

Table S1: Included studies methodological and numerical data

Study	Study Period	Imaging Modality	X-ray Exclusion Criteria	Number of CNN Pipeline Steps	CNN Sub-Purpose	Number of Patients	X ray views; per patient	Number of Images	Type of Fracture	Standard of Reference for Diagnosis	Human Comparison
Langerhuizen et al [34]	NA	X-Ray	Incomplete, distorted by cast or splint materials	1	Fracture Detection	300	4 (AP, PA, Lateral, Oblique Projections)	1200	Visible Fracture = 127 Occult Fracture = 23 No Fracture = 150	MRI-Confirmed or CT-Confirmed	5 Board-Certified Orthopedic Surgeons
Ozkaya et al [16]	NA	X-Ray	NA	1	Fracture Detection	390	1 (AP Projection)	390	Fracture = 192 No Fracture=198	Follow up CT scans	1 Emergency Doctor and 2 Orthopedic Hand Surgeons
Yoon et al [35]	January 2001 to December 2019	X-Ray	Poor quality, fractures older than 4 weeks, chronic hand conditions with bony changes around the scaphoid, psoriatic arthritis, rheumatoid arthritis, external immobilization (casts, splints, external fixations), hardware (screws, plates, wires, pins	3	Scaphoid Segmentation Model	4583	Exact number per patient not elaborated (PA or Scaphoid View)	11838	NA	A group of senior musculoskeletal radiologists provided final image interpretations. Radiographs with ambiguous or conflicting reports were reviewed by a hand surgeon. Final diagnoses were made based on the surgeon's interpretation. some cases had CT or MRI scans	NA
					Apparent Fracture Detection Model	4583			Fracture = 3991 No Fracture = 5542		
					Occult Fracture Detection Model	547			Occult Fracture = 139 No Fracture = 556		
					Full Pipeline	NA			Apparent Fracture = 904 Occult Fracture = 22 No Fracture = 1379		
Hendrix et al [36]	2017-2019	X-Ray	Old scaphoid fractures, screws or other implants, resection, excessive damage or malformation, cast	2	Scaphoid Segmentation	775	Exact number per patient not elaborated (AP or PA Projections of Hand, Wrist and Scaphoid Radiographs)	1039	NA	NA	NA
	2003-2019				Fracture Detection	1846		3000	Fracture = 1165 No Fracture = 1835	Original radiology reports. Dubious scaphoid fractures were reevaluated by a radiologist	NA
	2011-2020				Full Pipeline	190		190	Fracture = 95 No Fracture = 95	As Above and follow up CT scans	11 Radiologists, including: 3 residents, 6 Musculoskeletal (MSK) specialized senior Radiologists, 2 Non-MSK Senior radiologists
Tung et al [37]	NA	X-Ray	NA	2	Scaphoid Segmentation Fracture Detection	154	Wrist Radiographs (Frontal or Lateral)	356	Fracture = 178 No Fracture = 178	Radiology Report	NA

Yang et al [38]	NA	X-Ray	NA	2	Scaphoid Segmentation	280	NA	361	Fracture = 167	Surgical Verification	NA
					Fracture Detection	280		360	No Fracture = 194		
Li et al [2]	2010-2020	X-Ray	Poor quality, screw or other implants, tumors affecting wrist bones, tuberculous or rheumatoid arthritis	2	Scaphoid Segmentation	500	Wrist Radiographs (PA or Scaphoid View)	930	Fracture = 513	For Occult Fractures: 61 had follow up CT Scans 11 had follow up MRI Scans	Majority decision of 4 hand surgeons In case of no majority, a 5th hand surgeon joined to obtain majority
						Apparent Fracture -170					
						Occult Fracture – 60					
						No Fracture –270					
					Fracture Detection	100		209	No Fracture = 626		
						Apparent Fracture -38					
						Occult Fracture – 12					
					Full Pipeline	No Fracture – 50		1139			
						600					
						Apparent Fracture -208					
Occult Fracture - 72											
Hendrix et al [39]	2003-2019	X-Ray	Occlusion by casts, Occlusion by metallic osteosynthesis implants, Resection, not grown, unrecognizably deformed, non-acute fractures and severe scapholunate advanced collapse	2	Scaphoid Segmentation (Including laterality and x-ray projection detection)	394	Hand, Wrist and Scaphoid Radiographs (AP/PA, Ulnar-deviated AP/PA, Oblique and Lateral)	1117	NA	2 MSK-specialized radiologists. Cases were independently reviewed. Disagreements were resolved by consensus reading. Some cases had follow up CT or MRI scans	5 experienced MSK radiologists with 5, 7, 22, 24 and 26 years of experience Clinical Validation study: Interval between non-assisted radiograph reading and AI-assisted reading = 4 months
					Fracture -NA						
					Fracture Detection	No Fracture -NA		688 (different x-ray views) 219 Scaphoids (grouped by patients)	Fracture = 65		
						209					

