload.	Retrospective Cohort Study	1050 radiographs	Development of a transfer learning-based ensemble model, using Sobel filtering to preprocess radiographs, selecting six models for transfer learning (EfficientNet B5, B6, B7, DenseNet 121, MobileNet V1, and V2), and creating two ensemble models based on the three and five models with the highest accuracy.	Limitations include the rarity of atypical femoral fractures, the absence of presented fracture probability affecting treatment guidelines, and the need for further research to evaluate detection in exceptional situations, such as femurs with implants or severe deformation.	Transfer learning-based models effectively classified and detected incomplete atypical femoral fractures with high accuracy, demonstrating the potential of the developed AI-assisted diagnostic application to support decision-making and alleviate clinician workload.
Langerhuizen et al. (2020)	Is Deep Learning on Par with Human Observers for Detection of Radiographically Visible and Occult Fractures of the Scaphoid?	To evaluate the diagnostic accuracy, sensitivity, and specificity of a deep learning algorithm in detecting radiographically visible and occult scaphoid fractures, assess the impact of adding patient demographic information on diagnostic performance, compare the diagnostic abilities of orthopedic surgeons with deep learning, and analyze interobserver reliability among human observers and between human consensus and the deep learning algorithm.	Retrospective Cohort Study	300 patients radiographic scaphoid series	The dataset, divided into training, validation, and test groups, included radiographs of scaphoid fractures and non fractures. A convolutional neural network (CNN), pretrained on non-medical images, was employed to analyze the radiographs, and the model's performance was evaluated using metrics such as area under the receiving operating characteristic curve, accuracy, sensitivity, and specificity. Additionally, the study compared the CNN's performance with that of five human observers, assessing interobserver reliability and considering demographic factors in the analysis.	Limitations include a spectrum bias introduced by selecting patients from readily available radiology reports, a relatively small sample size of 300 patients, reliance on radiologist interpretations for ground truth labels, potential bias in manual cropping and resizing of radiographs, and the exclusion of injury details, signs, and symptoms from the analysis.	The deep learning algorithm performed less effectively than human observers in identifying scaphoid fractures on radiographs, suggesting the need for further investigation with larger datasets and algorithm refinement.

Lee et al. (2023)	Clinical Validation of an Artificial Intelligence Model for Detecting Distal Radius, Ulnar Styloid, and Scaphoid Fractures on Conventional Wrist Radiographs	To assess the feasibility and performance of an artificial intelligence (AI) model for detecting three common wrist fractures: distal radius, ulnar styloid process, and scaphoid.	Retrospective Cohort Study	4432 radiographic wrist images	The study employed a four-step process, involving data input and preprocessing, deep learning-based detection using the RetinaNet model, segmentation and fracture classification using the DeepLab v3 and NasNet models, and integration of results to provide a final decision; the model was trained on a combination of the MURA dataset and local images, and its clinical efficacy was assessed in a cohort of 593 patients by comparing its performance to human experts, demonstrating high sensitivity, specificity, and accuracy for detecting distal radius and ulnar styloid fractures, as well as improved diagnostic capabilities in scaphoid fractures when assisting novice radiologists.	Limitations include single-center retrospective design primarily involving adult subjects, potentially overlooking wrist fractures in children, and the inability to detect fractures beyond specific types, necessitating radiologist interpretation.	The AI model was found to be reliable for detecting wrist fractures, particularly for scaphoid fractures, which are commonly missed.
Lind et al. (2021)	Artificial intelligence for the classification of fractures around the knee in adults according to the 2018 AO/OTA classification system	To evaluate how well an AI can classify knee fractures according to the detailed 2018 AO-OTA fracture classification system.	Retrospective Cohort Study	6003 radiographic images	ResNet-based neural network was trained on a dataset of knee radiograph exams, labeled according to AO/OTA classification, with a test set of 600 exams independently reviewed by senior orthopedic surgeons; network training involved alternating between passive and active learning sessions, and the performance was measured using area under curve (AUC), sensitivity, specificity, and Youden Index, with integrated gradients used to interpret image features contributing to the network's output.	Limitations included incomplete avoidance of bias due to excluding images with poor quality, a bias towards rare fractures, overrepresentation of fractures, lack of sophisticated ground truth establishment, potential misclassification bias, and variations in the interpretation of the AO/OTA classification system, highlighting the need for caution in generalizing findings beyond the specific hospital dataset.	Neural networks can be used not only for fracture identification but also for more detailed classification of fractures around the knee joint.