Cohen et al [40]	2017-2019	X-Ray	Braces or casts, technical limitations	Commercially available AI system ("BoneView" by Gleamer)	Scaphoid Segmentation & Fracture Detection are integral	637 Study tested for various wrist fractures, including scaphoid fractures	NA	1917	Scaphoid Fractures = 25	3 MSK-specialized radiologists with 5, 5 and 7 years of experience. Results were obtained by consensus.	Initial radiology results made by a total of 41 radiologists with various levels of experience, including: 29 residents (4th or 5th year of residency), 8 fellows in radiology, and 4 attendings. Clinical Validation study Comparison of diagnostic performance in 3 groups: AI, IRR & IRR+AI. IRR+AI observation was considered positive when it was detected by either the AI or the IRR, regardless of the other's group result.
Lee et al [41]	2010-2020	X-Ray	Anatomic variation, bone abnormality, except for fractures (for example - tumor), splint or cast, orthopedic hardware, poor image quality	2 Parallel CNNs integrated to final result	Scaphoid Segmentation Fracture Detection	593 Study tested for various wrist fractures, including scaphoid fractures	Exact number per patient not elaborated. Wrist Radiographs (AP, Lateral, Oblique)	1186	Scaphoid Fractures = 32	1 MSK-specialized radiologist and 1 orthopedic upper-limb surgeon; both with 19 years of experience.	2 radiologists: 2-year-fellowship-trained musculoskeletal radiologist 1-year-trained radiology resident Clinical Validation study Interval between non-assisted radiograph reading and AI-assisted reading = 3 weeks

NA = Not Available
MSK =Musculoskeletal
AI = Artificial Intelligence
IRR = Initial Radiology Reports

Table S2: Included Studies artificial intelligence characteristics

Study	Fracture Detection - CNN Architecture Type	Scaphoid Segmentation	ROI or important region labeling	Data Input proportion in training / validation / test	Augments
Langerhuizen et al [34]	VGG16	Rectangular ROI wrapping the scaphoid. Manually cropped and resized.	NA	Training - 180	10-fold augmentation of the training and validation set by using rotation (-15° and +15°), shifting of height and width (10%), zooming (between 0.8 and 1.1), and horizon flipping. The test set only composed of original radiographs.
				Validation - 20	
				Test - 100	
Ozkaya et al [16]	ResNet50*	Rectangular ROI wrapping the scaphoid. Marked by Expert radiologist	NA	Training - 203 Validation - 87 Test - 100	NA
Yoon et al [35]	EfficientNetB3	Scaphoid detection model based on Cascade Region-based Convolutional Neural Network; was trained and used to isolate the scaphoid in a bounding box in hand radiographs	Highlight important regions = Grad-CAM	Total Pipeline	random flip, scale, random brightness, rotation (<15°), flip, resizing, and standardization.
				Training - 8356	
				Validation - 1177	
				Test - 2305	
				Apparent Fracture Model	
				Training - 8306	
				Validation - 1151	
				Occult Fracture Model	
				Training - 565	
Validation - 130					
Hendrix et al [36]	DenseNet-121	"Segmentation CNN"	Highlight important regions = Grad-CAM	NA	Horizontal Flip, Image Shift, Bounding Box Shift, Scale, Rotate, Grid Distortion, Elastic Transform, Brightness, Contrast
Tung et al [37]	VGG16, VGG19, ResNet50, ResNet101, ResNet152, DenseNet121, DenseNet169, DenseNet201, Inception-V3, EfficientNetB0	YOLOv4	NA	Training + Validation - 70% Test - 30%	Flip, Rotaion +/- 15
Yang et al [38]	ResNet-152	RCNN +ResNet50 Cascade R-CNN (Region-based Convolutional Neural Network) marked the fracture in a bounding box	NA	NA	Contrast limited adaptive histogram equalization (CLAHE), Random horizontal flip with 50% probability, Random contrast with 50% probability
Li et al [2]	MobileNetV3	YOLOv3	Highlight important regions = Grad-CAM	Training + Validation - 930 Test -209	NA
Hendrix et al [39]	Inception-V3	YOLOv5s	NA	Scaphoid localizer and laterality classifier: Training + Validation – 12990 9% overlap with prior work by Hendrix et al [36]	Horizontal flipping Rotation Gaussian Noise
				Scaphoid fracture detector: Training + Validation – 4316 82% overlap with prior work by Hendrix et al [36] Test – 688 5% overlap with prior work by Hendrix et al [36]	

Cohen et al [40]	Detectron 2 Framework	NA	Zone of interest, integral to the AI system: "solid line box" considered as a certain fracture with a high level of specificity "dotted line box" considered as a probable fracture with a high level of sensitivity	"BoneView" by Gleamer (commercially available AI system) 60170 radiographs: Training - 70% Validation - 10% Internal tests - 20%	NA
Lee et al [41]	NasNet	RetinaNet	Heat Map	MURA dataset (public radiographic dataset) – 3791 for distal radius & ulnar styloid fractures Hospital dataset – 641 for scaphoid fractures validation set - 20%	Contrast limited adaptive histogram equalization (CLAHE) Normalization Histogram Matching Sharpening

* Replaced the uppermost layers with a set of layers that performed two class classifications.