Lindsey et al. (2018)	Deep neural network improves fracture detection by clinicians	To develop and evaluate a deep neural network for detecting and localizing fractures in wrist radiographs.	Retrospective Cohort Study	132,435 radiographic images	The study developed and trained a deep learning model to detect and localize fractures in wrist radiographs using a dataset from a specialty hospital, and subsequently evaluated the model's performance on two test datasets. Additionally, a controlled experiment involving emergency medicine clinicians assessed their ability to detect fractures with and without the assistance of the trained model.	Limitations include retrospective nature, the need for a prospective study in a real clinical environment, the focus on visible radiograph information, potential impact of the model's output display on clinician performance, and the study's proof-of-concept approach targeting wrist fractures, despite the models having the potential to learn and detect various conditions on radiographs.	The deep neural network demonstrated a significant improvement in diagnostic accuracy for emergency medicine clinicians in detecting and localizing fractures in wrist radiographs when aided by the deep learning model.
Nguyen et al. (2022)	Assessment of an AI aid in detection of paediatric appendicular skeletal fractures by senior and junior radiologists	To investigate the ability of artificial intelligence (AI) to improve the detection of fractures by radiologists in children and young adults.	Retrospective Cohort Study	300 pediatric radiographic images		Limitations include an artificially balanced dataset, potentially overrepresentation of uncommon fractures and challenging scenarios, a retrospective design without access to clinical information, creating a context bias, and the absence of a washout period between reading with and without AI, possibly introducing bias in immediate AI-assisted readings.	Sensitivity increased by an average of 10% without significantly decreasing specificity in AI fracture detection in a predominantly pediatric population.
Mawatari et al. (2020)	Research article The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs	To develop deep convolutional neural network (DCNN) for detecting hip fractures using CT and MRI as a gold standard, and to evaluate the diagnostic performance of 7 readers with and without DCNN.	Retrospective Cohort Study	327 patients with proximal femoral fractures	All radiographs were manually checked and annotated by radiologists referring to CT and MRI for selecting ROI. At first, a DCNN with the GoogLeNet model was trained by 302 cases. The remaining 25 cases and 25 control subjects were used for the observer performance study and for the testing of DCNN. Seven readers took part in this study. A continuous rating scale was used to record each observer's confidence level. Subsequently, each observer interpreted with the DCNN outputs and rated them again. The area under the curve (AUC) was used to compare the fracture detection.	Limitations of the study include the potential omission of non-hip fractures, a small number of cases detected exclusively by MRI, the model providing only the probability of hip fractures without specifying their location, restrictions on input image sizes for the deep learning model, and potential selection bias in choosing fracture cases for accuracy testing.	DCNN developed using CT and MRI as a gold standard by radiologists and improved the diagnostic performance including the experienced readers.

Murphy et al. (2022)	Machine learning outperforms clinical experts in classification of hip fractures	To create a machine learning method for identifying and classifying hip fractures, and to compare its performance to experienced human observers.	Retrospective Cohort Study	3659 hip radiographs	Radiographs were processed using a two-stage convolutional neural network (CNN) approach. The first CNN (CNN1) automatically extracted two regions of interest (ROIs) containing hip joints, reducing the radiograph size. The second CNN (CNN2) classified hip fractures using these ROIs, with transfer learning from GoogLeNet. Training, testing, and validation data sets were split randomly, and performance was assessed using various metrics, including accuracy, agreement, precision, recall, F1 score, and Receiver Operating Characteristic (ROC) curves.	Subtrochanteric fractures were excluded due to lack of availability	The trained neural network can classify hip fractures with 19% increased accuracy compared to human observers with experience of hip fracture classification in a clinical setting.
Oakden-Rayner et al. (2022)	Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study		Retrospective Cohort Study	400 radiographic images	The model's performance was evaluated on a primary validation dataset from the Royal Adelaide Hospital, an MRMC dataset for reader study, and an external validation dataset from Stanford University Medical Center, demonstrating its capability in detecting proximal femoral fractures.	Limitations include the deep learning model's inability to handle cases with implanted metalwork, a relatively modest sample size in the MRMC study determined by reader availability, and a lack of racial and ethnic data for subgroup testing. Additionally, the audit findings and subgroup tests may lack statistical reliability due to individual human interpretation and small subgroups.	The model outperformed the radiologists tested and maintained performance on external validation, but showed several unexpected limitations during further testing.
Ozkaya et al. (2022)	Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography	To determine the diagnostic performance of artificial intelligence with the use of convolutional neural networks (CNN) for detecting scaphoid fractures on anteroposterior wrist radiographs.	Retrospective Cohort Study	390 AP wrist radiographs	The diagnostic performance of the CNN, ED physician and two orthopedic specialists (less experienced and experienced) as measured by AUC, sensitivity, specificity, F-Score and Youden index, to detect scaphoid fractures was evaluated and compared between the groups.	Limitations include its retrospective nature, a relatively small sample size, and the reliance on diagnosing scaphoid fractures solely from AP wrist radiographs, neglecting the comprehensive assessment typically done with standard wrist X-rays (AP and lateral) or scaphoid series.	The deep learning algorithm has the potential to be used for diagnosing scaphoid fractures on radiographs.

Suzuki et al. (2022)	Detecting Distal Radial Fractures from Wrist Radiographs Using a Deep Convolutional Neural Network with an Accuracy Comparable to Hand Orthopedic Surgeons	To evaluate the ability of CNN to diagnose distal radius fractures (DRFs) using frontal and lateral wrist radiographs.	Retrospective Cohort Study	503 patients with distal radius fracture	Limitations include a binary classification without localization of the pathological region, reliance on a relatively small dataset, especially for pediatric cases, potential challenges for less experienced clinicians in trusting broad classification labels, and the disadvantage of hand orthopedic surgeons judging small cropped images without additional clinical information available in typical settings.	Frontal and lateral views of wrist radiographs were manually cropped and trained separately. Fine-tuning was performed using EfcientNets. The diagnostic ability of CNN was evaluated using 150 images with and without fractures from anteroposterior and lateral radiographs. The CNN model diagnosed DRF based on three views: frontal view, lateral view, and both frontal and lateral view. The sensitivity, specificity, and accuracy of the CNN model was determined, and plotted a receiver operating characteristic (ROC) curve, and calculated the area under the ROC curve (AUC).	The CNN model exhibited high accuracy in the diagnosis of distal radius fracture with a plain radiograph.
Tanzi et al. (2020)	Hierarchical fracture classification of proximal femur X-Ray images using a multistage Deep Learning approach	To design a Deep Learning-based tool able to help doctors in diagnosis of bone fractures, following the hierarchical classification proposed by the Arbeitsgemeinschaft für Osteosynthesefragen (AO) Foundation and the Orthopaedic Trauma Association (OTA).	Retrospective Cohort Study	2453 annotated radiographic images	Two methods were employed: a fine-tuned InceptionV3 convolutional neural network (CNN) as a baseline and a multistage architecture of successive CNNs tailored to the hierarchical structure of the AO/OTA classification; evaluation metrics included accuracy, area under the receiver operating characteristics curve (AUC), recall, precision, and F1-score for the CNN, and accuracy and Cohen's Kappa coefficient for orthopedists both with and without the assistance of the CNN, with relevant areas visualized using Gradient Class Activation Maps (Grad-CAM).	None noted	Use of a CAD system based on CNN improves diagnosis accuracy.

Twinprai et al. (2022)	Artificial intelligence (AI) vs. human in hip fracture detection	To assess the diagnostic accuracy and sensitivity of a YOLOv4-tiny AI model for detecting and classifying hip fractures types.	Retrospective Cohort Study	1000 hip and pelvic radiographic images	A deep convolutional neural network (YOLOv4-tiny) was trained on an augmented dataset of hip images with bounding box annotations for normal, femoral neck fracture, intertrochanteric fracture, or subtrochanteric fracture labels, and its performance was assessed on a separate testing set, with subsequent comparison to human doctors' evaluations, ensuring no crossover data in the testing set.	Limitations include limited generalizability as the model was trained exclusively at a single institution, a small dataset of 1000 images, and challenges in model accuracy arising from false-positive cases related to poor film positioning and artifacts, as well as false-negative cases involving non-displaced fractures.	The model showed hip fracture detection sensitivity comparable to well-trained radiologists and orthopedists and classified hip fractures highly accurately.

Table 2: Summary of the articles included in the review (N = 36)