NA = Not Available

AI = Artificial Intelligence

Table S3: Included Studies Diagnostic Test Accuracy, and comparison to human performance

Study	AI Fracture Diagnosis					Human Fracture Diagnosis					
	CNN	Accuracy	AUC	Sensitivity	Specificity	Experience	Accuracy	AUC	Sensitivity	Specificity	
Langerhuizen et al [34]	VGG16	72.00%	0.77	84.0%	60.0%	Board-Certified Orthopedic Surgeons	84.0%	NA	76.0%	93.0%	
Ozkaya et al [16]	ResNet50*	NA	0.84	76.0%	92.0%	Experienced Orthopedic Surgeon	NA	0.92	86.0%	98.0%	
						Less Experienced Orthopedic Surgeon	NA	0.82	72.0%	92.0%	
						Emergency Doctor	NA	0.76	62.0%	90.0%	
Yoon et al [35]	EfficientNetB3	NA	Total Pipeline			NA					
			X	97.2%	66.0%						
			Apparent Fracture Model								
			0.955	87.1%	92.1%						
Hendrix et al [36]	DenseNet-121	NA	Fracture Detection			NA					
			0.86	66.0%	90.0%						
			Entire System Test								
			0.87	78.0%	84.0%	Average of 11 Radiologists [Range]	NA	0.83 [0.79-0.87]	NA [54.0-80.0%]	NA [75.0-95.0%]	
Tung et al [37]	Fracture Detection					NA					
	VGG 16	83.3%	0.86	86.1%	80.6%						
	VGG19	77.8%	0.87	83.3%	72.2%						
	ResNet50	86.1%	0.91	88.9%	83.3%						
	ResNet101	88.9%	0.95	88.9%	88.9%						
	ResNet152	80.6%	0.88	80.6%	80.6%						
	DenseNet121	87.5%	0.93	91.7%	83.3%						
	DenseNet169	88.9%	0.89	91.7%	86.1%						
	DenseNet201	90.3%	0.91	94.4%	86.1%						
	Inception-V3	88.9%	0.93	88.9%	88.9%						
EfficientNetB0	86.1%	0.92	94.4%	77.8%							
Yang et al [38]	ResNet-152	Detection of Fracture Area			NA						
		85.3%	0.92	78.9%						90.0%	
		Fracture Detection									
			82.9%	0.917	73.5%	92.0%					
Li et al [2]	MobileNetV3	NA	0.919	82.0%	94.0%	Majority between hand surgeons	NA	NA	76.0%	96.0%	

Hendrix et al [39]	Inception-V3	NA	All available X-ray projections			Average of 5 MSK Radiologists [Range]	NA	0.87	[75.0%-83.0%]	[81.0%-94.0%]
			0.88	72.0%	93.0%					
			PA							
			0.79	51.0%	93.0%					
			PA + ulnar-deviated PA							
			0.84	59.0%	93.0%					
			PA + oblique							
0.85	66.0%	93.0%								
PA + lateral										
0.83	55.0%	93.0%								
Cohen et al [40]	"Bone View" by Gleamer	NA	NA	For scaphoid subset = 84.0%	NA	Initial Radiology Reports	NA	NA	80.0%	NA
Lee et al [41]	NasNet	74.0%	0.81	87.0%	74.0%	2-year-fellowship-trained MSK radiologist without AI assistance	98.0%	0.75	50.0%	99.0%
						2-year-fellowship-trained MSK radiologist with AI assistance	98.0%	0.85	72.0%	98.0%
						1-year-trained radiology resident without AI assistance	94.0%	0.71	47.0%	96.0%
						1-year-trained radiology resident with AI assistance	93.0%	0.80	66.0%	94.0%

* Replaced the uppermost layers with a set of layers that performed two class classifications.

NA = Not Available

MSK = Musculoskeletal

AI = Artificial Intelligence

Table S4: Key for stratification and grading of the included studies using the QUADAS-2 tool

Patient Selection	High	Exclusion of "difficult to diagnose" cases, for example radiographs with casts.
	Unclear	No specification of exclusion criteria.
	Low	Full random sampling of eligible patients /radiographs.
Index Test	Low	The AI systems are not aware of the "ground truth" results, so they were all deemed low.
Reference Standard	High	Not stated clearly if fractures and non-fractures were confirmed by either CT or MRI, or by expert/s.
	Unclear	Fractures and non-fractures were assessed by expert/s. Some cases or none, were confirmed by CT or MRI.
	Low	Fractures and non-fractures were confirmed by either CT or MRI.
Flow and Timing	High	Clinical validation studies with a relatively short period of time (less than a month) between non-AI assisted readings, compared with AI-assisted readings.
	Unclear	Clinical validation studies with over a month time between non-AI assisted readings, compared with AI-assisted readings.
	Low	Studies without clinical validation, in which there is no need for an appropriate interval in between consecutive readings of radiographs.